Metric Regression Forests for Human Pose Estimation

Gerard Pons-Moll[12]
http://www.tnt.uni-hannover.de/~pons/
Jonathan Taylor[13]
jtaylor@cs.toronto.edu
Jamie Shotton[1]
jamiesho@microsoft.com
Aaron Hertzmann[14]
hertzman@adobe.com
Andrew Fitzgibbon[1]
awf@microsoft.com

[1] Microsoft Research
Cambridge, UK
[2] TNT
Leibniz University of Hannover,
Germany
[3] University of Toronto
[4] Adobe Research

Abstract

We present a new method for inferring dense data to model correspondences, focusing on the application of human pose estimation from depth images. Recent work proposed the use of regression forests to quickly predict correspondences between depth pixels and points on a 3D human mesh model. That work, however, used a proxy forest training objective based on the classification of depth pixels to body parts. In contrast, we introduce Metric Space Information Gain (MSIG), a new decision forest training objective designed to directly optimize the entropy of distributions in a metric space. When applied to a model surface, viewed as a metric space defined by geodesic distances, MSIG aims to minimize image-to-model correspondence uncertainty. A naïve implementation of MSIG would scale quadratically with the number of training examples. As this is intractable for large datasets, we propose a method to compute MSIG in linear time. Our method is a principled generalization of the proxy classification objective, and does not require an extrinsic isometric embedding of the model surface in Euclidean space. Our experiments demonstrate that this leads to correspondences that are considerably more accurate than state of the art, using far fewer training images.

1 Introduction

A key concern in a number of computer vision problems is how to establish correspondences between image features and points on a model. An effective method is to use a decision forest to discriminatively regress these correspondences [14, 24, 28]. So far, these approaches have ignored the correlation of model points during training, or have arbitrarily pooled the model points into large regions (parts) to allow the use of a classification training objective. In this work, we propose the Metric Space Information Gain (MSIG) training objective for decision forests, that, instead, naturally accounts for target dependencies during training and does not require the use of artificial parts. Our MSIG objective assumes that the model points lie

© 2013. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

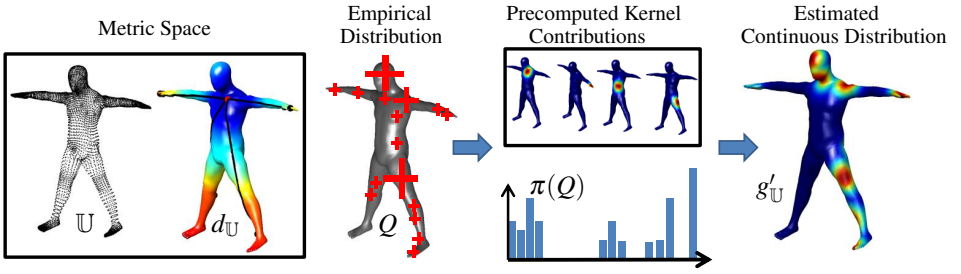

Figure 1: We propose a method to quickly estimate the continuous distributions on the manifold or more generally the metric space induced by the surface model. This allows us to efficiently train a random forest to predict image to model correspondences using a continuous entropy objective. Notation is explained in Sec. 3.

in a space in which a metric has been defined to encode correlation between target points. Among the larger class of problems where MSIG could apply, we focus on the challenging application of general activity human pose estimation from single depth images.

Human pose estimation has been a very active area of research for the last two decades. Algorithms can be classified into two main groups, namely generative [20] and discriminative [26]. Generative approaches model the likelihood of the observations given a pose estimate. The pose is typically inferred using local optimization [4, 5, 13, 22, 27] or stochastic search [8, 10, 21]. Regardless of the optimization scheme used, such approaches are susceptible to local minima and thus require good initial pose estimates.

Discriminative approaches [3, 15, 17, 29] learn a direct mapping from image features to pose space from training data. Unfortunately, these approaches can struggle to generalize to poses not present in the training data. The approaches in [14, 23] bypass some of these limitations by discriminatively making predictions at the pixel level. This makes it considerably easier to represent the possible variation in the training data, but yields a set of independent local pose cues that are unlikely to respect kinematic constraints. To overcome this, recent work has fit a generative model to these cues [1, 12, 28]. The most relevant example of such a hybrid system is that of Taylor *et al.* [28] who robustly fit a mesh model to a set of image-to-model correspondence predicted by a decision forest.

Decision forests are a classic method for inductive inference that has recently regained popularity by yielding excellent results on a wide range of classification and regression tasks. The canonical example in pose estimation is [23] where a forest is used to segment the human body into parts. These parts are manually specified and the segmentation is used to define a per-pixel classification task. To train the forest, split functions are evaluated using a parts objective ('PARTS') based on discrete information gain. Specifically, the split is chosen to reduce the Shannon entropy of the resulting body part class distributions at the left and right child nodes. Motivated by the success of Hough forests [11] for object detection and localization, a follow-up paper [14] directly regressed at each pixel an offset to several joint locations. They showed, surprisingly, that retrofitting a forest for this task that had been trained using the PARTS objective [23] outperforms forests that had been trained using a standard regression objective based on variance minimization. The work of Taylor *et al.* [28] followed suit in retrofitting a PARTS trained classification forest to predict model-image correspondences. Despite these successes, the somewhat arbitrary choice to bootstrap using a PARTS objective, clashes with the experience of several authors [6, 16, 18] who show that

the objective function has a substantial influence on the generalization error of the forest.

We address this by showing that the image-to-model correspondences used in Taylor *et al*. [28], can be predicted with substantially higher accuracy by training a forest using the 'correct' objective – an objective that chooses splits in order to minimize the uncertainty in the desired predictive distributions. When the target outputs lie in a metric space, minimizing the continuous entropy in that space is the natural training objective to reduce this uncertainty. Our main contribution is in how this continuous entropy can be computed efficiently at every split function considered in the training procedure, even when using millions of training examples. To this end, we estimate the split distributions using Kernel Density Estimation (KDE) [19] employing kernels that are functions of the underlying metric. To make this computationally tractable, we first finely discretize the output space and pre-compute a kernel matrix encoding each point's kernel contribution to each other point. This matrix can then be used to efficiently 'upgrade' any empirical distribution over this space to a KDE approximation of the true distribution. Although staple choices exist for the kernel function (*e.g*. Gaussian), its underlying metric (*e.g*. Euclidean distance) and discretization (*e.g*. uniform), they can also be chosen to reflect the application domain. In our domain of human pose estimation, the targets are points on a 3D mesh model surface. Interestingly, our MSIG objective can encode the body part classification objective [23] by employing a non-uniform discretization. It is, however, much more natural to have a near uniform discretization over the manifold and to use the geodesic distance metric to encode target correlation on this manifold, see Fig. 1. As articulated shape deformations are $\varepsilon-$isometric with respect to the geodesic distance, all computations in this space are independent of pose which removes the need to find an extrinsic isometric embedding in the Euclidean space as used in [28].

Our experiments on the task of human pose estimation show a substantial improvement in the quality of inferred correspondences from forests trained with our objective. Notably, this is achieved with no additional computational burden since the algorithm remains the same at test time. We further observe that with orders of magnitude less training data, we can obtain state of the art human pose performance using the same fitting procedure as [28].

2 Forest Training

We employ the standard decision forest training algorithm and features. A forest is an ensemble of randomly trained decision trees. Each decision tree consists of split nodes and leaf nodes. Each split node stores a split function to be applied to incoming data. At test time, a new input will traverse the tree branching left or right according to the test function until a leaf node is reached. Each leaf stores a predictor, computed from the training data falling into that leaf. At training time, each split candidate partitions the set of training examples Q into left and right subsets. Each split function s is chosen among a pool $\mathcal{F}$ in order to reduce the average uncertainty of the predictions. This is achieved using a training objective $I(s)$ that assigns a high score if s reduces this uncertainty. Training proceeds greedily down the tree, locally optimizing I for each node, until some stopping criterion is met. In all of our experiments, we use the same binary split functions as [23] which consist of fast depth comparisons executed on a window centered at the input depth pixel $\mathbf{x}_i$. For more details, we refer the reader to [7].

As our main contribution, we propose Metric Space Information Gain (MSIG) as the natural objective to learn to regress image-to-model correspondences where the target domain is a metric space. This objective aims to reduce the continuous entropy of the data on the

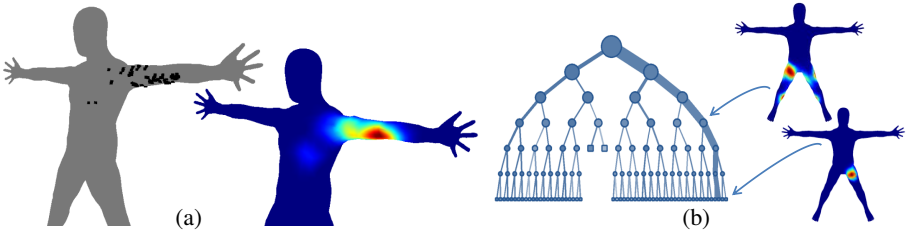

Figure 2: (a) On the left we show an example of an empirical distribution and on the right our estimated continuous distribution. (b) Examples of the continuous distributions induced by KDE at different levels of the tree. The MSIG objective reduces the entropy of the distributions through each split resulting in increasingly uni-modal and lower entropy distributions deeper in the tree.

metric space. In the case of a metric space induced by a reference 3D human mesh model with standard body proportions, this translates into the correspondence uncertainty over the model surface. To train a forest using MSIG we first need to define the metric for the target space which determines the correlation between the targets. Instead of assuming a uni-modal Gaussian distribution (*e.g.* [24]) we use KDE to approximate the density where the kernels are functions of the metric chosen; see Fig. 2. Informally, distributions with probability mass at nearby locations will result in lower entropies than distributions with probability mass spread to distant locations. As we will show, MSIG outperforms the PARTS [23, 28] and standard regression [14] objectives, and can be computed efficiently in linear time.

3 Metric Space Information Gain

We use the surface of a canonical human body to define the metric space $(\mathbb{U}, d_{\mathbb{U}})$ of our targets. Here, $\mathbb{U}$ denotes the continuous space of locations on this model and $d_{\mathbb{U}}$ denotes the geodesic distance metric on the manifold induced by the surface model. Let U denote a random variable with probability density p_U whose support is a set $\mathbb{U}$ and let $B(s)$ be a random variable that depends on a split function s and takes the values L or R. The natural objective function used to evaluate whether a split s reduces uncertainty in this space is the information gain,

$$I(s) = H(U) - \sum_{i \in \{L,R\}} P(B(s) = i) H(U|B(s) = i) \tag{1}$$

where $H(U)$ is the differential entropy of the random variable U. For a random variable U with distribution p_U this is defined as

$$H(U) = \mathbb{E}_{p_U(\mathbf{u})}[-\log p_U(\mathbf{u})] = -\int_{\mathbb{U}} p_U(\mathbf{u}) \log p_U(\mathbf{u}) d\mathbf{u}. \tag{2}$$

In practice the information gain can be approximated using an empirical distribution $Q = \{\mathbf{u}_i\}$ drawn from p_U as

$$I(s) \approx \hat{I}(s;Q) = \hat{H}(Q) - \sum_{i \in \{L,R\}} \frac{|Q_i|}{|Q|} \hat{H}(Q_i), \tag{3}$$

where $\hat{H}(Q)$ is some approximation to the differential entropy and $\|\cdot\|$ dennotes the cardinality of a set. One way to approach this is to use a Monte Carlo approximation of Eq. 2

$$H(U) \approx -\frac{1}{N}\sum_{\mathbf{u}_i \in Q} \log p_U(\mathbf{u}_i)\ . \tag{4}$$

As the continuous distribution p_U is unknown, it must also be estimated from the empirical distribution Q. One way to approximate this density $p_U(\mathbf{u})$ is using Kernel Density Estimation (KDE). Let $N = |Q|$ be the number of datapoints in the sample set. The approximated density $f_U(\mathbf{u})$ is then given by

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u}) = \frac{1}{N}\sum_{\mathbf{u}_j \in Q} k(\mathbf{u};\mathbf{u}_j), \tag{5}$$

where $k(\mathbf{u};\mathbf{u}_j)$ is a kernel function centered at $\mathbf{u}_j$. Plugging this approximation into Eq. 4, we arrive at the KDE estimate of entropy:

$$\hat{H}_{\mathrm{KDE}}(Q) = -\frac{1}{N}\sum_{\mathbf{u}_i \in Q} \log\left(\frac{1}{N}\sum_{\mathbf{u}_j \in Q} k(\mathbf{u}_i;\mathbf{u}_j)\right). \tag{6}$$

That is, one evaluates the integral at the datapoint locations $\mathbf{u}_i \in Q$ in the empirical distribution, a calculation of complexity N^2. To train a tree, the entropy has to be evaluated at every node of the tree and for every split function $s \in \mathcal{F}$. Thus this calculation could be performed up to $2^L \times |\mathcal{F}|$ times, where L is the maximum depth of the tree. Clearly, for big training datasets one cannot afford to scale quadratically with the number of samples. For example, the tree structures used in this paper are trained from 5000 images with roughly 2000 foreground pixels per image, resulting in 10 million training examples. Therefore, as our main contribution, we next show how to train a random forest with a MSIG objective that scales linearly with the number of training examples.

To this end, we discretize the continuous space into V points $\mathbb{U}' = (\mathbf{u}'_1, \mathbf{u}'_2 \ldots, \mathbf{u}'_V) \subseteq \mathbb{U}$. This discretization simplifies the metric to a matrix of distances $D_{\mathbb{U}} = \left(d_{\mathbb{U}}(\mathbf{u}'_i, \mathbf{u}'_j)\right)$ that can be precomputed and cached. Even better, the kernel functions can be cached for all pairs of points $(\mathbf{u}'_i, \mathbf{u}'_j) \in \mathbb{U}'$. For our experiments, we choose the kernel function on this space to be an exponential $k(\mathbf{u}'_i;\mathbf{u}'_j) = \frac{1}{Z}\exp\left(-\frac{d_{\mathbb{U}}(\mathbf{u}'_i,\mathbf{u}'_j)^2}{2\sigma^2}\right)$ where $d_{\mathbb{U}}(\mathbf{u}'_i, \mathbf{u}'_j)$ is the geodesic distance on the model and σ is the bandwidth of the kernel. The normalization constant Z ensures that the total amount of contribution coming from each point equals one and is thus invariant to the discretization. The geodesic distances are pre-computed on a high resolution triangulated mesh model using the Dikjstra algorithm [9] The discretization would ideally be uniformly distributed over the model surface, but we find that that simply using an appropriate sampling of the vertex locations of the original mesh sufficient to obtain good results.

In all the experiments shown in this paper we use $\sigma = 3$cm which roughly corresponds to the average nearest neighbor distance in the empirical distributions. A detailed discussion on kernel bandwidth selection can be found in [25]. Since the kernels fall off to zero, only a small subset of indices $\mathcal{N}_i \subseteq \{1., ..., V\}$ indicate neighboring points $\{\mathbf{u}'_j\}_{j\in\mathcal{N}_i}$ that contribute to $\mathbf{u}'_i$. Hence, for efficiency, we only store the significant kernel contributions for each discretized point $\mathbf{u}'_i$.

For ease of explanation in the following, we assume here that each point has a constant number of neighbors $|\mathcal{N}_i| = M$ for all $i \in \{1, ..., V\}$. Let $\mathcal{J}_{i,j}$ denote a look-up table that

contains the node index of the j-th neighbor of the i-th node. This leads to the following kernel matrix that is pre-computed before training:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{u}'_1;\mathbf{u}'_{\mathcal{J}_{1,1}}) & k(\mathbf{u}'_1;\mathbf{u}'_{\mathcal{J}_{1,2}}) & \dots & k(\mathbf{u}'_1;\mathbf{u}'_{\mathcal{J}_{1,M}}) \\ k(\mathbf{u}'_2;\mathbf{u}'_{\mathcal{J}_{2,M}}) & k(\mathbf{u}'_2;\mathbf{u}'_{\mathcal{J}_{2,2}}) & \dots & k(\mathbf{u}'_2;\mathbf{u}'_{\mathcal{J}_{2,M}}) \\ \vdots & & \ddots & \vdots \\ k(\mathbf{u}'_V;\mathbf{u}'_{\mathcal{J}_{V,1}}) & k(\mathbf{u}'_V;\mathbf{u}'_{\mathcal{J}_{V,M}}) & \dots & k(\mathbf{u}'_V;\mathbf{u}'_{\mathcal{J}_{V,M}}) \end{bmatrix}. \tag{7}$$

Thus, given a discretization $\mathbb{U}'$ we can smooth the empirical distribution over this discretization using the kernel contributions as

$$g_{U'}(\mathbf{u}'_i;Q) \simeq \frac{1}{N}\sum_{j\in\mathcal{N}_i} \pi_j(Q)k(\mathbf{u}'_i;\mathbf{u}'_j) \tag{8}$$

where the weights $\pi_j(Q)$ are the number of data points in the set Q that are mapped to the bin center $\mathbf{u}'_j$. In other words, $\{\pi_j(Q)\}_{j=1}^V$ are the unnormalized histogram counts of the discretization given by $\mathbb{U}'$. In this way, we can use a simple histogram as our sufficient statistic to estimate the density. The expression in Eq. 8 can be efficiently computed using the precomputed kernel matrix $\mathbf{K}$ in Eq. 7

$$g_{U'}(\mathbf{u}'_i;Q) = \frac{1}{N}\sum_{m=1}^{M} \pi_{\mathcal{J}_{i,m}}(Q)\mathbf{K}_{i,m}\ . \tag{9}$$

We can use this to further approximate the continuous KDE entropy estimate of the underlying density in Eq. 5 as

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u};Q) \simeq g_{U'}(\alpha(\mathbf{u});Q) \tag{10}$$

where $\alpha(\mathbf{u})$ maps $\mathbf{u}$ to a point in our discretization. Using this, we approximate the differential entropy of $p_U(\mathbf{u})$ using the discrete entropy of $g_{U'}$ defined on our discretization. Hence, our MSIG estimate of the entropy on the metric space for an empirical sample Q is

$$\hat{H}_{\text{MSIG}}(Q) = -\sum_{u_i\in\mathbb{U}'} g_{U'}(\mathbf{u}'_i;Q)\log g_{U'}(\mathbf{u}'_i;Q) \tag{11}$$

where the terms only need to be calculated when $g_{U'}(\mathbf{u}'_i;Q) \neq 0$.

Note that this is also equivalent to approximating the entropy defined in Eq. 2 by evaluating the integral only at the V points of the discretized space $\mathbb{U}'$. Note that in contrast to Eq. 4 we need to re-weight by $g_{U'}(\mathbf{u}'_i;Q)$ because we are sampling uniformly on a grid of points in the space as opposed to Eq. 4 where the samples are drawn from the empirical distribution Q. This is equivalent to importance sampling with a uniform proposal distribution.

The complexity of Eq. 11 is $V \times M$. When training a tree, each new split s requires a linear pass through the data to compute the left and right histograms. The total complexity of evaluating a split using Eq. 3 is thus $N+V\times M \ll N^2$ allowing trees to be trained efficiently.

Finally, to compute a correspondence using a forest trained with MSIG, we follow the lead of [28] in outfitting each leaf with a regression model.[1] Briefly, the training data that falls into a leaf defines an empirical distribution over the space and we use mean shift to find the most prominent mode $\hat{\mathbf{u}}$ and its mean-shift weight ω which we bundle as the regression model $(\hat{\mathbf{u}},\omega)$. At test time, each tree yields a separate regression model, and we predict the correspondence $\hat{\mathbf{u}}$ that yields the maximum weight ω.

[1] Naturally, we could also consider the maximum of the KDE fit but we opt to follow the strategy of [28] to reflect the improvement due to the forest structure itself in our results.

4 Pose Estimation

We now investigate the ability of MSIG trained forests to improve the accuracy of human-pose estimation. Hence, we follow the procedure of Taylor *et al*. [28] as closely as possible. Their procedure uses a 3D mesh model to explain a set of input depth pixels $\{\mathbf{x}_i\}_{i=1}^N \subseteq \mathbb{R}^3$. A vector of parameters $\theta \in \mathbb{R}^d$ representing a particular pose defines a global transform M that takes any point $\mathbf{u} \in \mathbb{U}$ on the model surface and outputs its 3D location $M(\mathbf{u};\theta) \in \mathbb{R}^3$. The goal is then to obtain a set of model correspondences $C = \{\mathbf{u}_i\}_{i=1}^N \subseteq \mathbb{U}$ and a corresponding pose θ so that the transformed correspondences $\{M(\mathbf{u}_i;\theta)\}_{i=1}^N \subseteq \mathbb{R}^3$ align with their respective input pixels. In practice, insufficient model capacity, noisy correspondences and local minima encountered during optimization make this goal nearly impossible to achieve. Therefore, a carefully designed energy function is used to robustly evaluate the quality of a pose θ and noisy correspondences C. This is given by:

$$E(\theta, C) = \lambda_{\text{vis}} E_{\text{vis}}(\theta, C) + \lambda_{\text{prior}} E_{\text{prior}}(\theta) + \lambda_{\text{int}} E_{\text{int}}(\theta) \tag{12}$$

where the first term E_{vis} accounts for both outliers and model visibility, E_{prior} encodes a Gaussian prior on pose learnt from motion capture data and the E_{int} penalizes self intersections of the model. We refer the reader to [28] for more details on these terms. Standard iterated closest point (ICP) approaches to minimizing an objective such as Eq. 12 alternate between optimizing C and θ, but convergence is unlikely without a good initial guess of one or the other. A key contribution of [28] was to demonstrate that the set of correspondences C could be initialized by their random forest and the pose θ effectively optimized in 'one shot'. Given this initial set of correspondences, a non-linear optimization of Eq. 12 with respect to pose θ is performed using a Quasi-Newton method (L-BFGS).

In contrast to [28], we also *consider* a further ICP optimization to achieve additional gains. Holding θ fixed, we update C by finding the closest visible model point to each depth pixel, instead of minimizing Eq. 12 exactly. This allows C to be updated efficiently using a k-D tree [2]. To update θ, the non-linear optimizer is restarted with the new correspondences.

5 Experiments

We evaluate our approach using the same test set of 5000 synthetic depth images as used in [28]. We examine both the accuracy of the inferred correspondences and their usefulness for single frame human pose estimation from depth images.

5.1 Setup

Forests. We use two forests in our experiments: MSIG and PARTS, indicating respectively that they were trained with our proposed MSIG objective or the standard PARTS based objective of [23]. Both forests contain three trees and were trained to depth 20. To learn the structure and split functions of each tree we use 5000 synthetic images per tree. The extra complexity in training a MSIG tree resulted in them taking roughly three times as long as the PARTS trees. This complexity does not exist at test time and thus speeds reported in [28] are obtainable using either type of tree. To populate the leaf distributions in both types of trees, we replicate the strategy of [28]: we push the training data from 20000 (depth, correspondences) image pairs through the trees and find the mode of the distribution in the extrinsic isometric embedding of a human shape (the 'Vitruvian' pose) using mean-shift.

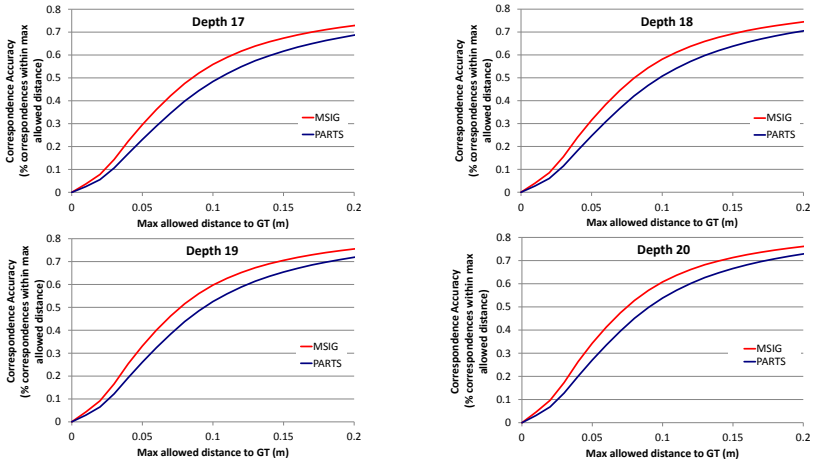

Figure 3: Correspondence error comparison of PARTS forest with the proposed MSIG forest. We evaluate the accuracy for forests of depths $17, 18, 19, 20$. It can be observed that our proposed method consistently produces considerably more accurate correspondences.

Pose estimation. For human pose estimation we parametrize a model using a skeleton. We predict the following 19 body joints: head, neck, shoulders, elbows, wrists, hands, knees, ankles, feet, and hips (left, right, center).

Metrics. To evaluate the accuracy of the inferred correspondences, we use the *correspondence error* defined as the geodesic distance between the prediction and the ground truth model location. We use a model with standard proportions and thus a correspondence error of 25 cm is roughly the length of the lower arm. To measure pose accuracy we use the challenging *worst joint* error metric introduced in [28]: the proportion of test scenes that have all predicted joints within a certain Euclidean distance from their ground truth locations.

5.2 Results

We evaluate the performance of our forest regressors to predict dense image to model correspondences. We quantify the proportion of predicted correspondences with an error less than a certain distance. We find that correspondences with an error of less than 15 cm tend to be useful for pose estimation whereas those with higher errors are usually treated as outliers. In Fig. 3 we show the correspondence accuracy for both the MSIG forest and PARTS forest at depths of $17, 18, 19$ and 20. As it can be seen, the MSIG forest produces correspondences that are consistently more accurate than those produced from the PARTS forest. This is very encouraging since forests trained using a PARTS objective had previously shown state of the art performance, far superior to those using other objectives such as the Hough-regression [14]. We attribute the better performance of our approach to the fact that MSIG favors distributions with mass concentrated (in the sense of the defined metric) in close locations.

Although the inferred dense correspondences can be used for a large number of tasks, we consider the task of single frame pose estimation as a motivational example. Therefore, we also show the impact in the pose accuracy again for forests of depth $17, 18, 19$ and 20. As one would expect, better correspondences translate into more accurate pose estimates. As can be seen in Fig. 4, the MSIG forest produces a small but significant improvement w.r.t.

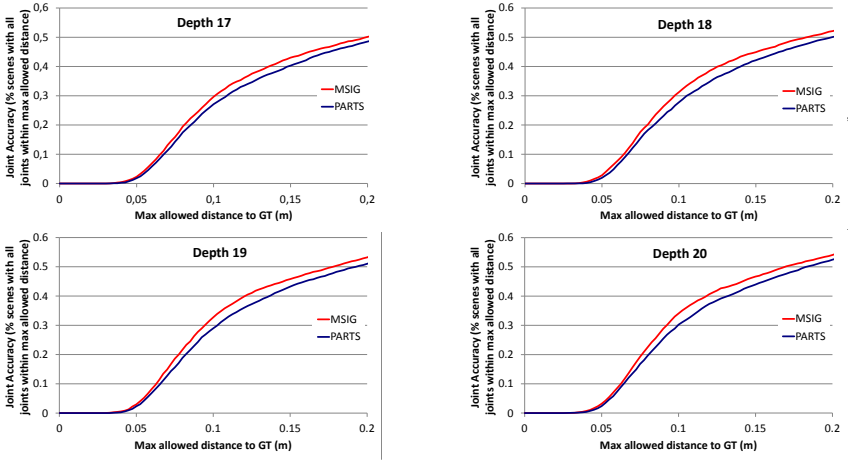

Figure 4: Pose accuracy comparison using correspondences from both PARTS and proposed MSIG forests at depths 17, 18, 19 and 20. For both forests, we use the pose estimation algorithm of [28] as explained in Sec. 4 and evaluate using the worst joint error metric.

to the PARTS forest. The smaller gains in pose accuracy are expected as the energy of [28] is designed to be robust to outliers from their forest. We also compare in Fig. 5 directly to the results provided by [28], which appears to be the state of the art for single frame pose estimation from depth images. Despite our MSIG forest using orders of magnitude less training images (300K images vs. 5K images per tree), we achieve equivalent performance.

We further demonstrate that our correspondences can be used to initialize classical registration methods such as articulated ICP as explained in Sec 4. Contrary to what was alluded to in [28] we find that using just 10 such ICP alternations provides an additional performance gain of up to 10% with both PARTS and MSIG correspondences as demonstrated in Fig. 5. Furthermore, it can be seen that the gap between the MSIG and PARTS is not washed out by this downstream ICP processing. The resulting MSIG poses after ICP refinement, thus represent the state of the art on this dataset.

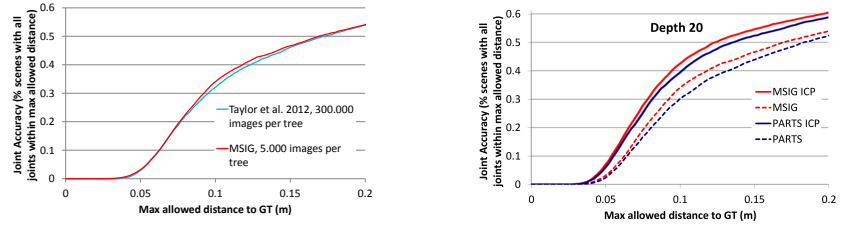

Figure 5: Left: Pose accuracy of our MSIG forest trained with 5000 images per tree compared to accuracy reported by [28] which used 300,000 training images. Right: Pose accuracy for both PARTS and MSIG forests after 10 iterations of ICP. Note that the curve labelled MSIG in both the left (solid red) and right (dashed red) plots are the same.

6 Conclusion

We have introduced MSIG, an objective function that evaluates a split function's ability to reduce the uncertainty over an arbitrary metric space using kernel density estimation. Using a discretization of this space, an efficient approximation to MSIG was developed as to facilitate its use in training random forests. Although the general framework can be tuned through the specification of an appropriate metric space, kernel function and discretization, natural choices exist making this approach widely applicable.

We employed MSIG in the context of human pose estimation to both simplify and enhance the inference of dense data to model correspondences by avoiding two arbitrary requisites of previous work: (i) our work does not require a segmentation of the human body into parts, and (ii) it does not require an extrinsic isometric embedding of the human shape. A number of experiments show that the more principled MSIG objective allows the inference of superior correspondences compared to those provided by standard training objectives. Additionally, these results translate into state of the art accuracy for single frame pose estimation using far fewer training images.

References

[1] A. Baak, M. Müller, G. Bharaj, H. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV*, pages 1092–1099. IEEE, November 2011.

[2] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, September 1975.

[3] L. Bo and C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *IJCV*, 87:28–52, 2010.

[4] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *IJCV*, 56(3):179–194, 2004.

[5] M.A. Brubaker, D.J. Fleet, and A. Hertzmann. Physics-based person tracking using the anthropomorphic walker. In *IJCV*, 2010.

[6] W. Buntine and T. Niblett. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8(1):75–85, 1992.

[7] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2): 185–205, 2005.

[9] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[10] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *IJCV*, 87:75–92, 2010.

[11] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 33(11):2188–2202, 2011.

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

[13] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real-time human pose tracking from range data. In *ECCV*, 2012.

[14] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, pages 415–422. IEEE, 2011.

[15] C.S. Lee and A. Elgammal. Coupled visual and kinematic manifold models for tracking. *IJCV*, 2010.

[16] W.Z. Liu and A.P. White. The importance of attribute selection measures in decision tree induction. *Machine Learning*, 15(1):25–41, 1994.

[17] R. Memisevic, L. Sigal, and D. J. Fleet. Shared kernel information embedding for discriminative inference. *PAMI*, 34(4):778–790, April 2012.

[18] S. Nowozin. Improved information gain estimates for decision tree induction. In *ICML*, 2012.

[19] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[20] G. Pons-Moll and B. Rosenhahn. Model-based pose estimation. *Visual Analysis of Humans*, pages 139–170, 2011.

[21] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV*, nov 2011.

[22] G. Pons-Moll, L. Leal-Taixé, Truong T., and B. Rosenhahn. Efficient and robust shape matching for model based human motion capture. In *DAGM*, 2011.

[23] 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 single depth images. In *CVPR*, pages 1297–1304. IEEE, 2011.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013.

[25] B.W. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

[26] C. Sminchisescu, L. Bo, C. Ionescu, and A. Kanaujia. Feature-based pose estimation. *Visual Analysis of Humans*, pages 225–251, 2011.

[27] C. Stoll, N. Hasler, J. Gall, H. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, pages 951–958. IEEE, 2011.

[28] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *CVPR*, 2012.

[29] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *CVPR*, pages 1–8. IEEE, 2008.