Conditional Regression Forests for Human Pose Estimation

Min Sun[†⋆*] Pushmeet Kohli[†] Jamie Shotton[†]

[†]Microsoft Research Cambridge [⋆]University of Michigan at Ann Arbor

Abstract

Random forests have been successfully applied to various high level computer vision tasks such as human pose estimation and object segmentation. These models are extremely efficient but work under the assumption that the output variables (such as body part locations or pixel labels) are independent. In this paper, we present a conditional regression forest model for human pose estimation that incorporates dependency relationships between output variables through a global latent variable while still maintaining a low computational cost. We show that the incorporation of a global latent variable encoding torso orientation, or human height, etc.*, can dramatically increase the accuracy of body joint location prediction. Our model also allows efficient and seamless incorporation of prior knowledge about the problem instance such as the height or orientation of the human subject which can be available from the problem context or via a temporal model. We show that our method significantly outperforms state-of-the-art methods for pose estimation from depth images. The conditional regression model proposed in the paper is general and can be applied to other problems where random forests are used.*

1. Introduction

In the last few years, random forests have become increasingly popular in computer vision. They have been successfully applied to problems such as image classification [16], object detection and segmentation [10, 22], human pose estimation [12, 29], action recognition [10], and even image completion [24]. Their popularity can be attributed to their simplicity and relatively low computational complexity at test time. In particular, [12, 29] have shown that random forests can be used to perform human pose estimation from depth images in real time.

The method proposed by Shotton et al. [29] works by classifying pixels into body parts (an intermediate representation). The body joints are then predicted as the modes of the parts density. They demonstrate that, given a 2D depth image, a large quantity of synthesized motion capture (mocap) training data, a simple pair-wise depth comparison feature, and an efficient random forest pixel classifier, it is sufficient to achieve impressive accuracy in real-time, even with the assumption that body joint locations are independent from each other.

Unlike Shotton et al. [29]'s classification based framework, Girshick et al. [12] formulate body joint prediction as a regression problem which avoids intermediate body part classification. Their method uses the same feature and mocap data to learn a regression forest which directly predicts the locations of body joints from all pixels. Most importantly, it proposes a compact regression model at each leaf-node of the forest which approximates the distribution by a Gaussian mixture model with only a few components (*e.g.* typically 2 components). This method obtains state-of-the-art results and is extremely efficient but does not achieve a perfect accuracy. The low computational complexity of the algorithm can be attributed to its assumption that locations of different body parts are independent from each other which is not true in practice.

In this paper, we present a conditional regression forest model for human pose estimation that incorporates dependency relationships among output variables through a global latent variable while still maintaining low computational cost. The latent variable can encode any property of the pose estimation problem instance such as the torso orientation, or the height of the human subject. When the global variable is uncertain/unknown, the body joint locations are no longer independent as assumed in both [29, 12]. We propose an efficient approach (Sec. 5) to jointly estimate the body joint locations and the global variable, and demonstrate that even when the global variable is completely unknown, the mean error reduced significantly (up to 6.6% in Sec. 6.1 which is a relative 25% reduction).

In many applications of pose estimation, some prior knowledge about the problem is known. For instance, while estimating the pose of a person playing a golf game, information about the player's height or orientation might be available. Similarly, in a surveillance application, we might know the walking directions of pedestrians. Unlike the regression forest based model of [12] that cannot handle such prior knowledge information, our model can incorporate it seamlessly. This ability also allows our model to exploit temporal consistency. For instance, the lengths of a person's limbs do not change while the person is playing the game. In such situations we can refine our prior belief about the lengths of the person's limbs by aggregating information across a video sequence and then using the refined belief to make body joint location predictions (Sec. 5.1). This process is extremely efficient and incurs almost no additional cost.

[*]This work was conducted while Min Sun was an intern in Microsoft Research Cambridge.

To summarize, the key differences between our model and the traditional random forests are: i) the posterior distribution of our model does not factorize over individual body joint locations, and is a joint distribution over the global latent variable and the body joint locations; ii) the regression model associated with each leaf node (codeword) in our model is conditioned on a global variable; iii) our model allows the incorporation of prior knowledge about the problem instance such as information about the player's height or torso orientation.

2. Related Work on Human Pose Estimation

Human Pose Estimation (HPE) (*i.e.* estimating body joint locations) is one of the most important problems in computer vision, with applications in gaming, human-computer interaction, security, telepresence, and even health-care [1, 33, 14, 23]. Naturally, the problem has attracted a lot of research (surveyed in [21, 26]). Researchers have tried to exploit a variety of different inputs, from 2D silhouettes (obtained from background subtraction) [9] to 2D intensity images [27, 7, 2].

The availability of high-speed depth sensors has led to tremendous progress in human pose estimation [13, 15, 30, 25, 11]. However, many real-time systems, achieving high speeds by tracking from frame to frame, struggle to re-initialize quickly and are not robust to tracking failure. Recently, two random forest based algorithms proposed by [29, 12] achieve super-real-time human pose estimation from a single depth image captured by the Kinect sensor [20]. The super-real-time algorithms are designed to complement any appropriate tracking algorithm [31, 13, 28, 11] that incorporates temporal and kinematic coherence, so that the overall system is much more robust to tracking failure. However, these algorithms are still far from perfect.

Methods performing human pose estimation in 2D images or video sequences, like mixture of experts [4], structure prediction models [3] and latent variable models [18, 19], can also be applied to depth data. However, unlike [29, 12] which utilize random forest, these methods cannot achieve super-real-time performance.

Conditional models Many researchers have shown that conditional models are very effective at improving accuracy. For example, in the field of object recognition, Felzenszwalb et al. [8] have shown how detectors conditioned on different aspect ratios of the object bounding box mitigate the appearance variation caused by viewpoint changes. Concurrent with our work, Dantone et al. [6] have used a conditional regression forests model for facial feature detection. They show that head pose can be used as a global variable to improve the detection accuracy. Their method works by first estimating the head pose and then using it to detect facial features. It is similar to our "*MaxA*" and "*Mar.A*" methods (described in Sec. 5) but different from our "*Joint*" method (described in Sec. 5) that tries to jointly infer the labelling of the global variable and the body joint locations. Conditioning a model, in the worst case, results in an increase in the model complexity which is linear in number of states of the global variable. Moreover, if each conditional model is trained independently, only a subset of training data is used to train each model. To overcome the first issue, we employ a shared model structure so that the model complexity typically increases much slower than the worse case (the *Partial* model in Sec. 3.1). The large synthetic training dataset (similar to that in [29, 12]) helps avoid the second issue.

3. Conditional Regression Model (C-RM)

This section provides a formal description of our model. In what follows, we will use I to denote the image data, L to denote a set of voting elements l (*e.g.* a patch or a pixel), and C to denote the set of leaf node ids (also called codewords $\{c_l\}$) that any voting element l can map to.

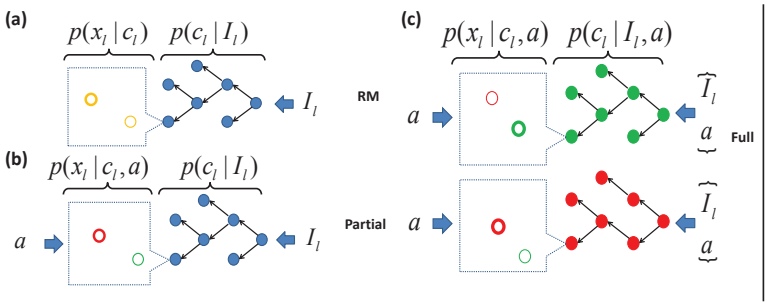
Given an image I, the method proposed in [12] regresses locations of body joints independently. For each body joint j, the method calculates a scoring function S defined over the 3D body joint location $x_j \in \mathbb{R}^3$. This function can be written as a sum of probabilistic votes contributed from all voting elements. The probabilistic vote $p(x_j, c_l|I_l)$ can be further decomposed into the distribution of 3D body joint locations for each codeword $p(x_j|c_l)$ and the probability that the image patch is mapped to a codeword $p(c_l|I_l)$ (referred to as the codeword mapping probability). More formally,

$$\begin{aligned} S(x_j|I) &= \sum_{l\in L}\sum_{c_l\in C} p(x_j, c_l|I_l) \\ &= \sum_{l\in L}\sum_{c_l\in C} p(x_j|c_l)p(c_l|I_l)\,. \end{aligned} \quad (1)$$

A random forest is utilized to discriminatively map each voting element to a specific codeword. The probability $p(c_l|I_l)$ is a delta function at the leaf (codeword) reached in the tree. The method of [12] outputs the value of x_i with the maximum score. To de-clutter the formal description, we will drop the subscript j in subsequent equations. In Fig. 1(a,b,c), one tree in the random forest is illustrated by the round nodes and directed edges, and the probabilistic votes are represented by the circles in the plate, where the location of the circle denotes the voting direction and the thickness of circle is proportional to the probability.

3.1. Conditioning on Global Variable

Some body joint locations are strongly dependent on global variables like torso orientation, human height, *etc*. Intuitively, conditioning the joint locations on these global variables would constrain them. We propose the following two models to capture the conditional property.



(d)

	Time	Memory
RM (Equations 1,10)	$T_t + T_v$	$S_t + S_v$
Full+Joint (Equations 2,4,5,11)	$AT_t + AT_v$	$AS_t + \hat{A}S_v$
Partial+Joint (Equations 3,4,5,11)	$T_t + AT_v$	$S_t + \hat{A}S_v$
Max A (Equations 4,5,14)	$T_t + T_v + T_a$	N.A.
Partial+Mar. A (Equations 3,4,5,14)	$T_t + T_v + T_a$	N.A.
Full+Mar. A (Equations 2,4,5,14)	$AT_t + T_v + T_a$	N.A.

Figure 1: We illustrate the different regression models, including (a) Regression Model (RM) [12], (b) partially conditional regression model (*Partial*), and (c) fully conditional model (*Full*). The tree graph in each panel illustrates the random forest codebook. The round nodes consisting of the tree structure represent the nodes in the random forest, where a splitting function stored at each internal node and a regression model is stored at each leaf node. The circles in the square plates represent the modes of the voting direction, where the thickness and the color of the circle indicates the confidence and the state of the global variable, respectively. Notice that the confidence of the green mode is stronger in the green forest rather than the red forest in the full model and visa versa. (d) Comparing the time and memory usage of different models (*Full* and *Partial* in Sec. 3.1) and different recognition approaches (*Joint*, *MaxA*, and *Mar.A* in Sec. 5). T_t and T_v are time for evaluating the tree and voting for the regression model. T_a is the time to evaluate the global variable scoring function in Eq. 14. S_t and S_v are the space for saving the tree structure and for accumulating the votes. A is the number of states of the global variable and $\hat{A} < A$ is number of time more voting modes stored in the regression model.

Full Model We define a new scoring function S which introduces dependencies between the body joint locations and a global variable $a \in \mathcal{A}$ as,

$$\begin{aligned} S(x|a,I) &= \sum_{l\in L}\sum_{c_l\in C} p(x,c_l|a,I_l) \\ &= \sum_{l\in L}\sum_{c_l\in C} p(x|c_l,a)p(c_l|a,I_l)\ . \end{aligned} \quad (2)$$

We call this the full conditional regression model (referred to as the *Full* model in Fig. 1(c)) since we model both the conditional distribution of 3D body joint locations for each codeword $p(x|c_l,a)$, and the conditional codeword mapping probability $p(c_l|a,I_l)$. Notice that the drawback of the *Full* model is that a separate tree structure per state of the global variable needs to be learned and stored during learning, and evaluated during recognition.

Partial Model We define a partial conditional regression model (referred to as the *Partial* model in Fig. 1(b)) which overcomes this computational expense by sharing tree structures (codebooks) for all states of the global variable as,

$$S(x|a,I) = \sum_{l\in L}\sum_{c_l\in C} p(x|c_l,a)p(c_l|I_l)\ . \quad (3)$$

This greatly reduces the model complexity and training time since we do not need to train and store separate random forest for each state of the global variable. Moreover, as indicated by our experimental results, it even produces the most accurate pose estimation results.

3.2. Prior Knowledge of Global Variable

In certain application scenarios, prior knowledge about attributes of the human subject such as height, gender or torso orientation is available. This information can be incorporated in our scoring function as a prior probability $p(a)$ over the corresponding global variable as,

$$S(x,a|I) = S(x|a,I)p(a)\ . \quad (4)$$

When no prior knowledge is available, we assume $p(a)$ is a uniform probability over all states for both the full and partial models. In this case, for the *Partial* model of a single body joint, the original scoring function in Eq. 1 is equivalent to the scoring function in Eq. 4 marginalized over the global variable a (*i.e.* $S(x|I) \propto \sum_a S(x,a|I) = \sum_{l\in L}\sum_{c_l\in C}\sum_a (p(x|c_l,a)p(a))p(c_l|I_l)$).

Dependency among body joint locations. Since all body joints depend on the shared global variable a, we define the following scoring function which is a function of all body joints and the global variable as,

$$S(X,a|I) = \sum_{j\in J} S(x_j,a|I) = p(a)\sum_{j\in J} S(x_j|a,I)\ . \quad (5)$$

where x_j is the 3D location of the j^{th} body joint and $X = \{x_j\}_{j\in J}$ is the set of all body joints. Hence, the dependency among body joint locations is established through sharing the global variable. Moreover, body joint locations become independent to each other when the state of the global variable is given.

3.3. Exploiting Temporal Consistency

The states of the global variable associated with the pose estimation problem remain the same in certain situations. For example, while estimating the pose of the same individual in multiple frames, we know that their body height would be the same. We can exploit this temporal consistency in the global variable to improve the estimate of its true state, and in turn the estimates of body part locations. We define a scoring function for multiple frames with a shared global variable as,

$$S(\mathcal{X}^F, a|\mathcal{I}^F) = \sum_{f\in\{1,\ldots,F\}} S(X^f, a|I^f)\ . \quad (6)$$

where $\mathcal{X}^F = \{X^f; f = 1\ldots F\}$, and $\mathcal{I}^F = \{I^f; f = 1\ldots F\}$ are the sets of body joints and image evidence from the 1^{st} frame to the f^{th} frame, respectively. Notice that all frames share the same global variable a since it is stationary in time.

4. Learning

This section describes how the different factors used in our model were learned.

4.1. Conditional Codeword Mapping

We partition the training data based on the value of the attribute corresponding to the specific instance in the training data. We then learn the random forests (to act as discriminative codebooks) corresponding to each partition of the training data. The conditional probabilistic mapping function can be represented as

$$p(c_l|a, I_l) = \frac{1}{T} \sum_{t \in \{1,\dots,T\}} \delta(c_t^{q(a)}(I_l) - c_l) \; . \tag{7}$$

where T is the number of trees, $c_t^{q(a)}$ is the mapped codeword for tree t in the forest corresponding to partition $q(a)$. Notice that each random forest is trained using the same procedure as described in [29] which minimizes the Shannon entropy for a classification task. Notice that the conditional codebook is only used in the full model but not in the partial model.

4.2. Conditional Distribution of Votes

We use the same algorithm as [12] to learn a compact regression model $p(x|c_l, a)$ for each codeword conditioned on each state of the global variable. Specifically, we use the mean-shift algorithm with a Gaussian kernel of bandwidth b^* to cluster the relative votes and obtain the largest K clusters. Each cluster consists of the relative vote Δ_{kca}, given by the mean-shift mode, and a confidence weight w_{kca}, given by the size of the cluster. The conditional distribution of votes is approximated as $p(x|c_l, a) \propto$

$$\sum_{k \in K} w_{kc_la} \cdot \exp\left(-\| \frac{x - (\Delta_{kc_la} + z_l)}{b} \|_2^2\right) \cdot \delta(\|\Delta_{kc_la}\|_2^2 \leq \lambda) \, ,$$

where z_l is the 3D location of the voting element l, b is the kernel bandwidth used to approximate the distribution, λ is the maximum distance threshold that a relative vote should be used.

4.3. Relating States of the Global Attribute

The different states of the global variable used in our model might be related to each other. For example, the torso orientation and person height are both intrinsically continuous variables. Intuitively, a person facing 15 degree left should share similar body joint locations with a person facing 10 degree left compared to a person facing 20 degree right. This structural information is lost when we model the global attribute using a set of discrete states. To reinstate these relationships, we utilize a vote transfer scheme which enables votes from other states of the global attribute to contribute to the scoring function,

$$p(x|c_l, a) = \sum_{\hat{a}} p(x|c_l, \hat{a}) p(\hat{a}|a) \; . \tag{8}$$

where $p(\hat{a}|a)$ is the transfer probability that re-weighted the votes ($p(x|c_l, \hat{a})$) conditioned on state $\hat{a} \in \mathcal{A}$. The transfer probabilities are learned using a model selection approach on a validation set. We use a symmetric and linearly[1] decreasing transfer probability defined as,

$$p(\hat{a}|a) \propto \left\{ \begin{array}{ll} 1 - \frac{d}{2D_a} & \text{if } d \leq D_a \\ 0 & \text{otherwise} \end{array} \right\} \; . \tag{9}$$

where $d = |a - \hat{a}|$ is the distance between the two states, and parameter D_a specifies the maximum vote transfer distance.

5. Inference

We now describe how to estimate the body joint locations and the state of the global variable from the scoring functions defined earlier. If the true state of the global variable is known, then we can estimate each body part location "independently" by using the mean-shift [5] mode finding algorithm to find the location with the maximum score as

$$x_j^* = \underset{x}{\text{argmax}} \, S(x_j|a, I); \forall j \in J \; . \tag{10}$$

where J is the set of all body joints and the scoring function is defined in Eq. 2 and 3. When the true state of the global variable is unknown, the part locations are not independent to each other and need to be inferred jointly with the global variable. We now describe our methods to solve this inference problem.

Jointly estimate X and a (*"Joint"*). Instead of independently estimating each body joint location, we use mean-shift algorithm to jointly find all body joint locations $\{x_j^*\}$ and the global variable a^* corresponding to the maximum score in Eq. 5 as

$$(X^*, a^*) = \underset{X,a}{\text{argmax}} \, S(X, a|I) = \underset{X,a}{\text{argmax}} \, p(a) \sum_{j \in J} S(x_j|a, I) \; . \tag{11}$$

The solution of the MAP inference problem defined in Eq. 11 can be efficiently solved by first calculating

$$x_j^*(a) = \underset{x_j}{\text{argmax}} \, S(x_j|a, I), S_j^*(a) = \max_{x_j} S(x_j|a, I) \tag{12}$$

for all a. Then, a^* and X^* are obtained as

$$a^* = \underset{a}{\text{argmax}} \sum_{j \in J} S_j^*(a), X^* = \{x_j^*(a^*); j \in J\} \; . \tag{13}$$

[1] We also tried a Gaussian shape model but found that it achieves accuracy similar to the linear one.

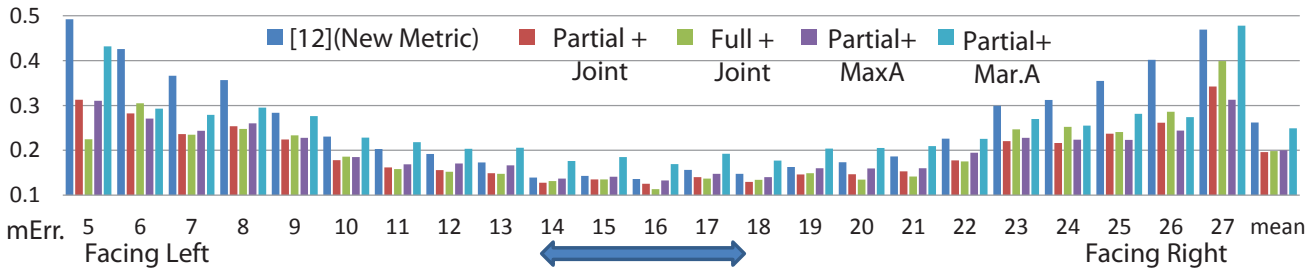


Figure 2: Mean body joint prediction error (y-axis) comparison between [12], *Joint* recognition approach on *Partial* and *Full* models (*Partial* + *Joint* and *Full* + *Joint*), and two other recognition approaches on *Partial* model (*Partial* + *MaxA* and *Partial* + *Mar.A*) for different torso orientations (x-axis).

By jointly estimating all body joint locations and the states of the global variable, the approach becomes more robust than the original regression approach. However, mean-shift needs to be applied A (*i.e.* the number of states of the global variable $|\mathcal{A}|$) times compared to once in the original regression approach (See row 2 and 3 in Fig. 1(d)). In order to be both robust and efficient, we also explore the following approaches.

Estimate a then X ("*MaxA*"). We propose to simplify the scoring function in Eq. 4 to a scoring function of global variable a as follows,

$$S(a|I) = \sum_{l \in L} \sum_{c_l \in C} p(c_l, a|I_l) = \sum_{l \in L} \sum_{c_l \in C} p(a|c_l) p(c_l|I_l) \ . \quad (14)$$

This allows us to estimate a^* efficiently as $\arg\max_a S(a|I)$. Given a^*, we can use Eq. 10 to estimate the 3D body joint locations.

Estimate $p(a)$ then marginalize over a ("*Mar.A*"). In this approach, we estimate $p(a)$ by assuming $p(a) \propto S(a|I)$. Given this prior over the global variable, we can obtain the following marginalized scoring function over a as $S(x|I) = \sum_a S(x, a|I)$. Hence, we can use the original recognition approach [12] to estimate the 3D location of each body joint independently.

The time and memory usage of different approaches are shown in Fig. 1. The estimation accuracy of all these approaches are reported in Sec. 6.

5.1. Temporally Stationary Global Variable

Instead of estimating sets of body joints and global variables at different frames independently, we jointly find sets of body joints $\mathcal{X}^{F*}$ and the global variable a^* that correspond to the maximum score in Eq. 6,

$$(\mathcal{X}^{F*}, a^*) = \underset{\mathcal{X}^F, a}{\operatorname{argmax}} S(\mathcal{X}^F, a|\mathcal{I}^F) \ . \quad (15)$$

Since frames are observed in a sequential order, the following efficient algorithm infers the body joint locations X^{f+1} and global variable a for the new frame $f+1$,

$$(X^{f+1*}, a^*) = \underset{X^{f+1}, a}{\operatorname{argmax}} (S(X^{f+1}, a) + S^{f*}(a)) \ . \quad (16)$$

where $S^{f*}(a)$ is the max-marginal score of the global variable a when the first f frames are observed, and it is defined recursively as

$$S^{f*}(a) = \max_{X^f} (S(X^f, a) + S^{f-1*}(a)) \ . \quad (17)$$

6. Experiments

We demonstrate the efficacy of our proposed models each conditioned on a global attribute: torso orientation or person height. Intuitively, given a specific orientation, possible space of body joint locations reduces due to the kinematic constraints. Similarly, person height captures pose variation caused by scale changes (*e.g.*the lengths of the limbs are very different for kids and adults). Following [29], we cast the body joint prediction problem as a detection problem, and average precision or average error (Err) (i.e.,1-AP) is reported. Notice that we use the same 0.1m tolerance criteria to calculate average precision or average error. Global variable estimation is treated as a classification problem, and mean accuracy (mAcc.) is reported. In all experiments, we compare our methods to the state-of-the-art method of [12], and the performances are evaluated on the MSRC dataset containing 5k synthetic depth images [29].

Implementation Details To enable a fair comparison of our models, our implementation uses the settings of [29, 12]: we use the depth comparison feature proposed in [29] as input to the random forest split function, our forest is trained using the same parameters for bagging and maximum tree depth, and we use the classification objective of [17] for training. We train the models using the 100k image training set from [29] and choose parameter values using their 5k image validation set. The values of hyper-parameters such as mean-shift bandwidth b^*, Gaussian kernel bandwidth b, and maximum distance threshold λ are also the same as the ones learned in [12].

6.1. Conditioning on Torso Orientation

We divide the torso orientation space ($360°$) into 32 non-overlapping partitions, each with $11.25°$. Since the torso orientations in the MSRC dataset are uniformly distributed between $-120°$ to $120°$, we train models conditioning on 23 unique states corresponding to torso orientations from the 5^{th} partition to the 27^{th} partition. In the following experiments, we demonstrate that our methods can reliably estimate the body joint locations and the states of the torso orientation.

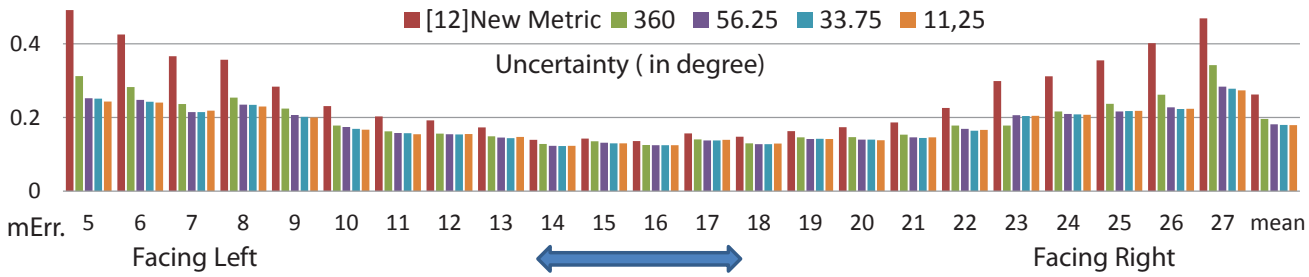


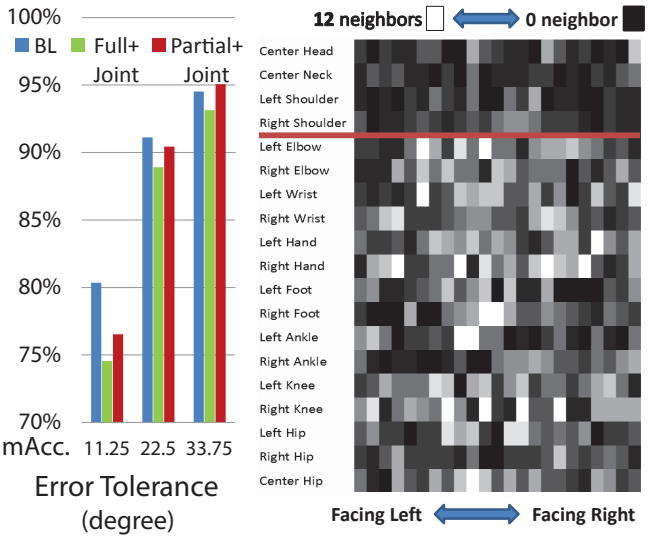
Figure 3: Mean body joint prediction errors (y-axis) for different torso orientations (x-axis) under different levels of torso orientation uncertainty.

Body Joint Prediction. We evaluate the *Full* and *Partial* conditional models (described in Eq. 2 and 3 respectively). To measure the generality of different approaches, we calculate the average error (Err) of body joints for every state of the global variable separately. We then obtain a mean error (mErr) across all body joints for each state. The final accuracy is summarized as the mean of the mErr across all states. This is slightly different from the metric used in [12], where AP of each body joint is calculated, and a mean AP across all body joints is reported as the overall accuracy. The body joint prediction accuracy comparison between [12], and our *Full* and *Partial* models is shown in Fig. 2. In the following, we use A+B to denote that B inference algorithm is applied on A model. *Joint* inference on both models achieves similar accuracy. In particular, *Partial* + *Joint* reduces the mean of mErr across all states of the global variable by up to 6.6% (which is a relative 25% reduction) compared to [12]. Moreover, we obtain 0.834 and 0.827 mAP for *Partial* + *Joint* and *Full* + *Joint*, respectively, using the metric in [12]. Both are much better than the 0.789 mAP obtained by [12]. Fig. 2 also shows that *MaxA* and *Mar.A* inferences on the *Partial* model produce inferior body part prediction accuracy compared to *Partial* + *Joint*. However, *Partial*+*MaxA* strikes a good balance between body joint prediction accuracy and efficiency.

Torso Orientation Estimation. As a baseline method (BL), we learn a classification forest (Eq. 14) using the tree structures in the *Partial* model. The comparison between BL, and *Full* + *Joint* and *Partial* + *Joint* is shown in Fig. 4a. Our methods achieve similar accuracy compared to the baseline method, while simultaneously predicting the body joint locations.

Model analysis. We now investigate the effect of the prior knowledge of the torso orientation on body joint prediction accuracy. In Fig. 3, we compare *Partial* + *Joint* with different levels of uncertainty of the torso orientation (*e.g.* set $p(a)$ to be uniform distributions across $360°, 56°, 33°, 11°$). When the uncertainty of the torso orientation reduces from $360°$ to $56°$, the body joint prediction error improves from 25% relative reduction to 30% relative reduction compared to [12]. In many real world applications such as gaming, a rough estimate of the torso orientation ($\sim 50°$) can be inferred from the context and can be used to significantly improve body joint prediction accuracy.

To understand the correlation between torso orientation and body joint location, we show the selected maximum



(a) Classification Accuracy (b) Model Selection

Figure 4: Panel (a) compares the mean accuracy (y-axis) between the baseline classification forest, *Full*+*Joint*, and *Partial*+*Joint*, when different errors in degree are considered as misclassified. Panel (b) shows the selected D_a in Eq. 9 (i.e., the number of neighbours for vote transfer) for different combinations of body joint (y-axis) and state of the torso orientation (x-axis). Notice that brightness-coded bins represent long range vote transfer is required (i.e.,larger D_a). Body joints above the red separation line are closely correlated with the torso orientation.

vote transfer distance (D_a) which produces the best prediction accuracy on the validation set for each unique pair of body joint and state of the torso orientation (Fig. 4b). Interestingly, the closer the body joint with respect to the torso, the smaller D_a is selected. This implies that there might be other global variables which can improve the body joint prediction accuracy for hands, foot, *etc*., more than the torso orientation.

6.2. Conditioning on Person Height

We divide the person height (ranging from about 0.5m to 2m) into 4 discrete partitions (*i.e.* ~ 0.4m each partition), and define models with 4 global states. The number of images in the MSRC dataset are distributed in different partitions with the proportion $(0.14 : 0.47 : 1 : 0.17)$. The number of training images in each partition affects the value of the scoring function. Therefore, before we select the parameters for vote transfer on the validation set, we first estimate a weight to be multiplied on the score of each state so that the scores are comparable (see more detail in our technical report [32]). Our learned weights $1.5433, 1.0230, 1, 2.8497$

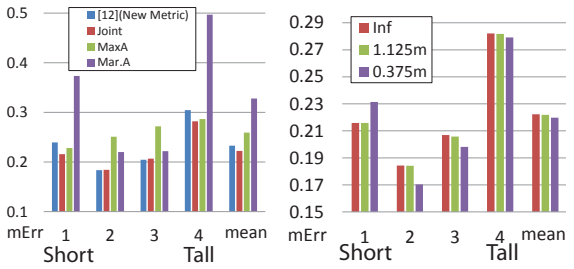
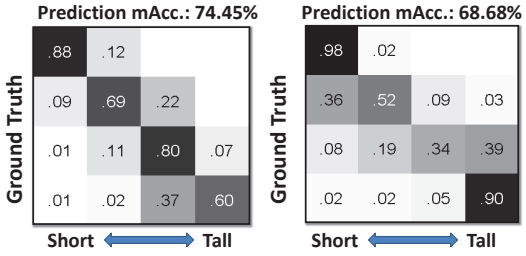


Figure 5: (Left) Mean body joint prediction error (y-axis) comparison between [12] and three recognition approaches (*e.g.Joint*, *MaxA*, and *Mar.A*) applied on our *Partial* model for different player heights (x-axis). (Right) Mean body joint prediction errors (y-axis) for different play heights (x-axis) under different level of player height uncertainty (color coded).



(a) *Partial* + *Joint* (b) BL

Figure 6: Person height estimation accuracy comparison between *Partial* + *Joint* (a) and the baseline classification forest (b).

are roughly proportional to the inverse of the size of each partition. We also observe that the scale of the relative body joint locations changes a lot when the person height is changed. However, the ratio between the body joint relative location and person height remains much more stable. Therefore, instead of transferring the absolute votes from neighbouring states, we first rescale the votes (see our technical report [32]). Other than these two additional steps, we follow the procedures described in Sec. 6.1 to train the conditional models.

Body joint prediction. Here we only evaluate the *Partial* model since it is compact and typically produces results similar to the full model. The body joint prediction comparison between [12] and different recognition approaches applied on the *Partial* model is shown in Fig. 5-Left. It can be seen that the *Joint* approach achieves the lowest error.

Person height estimation. The comparison between the classification forest Baseline (BL) and *Partial* + *Joint* is shown in Fig. 6. Our method achieves better accuracy (mAcc. 74.45%) compared to BL (mAcc. 68.68%), while simultaneously predicting body joint locations.

Model analysis. In Fig. 5-Right, we compare the accuracy of *Partial* + *Joint* with different levels of uncertainty of the player height. The accuracy improves when prior knowledge is available and the level of uncertainty is reduced.

Multiple frames. We now consider a new dataset which contains multiple frames for each human subject. Since person height is a stationary variable for each subject, our methods can use temporal consistency to obtain more accurate height estimates (Fig. 8b) and in turn body part locations (Fig. 8a). We observe that person height estimation accuracy improves significantly when more frames are given (except for the tallest person state (PH4)), and the body joint prediction error reduces for every state of the person height.

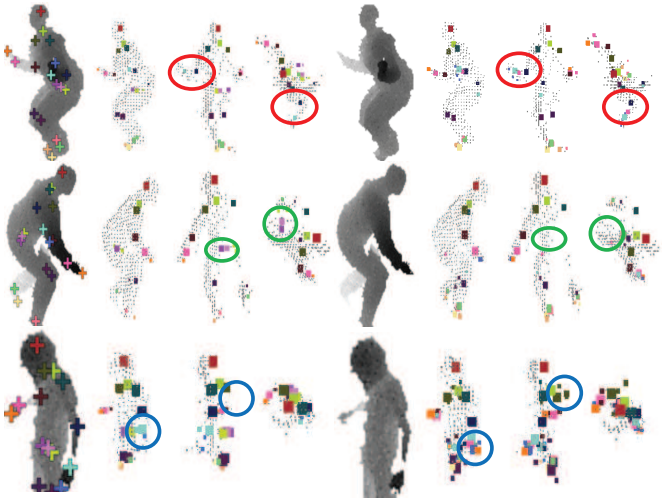
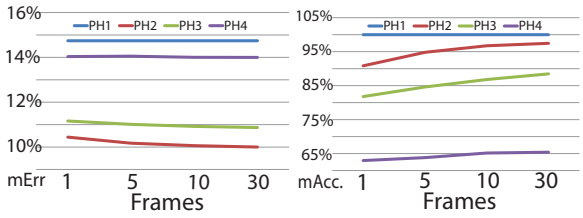


Figure 7: Predicted body joint locations. Each example shows an input depth image overlaid with color-coded ground truth body joint locations, and then inferred body joint locations from front, right, and top views. The size of the boxes indicates the inferred confidence. We compare the predictions of *Partial* + *Joint* (First-Column) with [12] (Last-Column). We observe that the method proposed in [12] confuses symmetric body joints in many images. For instance, left-wrist (pink) is confused with the right-wrist (light-blue) in the first example (highlighted in red circles). Our proposed model with a global variable encoding torso orientation does much better. Our method can also predict occluded body joints more accurately (e.g. the occluded hips are successfully predicted in the second row (highlighted in green circles)). When a variable encoding human height is used, we observe more accurate prediction especially for kids (e.g. our predictions in the last row is less noisy (highlighted in blue circles)).



(a) Body Joint Prediction Error (b) Person Height Estimation

Figure 8: Stationary player height across multiple frames (up to 30 frames). Panel (a) shows the mean body joints prediction errors (y-axis) when different number of frames (x-axis) are observed. Panel (b) shows the mean accuracies of predicted player heights (y-axis) when different number of frames (x-axis) are observed. In both panels, the accuracies of 4 states of player height are shown in curves with different color.

7. Conclusion and Future Work

We have presented a novel conditional regression model that significantly outperforms the state-of-the-art method for predicting human pose from depth images [12]. Our model incorporates dependency relationships between output variables through a global latent variable while still maintaining low computational cost. We show that the incorporation of a global latent variable encoding torso orientation, or human height, *etc*., can dramatically increase the accuracy of body joint location prediction. Further, it also allows efficient and seamless incorporation of prior knowledge about the problem instance. When there exists an effi-

cient way to estimate the global variable (*e.g*. torso orientation) with good accuracy, *Partial+MaxA* strikes a good balance between accuracy and efficiency. We also show how our model can efficiently utilize temporal consistency in the state of the global variable such as the height of the subject to improve performance. The method presented in this paper is general and can be applied to other problems where random forests are used.

Our results raise a number of interesting questions. How can one makes the model conditioned on continuous variables rather than having to discretize the domain of the global variables? Conditioning on different global variables implies different dependencies among locations of body parts. Can we infer the global variables that lead to the best accuracy for predicting body part locations? Similarly, for a general image labelling problem, like object segmentation, can we find the definition of the global latent variable that leads to the best performance? We believe all these questions are interesting directions for future work.

References

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *CVPR*, 2004. 2

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

[3] L. Bo and C. Sminchisescu. Structured output-associative regression. In *CVPR*, 2009. 2

[4] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3D prediction. In *CVPR*, 2008. 2

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002. 4

[6] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012. 2

[7] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 2

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 2

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 2

[10] J. Gall, A. Yao, N. Razavi, L. J. V. Gool, and V. S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 2011. 1

[11] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real time motion capture using a single time-of-flight camera. In *CVPR*, 2010. 2

[12] 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, 3, 4, 5, 6, 7

[13] D. Grest, J. Woetzel, and R. Koch. Nonlinear body pose estimation from depth images. In *DAGM*, 2009. 2

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

[15] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3D human body tracking with an articulated 3D body model. In *ICRA*, 2006. 2

[16] C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *ICCV*, 2009. 1

[17] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *CVPR*, 2005. 5

[18] Z. Lu, M. A. Carreira-Perpinan, and C. Sminchisescu. People tracking with laplacian eigenmaps latent variable models. In *NIPS*, 2009. 2

[19] R. Memisevic, L. Sigal, and D. J. Fleet. Shared kernel information embedding for discriminative inference. *PAMI*, 2012. 2

[20] Microsoft Corp. Redmond WA. Kinect for Xbox 360. 2

[21] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 2006. 2

[22] A. Montillo, J. Shotton, J. M. Winn, J. E. Iglesias, D. N. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *IPMI*, 2011. 1

[23] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV*, 2007. 2

[24] S. Nowonzin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. *ICCV*, 2011. 1

[25] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In *ICRA*, 2010. 2

[26] R. Poppe. Vision-based human motion analysis: An overview. *CVIU*, 2007. 2

[27] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, 2005. 2

[28] R.Wang and J. Popović. Real-time hand-tracking with a color glove. In *Proc. ACM SIGGRAPH*, 2002. 2

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011. 1, 2, 4, 5

[30] M. Siddiqui and G. Medioni. Human pose estimation from a single view point, real-time range sensor. In *CVCG at CVPR*, 2010. 2

[31] H. Sidenbladh, M. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV*, 2002. 2

[32] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. Technical report. http://www.eecs.umich.edu/~sunmin/. 6, 7

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