Accurate Static Pose Estimation Combining Direct Regression and Geodesic Extrema

Brian Holt, Eng-Jon Ong and Richard Bowden

Abstract— **Human pose estimation in static images has received significant attention recently but the problem remains challenging. Using data acquired from a consumer depth sensor, our method combines a direct regression approach for the estimation of rigid body parts with the extraction of geodesic extrema to find extremities. We show how these approaches are complementary and present a novel approach to combine the results resulting in an improvement over the state-of-the-art. We report and compare our results a new dataset of aligned RGB-D pose sequences which we release as a benchmark for further evaluation.**

I. INTRODUCTION

Automatic human pose estimation remains an active area of research in computer vision. A fast, robust and accurate solution to this challenging problem would have wide ranging impact in markerless motion capture systems, human computer interaction and other applications of the visual analysis of humans. While many solutions have been proposed, the problem remains difficult because of the highly deformable nature of the human body. Compounding the problem is the large anthropometric variability in the population, variable image capture conditions, complex background, clothing, camera viewpoint and occlusion of body parts (including self-occlusion).

The availability of depth information from real-time depth cameras has simplified the task of pose estimation [28], [10], [22], [13], [11], [20] over traditional image capture devices by supporting high accuracy background subtraction, working in low-illumination environments, being invariant to colour and texture, providing depth gradients to resolve ambiguities and providing a calibrated estimate of the scale of the object. However, even with these advantages, there remains much to be done to achieve a pose estimation system that is fast and robust.

We define pose as the 2D or 3D spatial configuration of body parts. In this paper we build on the continuous nonlinear regression approach of [12] by incorporating geometric information. The direct regression approach generates very high accuracy predictions for rigid body part locations, but suffers poor performance on highly deformable body parts such as the hands. In this paper we show how geometric information obtained by exploiting the geodesic structure supports accurate estimation of extremal points which correspond to the most deformable parts. We show how the estimation processes are complementary and yield significantly higher accuracy of deformable parts resulting in a 12% overall improvement in performance.

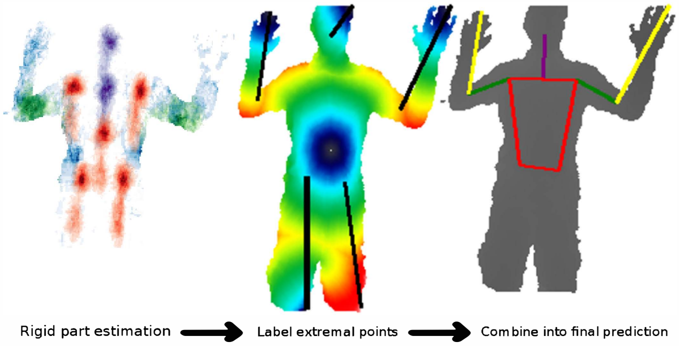

Fig. 1. **Overview**. Given a single input depth image, evaluate a bank of RRFs for every pixel and vote predictions into Hough accumulator images. Locate and label geodesic extrema, and combine with RRFs predictions to form final prediction.

Recently the approach of Shotton et al [22] has received much attention by demonstrating that a model learned on a large training corpus (1M images) can deliver reasonably accurate predictions at frame rates. Their work was extended in [11] showing that a direct regression approach could yield higher accuracy while requiring fewer training samples (300K). Our work follows this development while noting that accuracy remains relatively poor on deformable parts, and for human pose the articulations of the hand make this very challenging.

Our contributions are the following. First, we demonstrate that the performance of Random Regression Forests (RRFs) with Hough and Dijkstra-based geodesic finding are complementary approaches and that their combination results in significantly improved hand localisation. Secondly, this approach is evaluated against the state-of-the-art on existing datasets and a new dataset captured for this work is introduced. The paper is organised as follows: Section II discusses related work, Section III develops the theory and discusses the approach, Section IV details the experimental setup and results and Section V concludes.

II. RELATED WORK

A recent specialist publication surveying the visual analysis of humans is available in [16]. Broadly speaking, discriminative automatic pose estimation can be divided into global (feature-based) and local (part-based) approaches. Global approaches include direct regression of shape context features using Relevance Vector Machines [1], Parameter

All authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

Corresponding author is B. Holt b.holt@surrey.ac.uk

Sensitive Hashing to efficiently find similar poses[21], a manifold based approach using Random Forests trained by clustering similar poses hierarchically [19] and structured prediction using Gaussian Processes [5] and Spectral Latent Variable Models [4].

Local approaches tend to make use of either explicit models, where parts are detected then assembled, or implicit models where the model is geometrically encoded. Explicit part-based models typically consist of a part detection and an assembly stage. The part detection stage has been formulated as an object detection problem using classifiers based on Shape Contexts [2] and Histogram of Oriented Gradients (HOG) [7], [15]. Part models are geometric constraints applied to part detections. The star model, where parts are connected to but independently located near a root, is very common [7]. Pictorial Structures models [8], [2], [6] extend this approach by allowing child parts to be independently located relative to their parent while still allowing for efficient inference provided the model is tree structured. Non-tree models such as the fully connected model of [25] may yield better estimation results but require alternative less efficient approaches for inference. The task of assembly is often achieved by solving the inference problem with belief propagation [23], [2] and loopy belief propagation for cyclical models [27], [24]. Implicit models learn geometric relationships between body parts without the need to assemble detections directly. Examples of this are the semantic labelling approach of [22] and the generalised Hough Transform approaches of [11], [12], all of which use simple binary comparison features as their basis. Our approach applies advances made using RRFs reported recently in a wide range of computer vision problems. The direct regression approach using RRFs differs from traditional part-based approaches by not having part detectors at any scale. Features are computed per pixel and vote into accumulator images, and a final prediction is made by selecting the most likely hypothesis.

Geometric approaches that exploit anthropometric properties have been proposed with promising results. These approaches perform well on deformable body parts like the hands and feet provided the graph structure on which they are based is not degenerate. Plagemann et al [17] construct a graph mesh from depth data and then find and classify interest points to localise joints. Baak et al [3] extract the 5 geodesic extrema from a graph mesh and use their relative positions as an index into a database from which they lookup the pose. The ability to identify extrema efficiently is complementary to implicit models that have shown to perform well on rigid body parts but tend to have difficulty with highly deformable parts, which invariably are the extremal points.

III. Proposed Approach

The objective of our work is to estimate the configuration of a person in the 2D image plane parametrised by a set of body parts.

Our main novelty is an approach to pose estimation that combines the estimation of rigid body parts with estimation of deformable parts. Rigid parts can be accurately estimated using existing approaches, and deformable parts correspond to the extrema which can be found using geodesic distances. We show that rigid and deformable parts exhibit complementary characteristics and demonstrate how the hungarian algorithm can be used to align and label the extrema with rigid joint predictions.

Our secondary novelty is the introduction of a new dataset that is much larger and much more comprehensive than previous datasets for single viewpoint pose estimation. This dataset consists of aligned RGB and pointcloud map, with annotations in both the 2D image plane and in 3D real world coordinates. By releasing this dataset to the community, we feel that research into the state of the art in pose estimation can best be advanced by facilitating the comparison of depth-only, RGB only and hybrid techniques.

A. Preliminary Definitions

We now provide some preliminary definitions. A point-cloud map $\chi = \{\mathbf{c}_p \in \mathbb{R}^3\}$ is the set of $n_x n_y$ 3D points captured from the depth sensor where point $\mathbf{c}_p = (c_p^x, x_p^y, c_p^z)$ corresponds to the pixel at coordinates $p = (i, j) \in \mathbb{R}^2$. The depth values from χ are accessed by $I(p) = c_p^z$. Next, let the set of anatomical landmark labels of interest be defined by the set $\mathbb{Q} = \{$head, neck, shoulder$_L$, shoulder$_R$, hip$_L$, hip$_R$, elbow$_L$, elbow$_R$, hand$_L$, hand$_R\}$.

The set of estimated anatomical locations is defined as $\mathbf{B} = \{(\mathbf{b}_i, q_i)\}_{i=1}^{B}$, where $\mathbf{b_i} \in \mathbb{R}^2$ is the 2D position for the body part on the depth image and $q_i \in \mathbb{Q}$ is the body part label.

B. Image features

For the estimation of rigid body parts, we extract weak images features using the randomised comparison descriptor of [22] from I. Although the feature is weak, it is easy to visualise how the feature relates to the underlying data. Many such features are extracted around a pixel, $p \in \mathbb{R}^2$, and random offsets $\phi = (u, v) \quad |u| < w, |v| < w$ at a maximum window size w to define the feature

$$f_\phi(I, p) = I(p + \frac{u}{I(p)}) - I(p + \frac{v}{I(p)}) \qquad (1)$$

where $I(x)$ is the depth value (the range from the camera to the object) at pixel p in image I and $\phi = (p_1, p_2)$ are the offset vectors relative to p. As in [22], [11], [12], the offset vectors are scaled by a factor $\frac{1}{I(p)}$ to ensure that the generated features are invariant to depth. $I(p')$ is also defined to be a large positive value when p' is either background or out of image bounds.

A feature vector of the above image features for a single body part can be constructed by firstly constructing a fixed set of random offsets relative to the pixel location of the body part. In this paper, we define the *number of generated offsets* as F. Thus, body part, $q \in \mathbb{Q}$, will be associated with a set of random offsets: $\phi_q = (u_j, v_j)_{j=1}^{F}$. Given an image

(I), pixel location (p) and using Eq. 1, it is possible to obtain an F-dimensional *input feature vector* of depth differences:

$$S_q(I) = (f_{\phi_q}(I,p))_{j=1}^{F} \tag{2}$$

C. Pose Estimation using Regression Forests

In order to obtain an initial estimation of the body parts positions, forests of regression trees are used, where two separate regression forests are trained for each body part, one for predicting the row coordinate offset and another for the column coordinate offset. This is achieved by means of a regression forest used to estimate the locations of different body parts given different starting locations on the image. A decision tree is a non-parametric learner that can be trained to predict categorical or continuous output labels.

A regression tree, T, consists of a set of K number of terminal nodes and their output values: $M = \{\mu_1^T, ...\mu_K^T\}, \mu_i \in \mathbb{R}$. There are also binary non-terminal nodes, each associated with a decision rule on thresholded feature values. Specifically, the m^{th} non-terminal node consists of the following: a feature dimension $j_m \in [1, F]$; the threshold $\tau_m \in \mathbb{R}$ that is used in the decision rule: $\{S_{i,j_m} \leq \tau_m\}$ vs $\{S_{i,j_m} > \tau_m\}$. The regression tree T can then assign an F-dimensional input feature vector S with some output value $\mu \in M$ by applying the non-terminal node decision rules from the top to bottom of the tree. This function is denoted as: $\mu = g(S;T)$.

In order to learn the decision tree, a supervised training set consisting of N pairs: $\{(S_i, y_i)\}_{j=1}^{N}$, where $S_i \in \mathbb{R}^F$ is the input feature vector (Section III-B) and $y_i \in \mathbb{R}$ the output offset value. Next, we note that the regression tree induces a recursive partitioning on the dataset, where each non-terminal node splits a subset of the training dataset into two smaller subsets. The parameters of a non-terminal node (i.e. feature dimension(j_m) and threshold value(τ_m)) is configured such that the label mean squared error of the data partition it induces is minimised.

Given that trees have a strong tendency to overfit to the training data, they are often used within an ensemble of N_T number of trees:$\mathbf{T} = \{T_i\}_{i=1}^{N_T}$, where each tree is only allowed to use a random subset of input features. The individual tree predictions are then averaged to form a final prediction with demonstrably lower generalisation errors:

$$G(S;\mathbf{T}) = \frac{1}{N_T}\sum_{t=0}^{N_T} g(S;T_t) \tag{3}$$

We will now define the mathematical convention for identifying regression forests for the different body parts. Let $q \in \mathbb{Q}$ denote a body part, in order to predict its location from a particular pixel in the input image, two regression forests are trained: $\mathbf{T}^{q,1}$ and $\mathbf{T}^{q,2}$, representing the regression forest for predicting the x and y coordinate of the offset vector.

1) Synthesis of Training Data: In order to train the regression trees in the regression forest, it is necessary to extract features and labels from the training data. Firstly, we generate a dictionary of F random offsets $\phi_j = (u_j, v_j)_{j=1}^{F}$. Next, the training data and labels are constructed as follows: For each image in the training set, a random subset of P example pixels is chosen to ensure that the distribution over the various body parts is approximately uniform. For each pixel x_p in this random subset, the feature vector S is computed using Eq. 1 and the offset $o_i \in \mathbb{R}^2$ from every x to every body part q_i is:

$$o_i = x - \mathbf{b}_i \tag{4}$$

The training set is then the set of all training vectors and corresponding offsets. With the training dataset constructed, we train $2B$ RRFs $R_i^1 i \in 1..B$, to estimate the offset to the row of body part $\mathbf{b}_i$ and $2B$ RRFs $R_i^2 i \in 1..B$, to estimate the offset to the column of body part $\mathbf{b}_i$.

D. Hough Voting

The estimated positions for body parts from the random regression forests are combined together using Hough voting. Hough voting is a technique that has proved very successful for identifying the most likely hypotheses in a parameter space. It is a distributed approach to optimisation, by summing individual responses to an input in an parameter space. Here, at each pixel on the input image, the pair of regression forests of a body part outputs a body part location vote. These votes are accumulated into a hough voting image. The location of a particular body part is set at the location of global maxima in its corresponding hough accumulator image.

Our approach uses the two dimensional image plane as both the input and the parameter space. For each body part $q_j \in \mathbb{Q}$ we define a Hough accumulator $\{\mathbb{H}_q\}, \forall q \in \mathbb{Q}$, where the dimensions of the accumulator correspond to the dimensions of the input image I: $\mathbb{H} \in \mathbb{R}^{I_w} \times \mathbb{R}^{I_h}, \mathbb{H} = 0$ for all pixels.

The following algorithm is then used to populate the Hough parameter space $\mathbb{H}_q, \forall q \in \mathbb{Q}$: An example of the Hough voting step in our system can be seen in Figure 3 where the final configuration is shown alongside the accumulator images. This figure motivates our approach in that estimates for rigid body parts tend to be accurately clustered around the ground truth locations whereas estimates for highly deformable parts are distributed far more widely, making any prediction very difficult.

Algorithm 1 Compute probability distribution $\mathbb{H}_q$

```
Input: Image $I$,
  for each pixel $x$ do
    for each label $q \in \mathbb{Q}$ do
      $S = S_q(I)$ (Eq. 2)
      $o_i^1 \Leftarrow G(S; \mathbf{T}^{q,1}(x))$
      $o_i^2 \Leftarrow G(S; \mathbf{T}^{q,1}(x))$
      increment $\mathbb{H}_q(x + o_i^1, x + o_i^2)$
    end for
  end for
```

E. Identifying geodesic extrema with Dijkstra's Algorithm

In order to improve on the estimated positions for the hands and elbows, we use a novel combination of the results

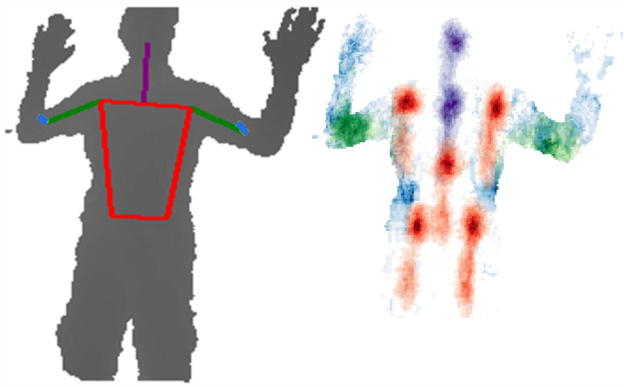

Fig. 2. **Hough Estimates**. Accumulator images for a typical pose are shown colorised. The rigid body part estimates (shoulder, head, torso) are tightly clustered around the real location whereas hand estimates are widely distributed over candidate locations.

from the RRFs and geodesic extrema that can be extracted efficiently using Dijkstra's algorithm.

Firstly, we address the problem of locating candidate locations for the hands. In order to achieve this, we observe that the hands lie on the extremal regions of the body. By representing the 3D real world coordinates from the depth sensor as a graph, the problem of finding extrema can be efficiently solved using Dijkstra's algorithm to compute the shortest paths from the centre of mass to every other vertex and selecting as the extrema the vertex farthest away.

Given the pointcloud map χ we construct a graph $G = (V, E)$ where $V = \mathbf{x}_p$ are the vertices and $E \subset V \times V$ are the edges . Two vertices share an edge if they lie on 8-neighbouring pixels and if the spatial distance between them is less than an empirically determined threshold δ. The set of edges is given by

$$\begin{aligned} E =\{ & x_{(i,j)}, x_{(k,l)} \in V \times V \mid \quad \| x_{(i,j)} - x_{(k,l)} \|_2 < \delta \\ & \wedge \| (i,j) - (k,l) \|_\infty \leq 1 \} \end{aligned}$$

where $\| \cdot \|_2$ is the Euclidean distance and $\| \cdot \|_\infty$ is the maximum norm, and $(i,j),(k,l)$ are 2D coordinates of $\mathbf{x}_{(i,j)}, \mathbf{x}_{(k,l)}$ of the pointcloud map. At every edge $e = (x_{(i,j)}, x_{(k,l)}) \in E$ we store a weight $w(e) = \| x_{(i,j)} - x_{(k,l)} \|_2$

A path is defined as a set of connected vertices. The shortest path between any 2 vertices can be efficiently computed using Dijkstras algorithm. In our method, we start with the vertex that is closest to the centre of mass of the 3D pointcloud. This is the origin from which Dijktra's algorithm runs. We compute the shortest path to every vertex and then define the geodesic extremum as the longest shortest path. Multiple extrema are found by adding a zero-cost edge from the origin to the extremum and repeating the process.

F. Body Part Assignment

From the extrema positions, we can recover which belongs to the left and right hands. To this end, the shortest geodesic distance estimated using Dijkstras algorithm between the left shoulder and right shoulder is used. Thus, the position of the left hand is the extrema location that has the smallest geodesic distance to the left shoulder and similarly for the right hand position.

The Hungarian algorithm [26] has a rich history in operations research where it is used to calculate the optimal distribution of tasks among a pool of workers. More generally the algorithm can be used to solve any assignment problem in polynomial time. The assignment problem is formulated as a minimisation task by creating a $n \times n$ cost matrix C representing the costs of each of n workers to perform any of n jobs and then finding minimum total cost incurred by assigning a worker to a task. The problem of labelling the geodesic extrema extracted from the graph representation of the pointcloud is a bipartite graph matching problem and is ideally suited to the Hungarian algorithm. Construct a cost matrix C where $C_{i,j} = dijkstra(E_i, B_j)$ where E is the set of extrema points, B is the set of body parts (head, left elbow, right elbow, left hip, right hip) and $dijkstra(E, B)$ is the length of the shortest path in undirected graph G from vertex E to vertex B.

The result of this algorithm is the alignment of the extrema with the predicted rigid points which allow us to assign labels to the extrema. We now simply use the extrema assigned left and right hand as the predictions for the left and right hands.

IV. EXPERIMENTAL RESULTS

In this section we evaluate our proposed method and describe the experimental setup and experiments performed. We compare our results to the state-of-the-art [13] on a publicly available dataset, and evaluate our results both quantitatively and qualitatively.

For each body part $q_i \in \mathbb{Q}$, a Hough accumulator likelihood distribution is computed using Algorithm 1. Unless otherwise specified, we construct our training set from 100 random pixels x per training image I, where each sample has $F = 2000$ features $f_\phi(I, p)$.

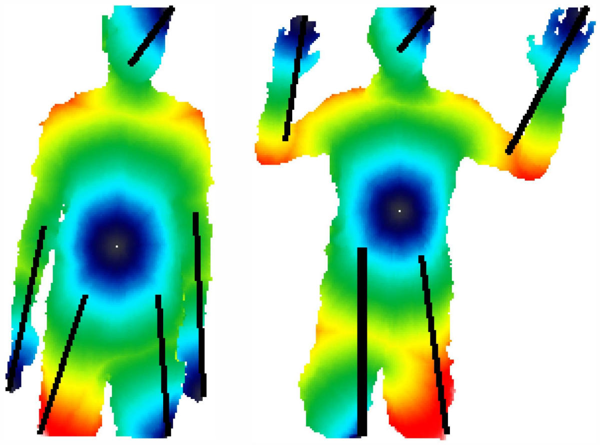

Fig. 3. **Body part assignment**. Examples of optimal assignment by applying the hungarian algorithm. Black lines indicate which extremum is assigned to which body part.

A. Dataset

A number of datasets exist for the evaluation of pose estimation techniques. Most consist of a few hundred images and differ in terms of input modalities, level of annotation, types of body configurations, scenarios and the level of precision required as output. Appearance datasets include *Buffy* [9], *People* [18], *Leeds Sports Poses* [14]. Depth datasets are limited to *CDC4CV Poselets* [13] with 345 training and 347 test frames at 640x480 pixels over 3 subjects and *Stanford ToF* [10] with 2284 frames at a resolution of 144x176 fo a single subject.

Such paucity of data on which to compare and evaluate methods that work on different underlying modalities has hampered the research effort. To this end we propose a new dataset consisting of 19000 training frames and 6500 test frames over 12 subjects. The data consists of of RGB aligned with the 3D pointcloud map with a usermask for background segmentation. The subjects are almost always in frame and are free to perform any movement including turning around. The data is captured against a varied office background. Annotations of 11 upper body parts is povided both in both the 2D image plane and in 3D real-world coordinates.

B. Evaluation

We report our results using the evaluation metric proposed by [9]: "A body part is considered as correctly matched if its segment endpoints lie within $r = 50\%$ of the length of the ground-truth segment from their annotated location." The percentage of times that the endpoints match is then defined as the Percentage of Correctly Matched Parts (PCP). A low value for r requires to a very high level of accuracy in the estimation of both endpoints for the match to be correct, and this requirement is relaxed progressively as the ratio r increases to its highest value of $r = 50\%$. In Figure 5 we show the effect of varying r in the PCP calculation, and we report our results at $r = 50\%$ in Table I as done by [9] and [12]. From Table I it can be seen that our approach represents an improvement on average of 50% for the forearm, upper arm and waist over [12], even though our approach makes no use of kinematic constraints to improve predictions.

Example predictions including accurate estimates and failure modes are shown in Figure 7. Like other geometric approaches, the method presented here tends to work well when the limbs are not occluding other body parts or causing edges between body parts that are not anatomically connected (such as the head and hands). Further work will investigate how to detect and correct degenerate graphs.

Our implementation in python runs at ~ 3 seconds per frame on a single core modern desktop CPU with most of the computation used to construct the graph G.

V. CONCLUSIONS AND FUTURE WORK

We have shown how Random Regression Forests can be combined with a Hough voting framework to achieve robust body part localisation with minimal training data. We use data captured with consumer depth cameras and efficiently compute depth comparison features that support our goal of non-linear regression. We show how Random Regression Forests are trained, and then subsequently used on test image with Hough voting to accurately predict joint locations. We demonstrate our approach and compare to the state-of-the-art on a publicly available dataset. Even though our system is implemented in an unoptimised high level language, it runs in seconds per frame on a single core. As future work we plan to apply these results with the temporal constraints of a tracking framework for increased accuracy and temporal coherency. Finally, we would like to apply these results to other areas of cognitive vision such as HCI and gesture recognition.

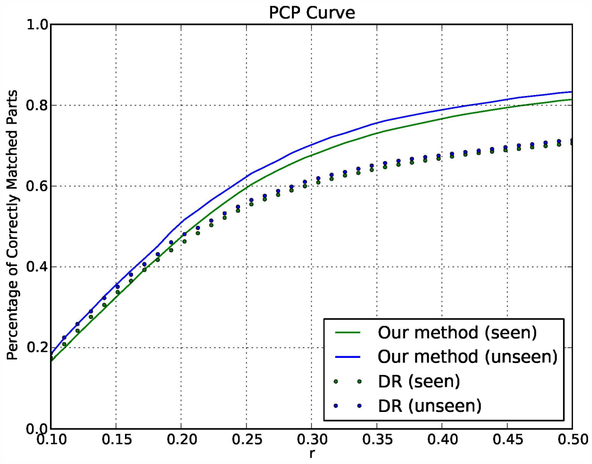

Fig. 5. PCP error curve against [12]. Our method clearly beats theirs for all values of r on both seen and unseen data.

ACKNOWLEDGEMENTS

This work was funded by EU FP7 Project Dicta-Sign, EPSRC Project EP/I011811/1 and the British Machine Vision Association.

REFERENCES

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 28(1):44 – 58, 2006.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, pages 1014 –1021, Miami, FL, USA, June 2009.

[3] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Proc. ICCV*, Barcelona, Spain, November 8 – 11 2011.

[4] L. Bo and C. Sminchisescu. Supervised spectral latent variable models. In *International Conference on Artifical Intelligence and Statistics*, 2009.

[5] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 87:28–52, 2010.

[6] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. In *Proc. BMVC*, volume 2, page 6, London, UK, September 7 – 10 2009.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, pages 1 –8, Anchorage, AK, USA, June 2008.

[8] P.F. Felzenszwalb and D.P Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55 – 79, 2005.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, pages 1 – 8, Anchorage, AK, USA, June 23 – 28 2008.

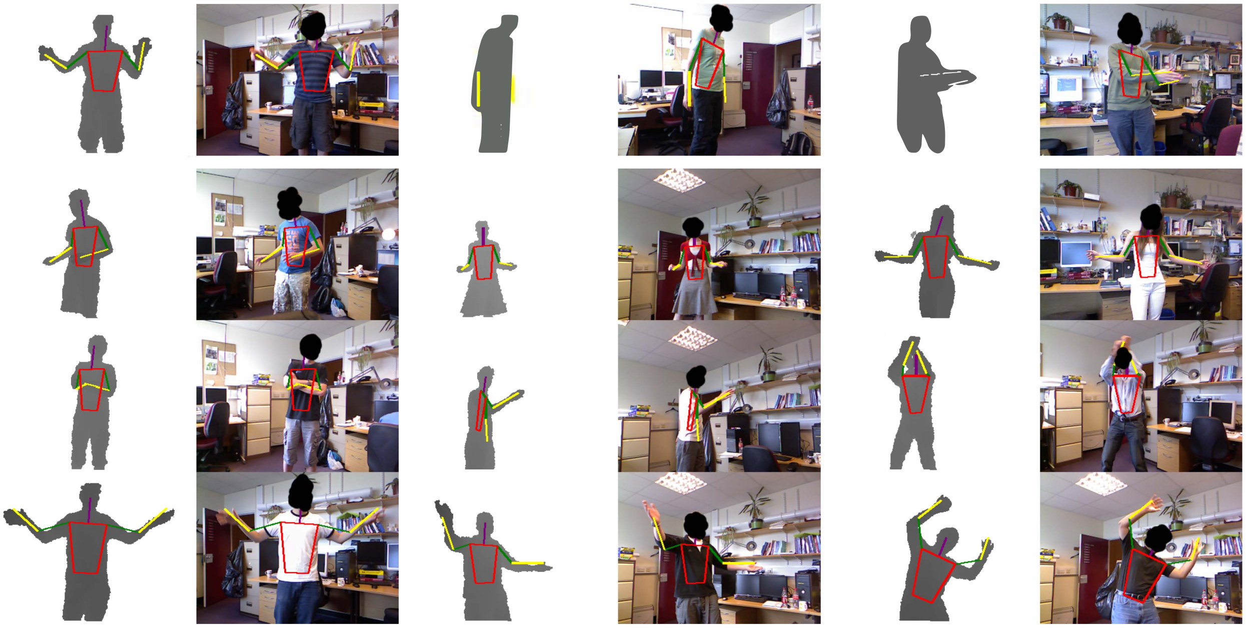

Fig. 4. *New Pose Dataset* Examples taken from the pose dataset showing the complexity of poses against a cluttered background. **Faces blanked out for review.**

	Head	Shoulders	Side		Waist	Upper arm		Forearm		Total
Our method (unseen)	–	–	–	–	–	–	–	**0.59**	**0.57**	**0.83**
Our method (seen)	–	–	–	–	–	–	–	**0.62**	**0.65**	**0.81**
[12] (unseen)	0.93	0.96	0.97	0.98	0.90	0.82	0.80	0.04	0.02	0.71
[12] (seen)	0.96	0.95	0.98	0.98	0.87	0.72	0.71	0.05	0.10	0.71

TABLE I

PERCENTAGE OF CORRECTLY MATCHED PARTS. WHERE TWO NUMBERS ARE PRESENT IN A CELL, THEY REFER TO LEFT/RIGHT RESPECTIVELY. OUR METHOD PROVIDES A DRAMATIC IMPROVEMENT IN HAND LOCATION ACCURACY.

[10] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real time motion capture using a single time-of-flight camera. In *Proc. CVPR*, pages 755 –762, San Francisco, USA, June 2010.

[11] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. ICCV*, Barcelona, Spain, November 8 – 11 2011.

[12] B. Holt and R. Bowden. Static pose estimation from depth images using random regression forests and hough voting. In *7th International Conference on Computer Vision Theory and Applications (VISAPP)*, February 2012.

[13] B Holt, E J Ong, H Cooper, and R Bowden. Putting the pieces together: Connected poselets for human pose estimation. In *Proc. ICCV (Workshop)*, Barcelona, Spain, November 2011.

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*, Aberystwyth, UK, August 31 September 10 2010.

[15] M.P. Kumar, A. Zisserman, and P.H.S. Torr. Efficient discriminative learning of parts-based models. In *Proc. ICCV*, pages 552 –559, Kyoto, Japan, September 29 – October 2 2009.

[16] Th.B. Moeslund, A. Hilton, V. Krger, and L. Sigal. *Visual Analysis of Humans: Looking at People*. Springer, 2011.

[17] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Realtime identification and localization of body parts from depth images. In *Proc. ICRA*, Anchorage, Alaska, USA, 2010.

[18] D. Ramanan. Learning to parse images of articulated bodies. In *Proc. NIPS*, volume 19, page 1129, Vancouver, B.C., Canada., 2006. Citeseer.

[19] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. S. Torr. Randomized trees for human pose detection. In *Proc. CVPR*, pages 1–8, Anchorage, AK, USA, June 23 – 28 2008.

[20] L.A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Estimating human 3D pose from time-of-flight images based on geodesic distances and optical flow. In *Proc. FG*, pages 700–706, Santa Barbara, CA, USA, March21 - 25 2011.

[21] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. ICCV*, page 750, Nice, France, October 14 – 18 2003.

[22] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from a single depth image. In *Proc. CVPR*, Colorado Springs, USA, June 20 – 25 2011.

[23] Vivek Kumar Singh, Ram Nevatia, and Chang Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *Proc. ECCV*, volume 6313 of *Lecture Notes in Computer Science*, pages 314 – 327, Heraklion, Crete, September 5 – 11 2010. Springer.

[24] Tai-Peng Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *Proc. CVPR*, pages 81 –88, San Francisco, USA, June 13 – 18 2010.

[25] Duan Tran and David Forsyth. Improved human parsing with a full relational model. In *Proc. ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 227–240, Heraklion, Crete, September 5 – 11 2010.

[26] Kuhn H. W. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83 – 97, 1955.

[27] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proc. ECCV*, Marseille, France, October 12 – 18 2008.

[28] Y. Zhu and K. Fujimura. A bayesian framework for human body pose tracking from depth image sequences. *Sensors*, 10(5):5280 – 5293, 2010.

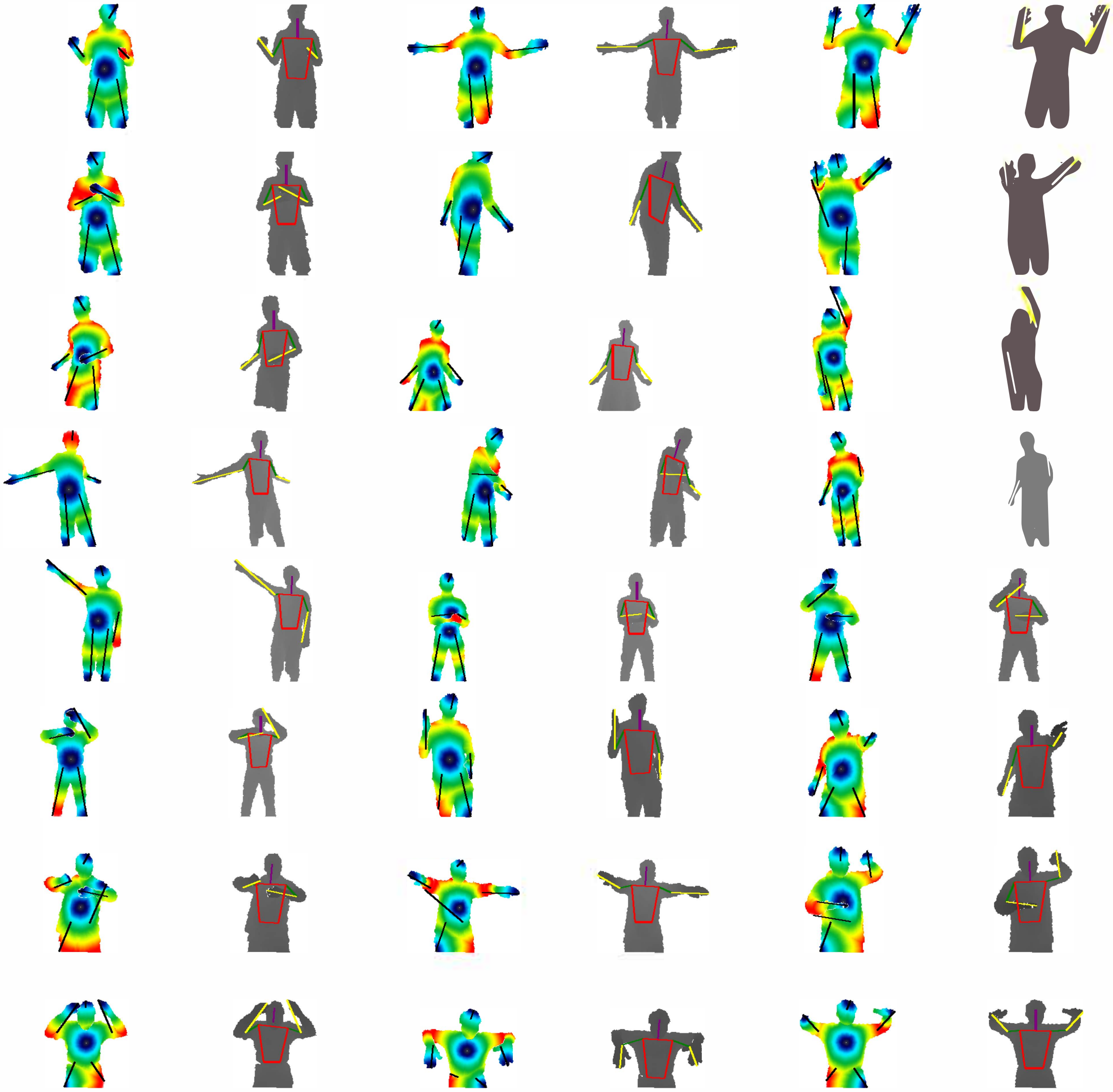

Fig. 6. **Example predictions** Assignment visualisations with corresponding final predictions for a variety of subjects.

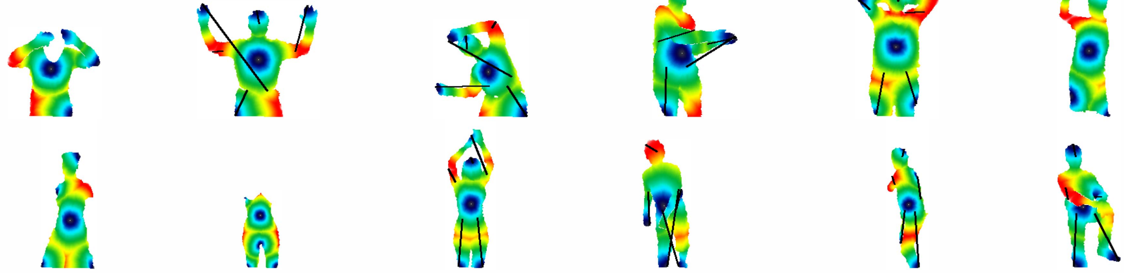

Fig. 7. **Failure modes** Examples where the graph is degenerate leading to prediction failures.