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

Traditionally, human pose estimation algorithms could be classified into generative [2] and discriminative [4] approaches. Generative approaches model the likelihood of the observations given a pose estimate, however, they are susceptible to local minima and thus require good initial pose estimates. Discriminative approaches learn a direct mapping from image features to pose space from training data, however, they struggle to generalize to unseen poses. Building on previous work [3], Taylor *et al.* [5] bypass some of these limitations using a hybrid-approach that discriminatively predicts, for each pixel in a depth image, a corresponding point on the surface of a humanoid mesh model. This mesh model is then robustly fit to the resulting set of correspondences using local optimization. Surprisingly though, these correspondences are actually inferred using a random forest whose structure was trained using a *classification* objective that arbitrarily equates target model points belonging to the same predefined body part [3].

In this paper, we address Taylor *et al.*'s use of this proxy classification objective by proposing Metric Space Information Gain (MSIG), a replacement objective function for training a random forest to directly minimize the uncertainty over the target model points, naturally encoding the correlation between these points as a function of the geodesic distance. To this end, we view the surface of the model $\mathbb{U}$ as a metric space $(\mathbb{U}, d_{\mathbb{U}})$ defined by the geodesic distance metric $d_{\mathbb{U}}$ (see first panel of Figure 1). The natural objective function to minimize the uncertainty in the resulting true distributions that result from a split function s in such a space, is the information gain $I(s)$ [1]. This is generally approximated using an empirical distribution $Q = \{\mathbf{u}_i\} \subseteq \mathbb{U}$ drawn from the true unsplit distribution p_U as

$$I(s) \approx \hat{I}(s;Q) = \hat{H}(Q) - \sum_{i\in\{L,R\}} \frac{|Q_i|}{|Q|}\hat{H}(Q_i), \tag{1}$$

where Q_L and Q_R are the two resulting empirical distributions from applying s, and $\hat{H}(Q)$ is some approximation to the differential entropy

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

of the distribution p_U on $\mathbb{U}$ from which Q arose. We provide this approximation by first estimating the true continuous distribution $p_U(\mathbf{u})$ 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

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u}) = \frac{1}{N}\sum_{\mathbf{u}_j\in Q} k(\mathbf{u};\mathbf{u}_j), \tag{3}$$

where $k(\mathbf{u};\mathbf{u}_j)$ is a kernel function centered at $\mathbf{u}_j$. Unfortunately, the obvious way to estimate (2) using Monte Carlo is quadratic in N (see full paper) and thus a key contribution of this work is to demonstrate how to efficiently estimate it in linear time.

To this end, we discretize the space as $\mathbb{U}' = (\mathbf{u}'_1, \mathbf{u}'_2 \ldots, \mathbf{u}'_V) \subseteq \mathbb{U}$. The main advantage is that the discrete metric simplifies 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 beforehand. 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 $d_{\mathbb{U}}\left(\mathbf{u}'_i, \mathbf{u}'_j\right)$ is the geodesic distance on the model and σ is the bandwidth of the kernel.

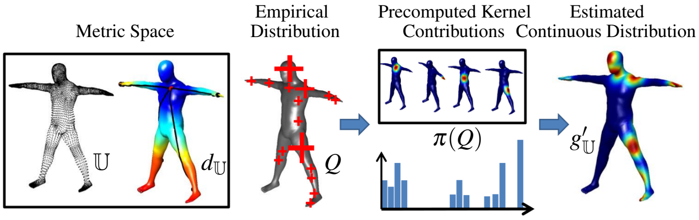

Figure 1: We propose a method to quickly estimate the continuous distributions on the manifold or more generally the metric space induced by the surface model. This allows us to efficiently train a random forest to predict image to model correspondences using a continuous entropy objective.

Using our discretization $\mathbb{U}'$ we then smooth the empirical distribution provided by Q over this discretization using the pre-computed kernel contributions as

$$g_{U'}(\mathbf{u}'_i;Q) \simeq \frac{1}{N}\sum_{j\in\mathcal{N}_i}\pi_j(Q)k(\mathbf{u}'_i;\mathbf{u}'_j) \tag{4}$$

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 $\mathbb{U}'$. We can use this to further approximate the continuous KDE entropy estimate of the underlying density in Eq. 3 as

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u};Q) \simeq g_{U'}(\alpha(\mathbf{u});Q) \tag{5}$$

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

$$\hat{H}_{\text{MSIG}}(Q) = -\sum_{u_i\in\mathbb{U}'} g_{U'}(\mathbf{u}'_i;Q)\log g_{U'}(\mathbf{u}'_i;Q)\ . \tag{6}$$

Only the calculation of the histogram counts scales with the number of training examples and thus, the complexity of calculating (6) is linear.

We find that forests trained using our MSIG objective function can provide substantially better correspondences in comparison to the forests trained using the objective from [5]. These improved correspondences translates into modest improvements in pose estimation that allows us to achieve state of the art pose estimation results with orders of magnitude less training data.

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

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

[3] 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.

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

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