Putting the pieces together: Connected Poselets for Human Pose Estimation

Brian Holt, Eng-Jon Ong, Helen Cooper and Richard Bowden
CVSSP, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom
b.holt,e.ong,helen.cooper,r.bowden@surrey.ac.uk

Abstract

We propose a novel hybrid approach to static pose estimation called Connected Poselets. This representation combines the best aspects of part-based and example-based estimation. First detecting poselets extracted from the training data; our method then applies a modified Random Decision Forest to identify Poselet activations. By combining keypoint predictions from poselet activitions within a graphical model, we can infer the marginal distribution over each keypoint without any kinematic constraints. Our approach is demonstrated on a new publicly available dataset with promising results.

1. Introduction

Unconstrained human pose estimation is one of the major topics in computer vision, specifically in static images. This is a challenging problem, to which a solution would have far reaching implications for Human Computer Interaction, gaming, and Gesture/Action Recognition. The problem is non-trivial for a number of reasons: the significant variation in body shape among the population, the variability of the visual appearance of humans, the human body being a highly deformable object (bounded by kinematic constraints), variable image capture conditions and lighting, camera viewpoint, occlusion and background.

Recently, static pose estimation has attracted much attention, with the release of the Kinect demonstrating that discriminative pose estimation is capable of being both robust and operating in real-time. The fact that this research has matured into a viable commercial product within such a short period of time is a testament to the progress made by the computer vision community over the past few years. However, there still remain significant research challenges. Most pose estimation datasets consist of only a few hundred frames making it hard to compete with a system trained on nearly 1 million frames as used in the Kinect. There is an open question whether the approach will scale to more varied scenarios and if so, how much data would be required for generalisation. Since much human computer interaction and gaming takes place seated on a sofa or a chair, or standing where the majority of the information is derived from the upper body only, there remains a large scope for work on pose estimation within these contexts.

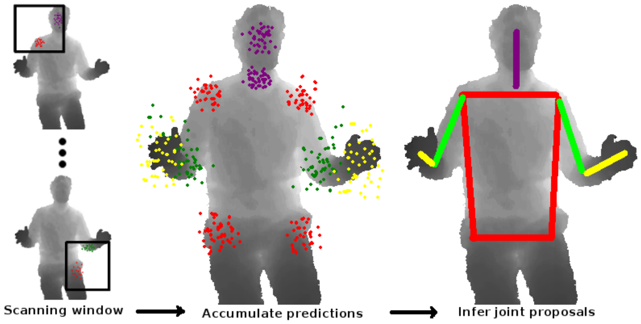

Figure 1. **Overview**. For an input single depth image, run a multiscale scanning window. Each window is evaluated by a Random Forest classifier to detect poselet activations. Each activated poselet makes local predictions for the body parts it was trained on. The overall configuration is inferred by combining keypoint predictions within a graphical model, and finding the maximum over the marginal distributions.

Fundamentally, the objective of our work is to achieve generalisation with a small training set such that we can widen the applicability of the work. Our work builds on the well established traditions of part-based models and example-based pose estimation, which we extend by proposing a novel method called Connected Poselets for articulated static pose estimation. See Figure 1 for an overview of our approach. A multiscale scanning window is applied over the test image, and trained Poselet detectors activate and predict in combination likelihood maps of the key body part locations. Furthermore, by employing an inference step using the natural hierarchy of the body, limb estimation is improved.

An overview of the field is given in Section 2. In Section 3 we formally define our poselet representation which bridges the gap between example-based and part-based pose estimation. We use binary features and a Random Decision Forest (RDF) classifier to detect poselet activations on depth data as opposed to the more traditional approach of Histogram of Oriented Gradients (HOG) with an SVM classifier on appearance data and we propose an adaptation to the traditional Classification and Regression Tree (CART) framework to compare two dimensions at a tree node as the splitting criteria in Section 4. In Section 5 we describe how we apply a bayesian network to improve results, before we present our experimental approach and results in Section 6.

2. Related work

Initially, most approaches to the problem of static pose estimation applied a generative strategy that found the most likely configuration through an iterative synthesis and analysis procedure (see [17] for a survey). Within the last few years, the focus has shifted towards a discriminative approach whereby either the global pose is estimated directly using a classifier [23, 1, 3] or a local approach centred on detection and assembly of parts [10, 2, 29, 26, 22, 15]. Example-based approaches benefit from being fast, yet the many degrees of freedom in the human body means that a large training set containing all the expected poses is required, making it infeasible to cover the entire pose space. Interpolation between poses is challenging given that the poses are distributed on a non-linear manifold, although some have attempted to model this directly [23, 20].

Part based approaches have shown to be very effective. These approaches have two stages: firstly the detection, and secondly the assembly of detected parts into a global configuration. Part detectors have included cascaded classfiers based on Haar features [16], classifiers based on Shape Contexts [2] and HOG [8] within a Support Vector Machine (SVM) classifier [11, 15]. Approaches to part assembly have typically used graphical models, of which Pictorial Structures [13, 10, 2] are an elegant method of relating body parts within a tree structure that supports direct inference of the marginals. Loopy belief propagation models [25, 29, 27] and fully connected models [28] require approximations to infer the marginals. Model parameters can be trained iteratively [2], discriminatively [21] or as an optimisation problem [15]. Further improvements have been made by reducing the search space, with colour models [9], using a branch and bound algorithm [26], or by applying a weak model initially and then concentrating search efforts in subsequent iterations [12].

Depth information has been used before as the basis of static pose estimation with a heuristics approach [18], model based approach [30], a part-based approach with bayesian inference [31], a hybrid model and example-based approach [14] and the object category recognition approach of Shotton et al [24]. For this paper, we also use depth information but our training set is many orders of magnitude smaller than [24], and our learning and inference also differs from [24] in that we estimate the keypoint locations directly and use belief propagation to infer the max marginal rather than visual category segmentation approach.

3. Poselet Representation

The objective of our work is to estimate the configuration of a person in the 2D image plane parameterised by B body parts by making use of a small training set. We define the set of body parts $\mathbb{B} = \{\mathbf{b}_i\}_{i=1}^{B}$ where $\mathbf{b}_i \in \Re^2$. The labels of corresponding to $\mathbb{B}$ comprise $\mathbb{Q} = \{$head, neck, shoulder$_L$, shoulder$_R$, hip$_L$, hip$_R$, elbow$_L$, elbow$_R$, hand$_L$, hand$_R\}$ where $|\mathbb{Q}| = B$.

A poselet, by the definition of Bourdev et al [4, 5], is a set of parts that are "tightly clustered in configuration space and appearance space". The name *poselet* reflects the fact that it is related to concept of a pose (a specific configuration of the human body), but is only a subset of the overall configuration. The basic assumption through part-based pose estimation literature is that a "part" should correspond closely to an anatomical subdivision of the body such as "hand" or "forearm", but is not necessarily the most salient feature for visual recognition especially if the part is itself is highly deformable, making it susceptable to high levels of false positive detections. In contrast, a description such as "half a frontal face and shoulder" or "legs in a scissor shape" may be far easier to detect reliably. While [4] proposed poselets for human detection, this paper proposes to apply the concept of a poselet to pose estimation.

Given that the definition of poselet covers any subset of configurations of the human body, the term could be applied to a part described by single joint or to the configuration of the entire body and anywhere in between. Using just the smallest poselets (anatomically defined parts) corresponds to the part-based pose estimation approach, whereas using the entire configuration corresponds to the example-based approach to pose estimation. The use of poselets can be seen as a hybrid between these two major approaches.

We formally define poselets as follows. Let $\mathbb{P} = \{\mathbf{p}_i\}_{i=1}^{P}$ be the set of P poselets. Each poselet $\mathbf{p}_i = (r_i, q_i, \mathbb{I}_i)$. The number of *instances* of poselet $\mathbf{p}_i$ is $|Pi|$. $r_i = (w_i, h_i)$ where w_i and h_i are the width and height of poselet $\mathbf{p}_i$, $q_i = (q_i^1, q_i^2)$ are the labels of the two body parts on which the poselet was extracted, where $q_i^1 \neq q_i^2$, $q_i^{1,2} \in \mathbb{Q}$ and the set of poselet instances $\mathbb{I}_i = \{\mathbf{i}_{ij}\}_{j=1}^{|Pi|}$. Let $\mathbf{i}_{ij} = (A_{ij}, c_{ij}^1, c_{ij}^2)$ where $A_{ij} \in \Re^D$ is the pixel intensities, subsampled to 24x18 pixels and vectorised. c_{ij}^1 is a hypothesis of the location for body part q_i^1. c_{ij}^2 is the potential location in rectangle r_i for body part q_i^2.

3.1. Extracting poselets

Each poselet $\mathbf{p}_i$ is treated as a semantic class. The goal is to find a set of poselets $\mathbb{P}$ from the training data that maximally span the configuration space. Our algorithm for selecting poselets is very similar to that of [4]. A seed window r_i for poselet $\mathbf{p}_i$ is randomly chosen within a randomly selected training image, and other examples are found by searching each of the other training images for a patch where the local conguration of keypoints is similar to that in the seed. However, since our goal is pose estimation and not person detection, we propose to use a different distance measure, having found empirically that poselets generated using the algorithm of [4] tend to concentrate around the torso at the expense of arm-centric poselets. By applying a euclidean distance over the seed poselet body part locations $c_{i0}^{1,2}$ and new candidate poselet $c_{ij}^{1,2}$, we found that the extracted poselets to be widely distributed over the pose space. A probability distribution over the keypoints for each poselet is computed, which is used at test time to make predictions of the locations of keypoints.

Poselet extraction is essentially an unsupervised clustering step in which the data is preprocessed for the classification task to follow.

4. Learning

Many part detection approaches make use of a HOG feature descriptor [8] within a SVM classifier. While this has been demonstrated to work well, computation of the descriptor is time consuming and its not clear from the litterature that this descriptor is applicable to depth data.

4.1. Decision Trees

Given that our images are sourced from a depth camera, we believe that the pixel relationships within a local window contain sufficient information to efficiently discriminate between poselets, especially with a tree based learner. RDF classifiers are an ensemble classifier, consisting of T individual binary decision trees, usually Classification and Regression Tree (CART) [7]. Breiman [6] demonstrated that using multiple trees trained on random partitions of the dataset are superior classifiers than individual trees.

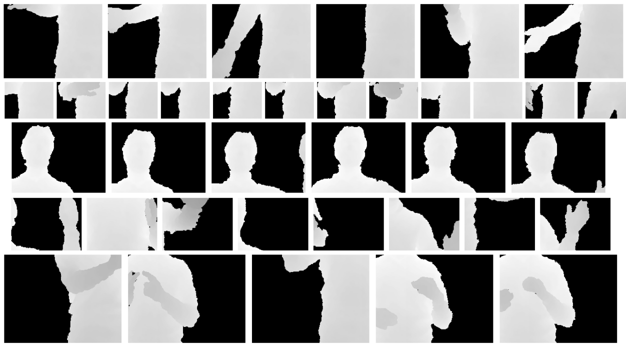

Figure 2. Example Poselets. Each row contains representative examples of a poselet.

The RDF is trained on datapoints and labels $(A, \ell) = \{A_{ij}, i\}_{i=1,j=1}^{P,|\mathbf{p}_i|}$ where $|A| = P \times \sum_{i=1}^{P} |\mathbf{p}_i|$. The labels ℓ for the training data are the indices of the poselet $\mathbf{p}_i$ to which it belongs.

Given a labelled dataset (A, ℓ), Classification and Regression Tree (CART) learners approach the learning task recursively by finding at each node N_n a dimension $l \in 1, ..., D$ on which to split the input data A that minimises some measure of entropy. This leads to trees with binary decision nodes of the form $X_l \leq s_n$ where $s_n \in \Re$ is learned for node N_n. The dimension l and value s_n are learned in a greedy manner to make it computationally feasible, leading to the traditional Gini, information gain or misclassification rate criteria. Shotton et al [24] showed how decision trees can be very effective to predict class (body part) distributions using a binary feature centred on the pixel under evaluation. We also believe that binary pixel comparisons present an effective measure where the underlying image data A_{ij} is depth pixel intensities, but our approach differs from [24] in that the pair of pixels under evalution for a given window is not required to include the pixel on which the window is centred.

A multiscale sliding window paradigm utilising a RDF classifier is used to detect activations of poselets in a test image.

Training Let $\theta = (l, m)$ where $l, m \in 1, ..., D, \quad l \neq m$, $X \in \Re^D$. We define the region R_n based on the decision rule for node N_n as $X_l \leq X_m$, which is a specialisation of the linear combination split $\sum_{l=1}^{D}(a_l \times X_l) \leq s_n$. We say that $X \in R_n$ if $X_l < X_m$.

Given the k^{th} poselet, we define the proportion of observations of poselet i in node N_n as

$$\hat{p}_{ni} = \frac{1}{|N_n|} \sum_{A_{ij} \in R_n} I(i = k) \tag{1}$$

The Gini index of node impurity $Q_n(\theta)$ is then given as

$$Q_n(\theta) = \sum_{k=0}^{K-1} \hat{p}_{ni}(1 - \hat{p}_{ni}) \tag{2}$$

At each node N_n, select θ_n^* such that

$$\theta_n^* = \arg\max_{\theta} Q_n(\theta) \tag{3}$$

This algorithm is applied recursively until all the class distribution at a node is pure ($Q_n(\theta) = 0$) or until the maximum allowable depth of the tree is reached.

Test For a given window position and scale $\mathbf{x}$, resample and vectorise to yield $X \in \Re^D$. Evaluating the RDF on X with P poselets, each tree returns a probability distribution over the poselet set $\alpha_{ti} = (\alpha_{t1}, \ldots, \alpha_{tP})$, where $\sum_{i=1}^{P} \alpha_t i = 1$. α_{ti} is therefore probability that the i^{th} is activated at the given window $\mathbf{x}$.

The output of the decision is averaged to yield the final output for the RDF.

$$P(\alpha|X) = \frac{1}{T}\sum_{t=1}^{T} \alpha_{ti} \tag{4}$$

A poselet $\mathbf{p}_{i^*}$ is considered as detected if

$$i^* = \arg\max_i \alpha_i \quad \text{and} \quad \alpha_{i^*} > c \tag{5}$$

where c is an empirically derived threshold.

5. Predictions and Inference

Given an input image I of resolution $I_w \times I_h$ pixels, for each body part $q_j \in \mathbb{Q}$ we define a probability distribution $\{\mathbb{Y}_q\}, \forall q \in \mathbb{Q}$ where $\mathbb{Y} \in \Re^{I_w} \times \Re^{I_h}, \mathbb{Y} = 0$ for all pixels.

We aim to improve predictive quality by introducing a graphical model to model the relationship between upper body anatomical parts. We define a hierarchy of levels inspired by Pictorial Structures, where the torso is the root, connecting to the elbows and hands. At each level is a restriction set $\mathbb{L}_{L=1,2,3} \subset \mathbb{Q}$.

$\mathbb{L}1 = \{$head, neck, shoulder$_L$, shoulder$_R$, hip$_L$, hip$_R\}$
$\mathbb{L}2 = \{$elbow$_L$, elbow$_R\}$
$\mathbb{L}3 = \{$hand$_L$, hand$_R\}$

We apply the following algorithm to compute the probability distribution $\mathbb{Y}_q^L \forall q \in \mathbb{Q}$.

Algorithm 1 Compute probability distribution $\mathbb{Y}_q^L$

```
Input: Image $I$, level $L$,
  for all scanning windows $W(\mathbf{x})$ do
    for all pixel locations $x, y \in I$ do
      $X \Leftarrow linearise(resample(W(\mathbf{x}))$
      Obtain most likely poselet $\mathbf{p}_i^*$ from Eq. 4 and Eq. 5
      Retrieve $(r^*, q^*, \mathbb{I}^*)$
      for $j \in 0 \ldots |\mathbf{p}_i^*|$ do
        $\mathbb{Y}^L_{q_1^*}(c_j^{*1}) += f(\mathbb{Y}^{Lprev}_{q_1^*}, L)$
        $\mathbb{Y}^L_{q_2^*}(c_j^{*2}) += f(\mathbb{Y}^{Lprev}_{q_2^*}, L)$
      end for
    end for
  end for
  Normalise $\mathbb{Y}_q$
```

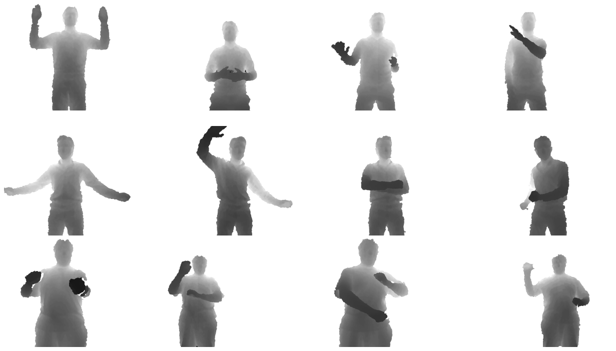
Figure 3. Examples from the dataset.

The function $f()$ is defined as 1 where $L == 1$, but where $L \geq 1$, $\mathbb{Y}^L_{q_1^*}$ is inferred using $\mathbb{Y}^{Lprev}_{q_1^*}$ at previous levels.

The most likely pose configuration $\hat{\mathbb{B}}$ is derived by applying non-maximal suppression to the probability distributions.

6. Experimental work

The objective of this work is to demonstrate a method for pose estimation based on a small training set that is capable of adapting to new poses. While there exist databases for static pose estimation, none are suitable for our work. The Buffy dataset [12] contains annotated upper body poses, but consists entirely of appearance data. Similarly the Image Parse dataset [19] is only appearance data, and consists of full body images. The authors are currently unaware of any dataset of depth images available for comparison of static pose estimation techniques, therefore we have captured our own dataset using the Kinect$^{\text{TM}}$on which to test our approach.

We call our dataset *CDC4CV Poselets*. The dataset consists of depth data and annotations of three participants performing a range of motions in front of the camera. The goal is to ensure that the upper body of each subject remains within the 640×480 window. The dataset is partitioned into training (with 345 images) and test (347 images) subsets (which is approximately the same size as the Buffy (748), Image Parse (305) and ETHZ Stickmen (549) datasets), and code to view the dataset and to compute error metrics is provided. Our training data is doubled by flipping the images and groundtruth horizontally prior to extracting poselets.

For each body part $q_i \in \mathbb{Q}$, a probability distribution is computed for that body part using Algorithm 1. Figure 5 shows the raw probability maps of the left/right elbow/hand respectively as insets as well as the maximum a posterior body part configuration overlaid on the depth image. We show 4 of the 10 possible distributions, that are affected by our graphical model. Figure 6 shows the same body part distributions with the inclusion of the graphical model. No-

	Head	Shoulders	Side		Waist	Upper arm		Forearm		Total
Without inference	0.99	0.78	0.93	0.73	0.65	0.52	0.58	0.14	0.21	0.61
With inference	0.99	0.78	0.93	0.73	0.65	0.69	0.66	0.22	0.33	**0.67**

Table 1. Percentage of Correctly Matched Parts. Where two number are present in a cell, they refer to left/right respectively.

tice the reduced ambiguity in the distributions that leads to a better final estimate of the pose.

We use the evaluation metric proposed by [12]: "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 Correctly Matched Parts (PCP) is then the percentage of these correct matches. We report our results at $r = 50\%$ in Table 1 and in Figure 4 we show the effect of varying r in the PCP calculation. A low value for r expects a very high level of accuracy in the estimation of both endpoints and relaxes this requirement as r increases to its highest value. Our method is 67% accurate overall, showing very high accuracy on the torso, with accuracy decreasing as we move from torso to hand where the benefits of the inference process can clearly be seen. These results are comparable in terms of overall accuracy to the results of [12, 21, 9, 2]. However, as the underlying datasets are different further comparisons are not possible. Importantly these results are achieved without any kinematic constraints on the estimate. We also plot the PCP curve in Figure 4. Inference doesn't provide fine corrections to posture, but is a strong prior that reduces false positive detections in favour of more likely poses.

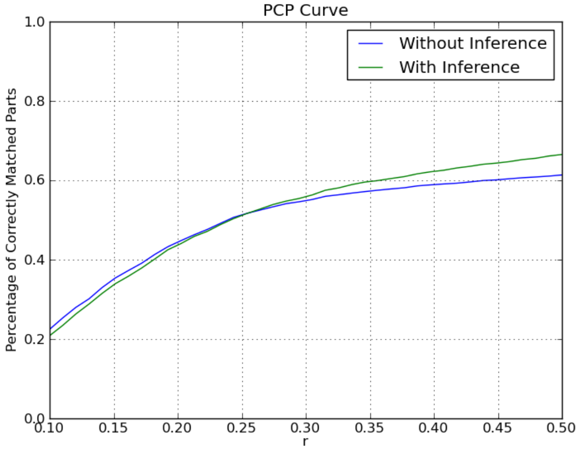

Figure 4. Pixel errors over test set.

7. Conclusions

In this paper we have presented connected poselets for human pose estimation. Each poselet is a class of semantically related parts, extracted by clustering image patches containing similar joint configurations. At its coarsest level a poselet may correspond to the entire body finest level, and at the finest level may correspond to an anatomically defined part. We apply Random Forests modified to support a special case of linear combination splits as the learning basis for the approach. Poselets for pose estimation present a method to bridge the gap between part-based and example-based pose estimation. We have shown that with a small amount of training data it is possible to estimate very accurately the locations of torso joints which can be used to yield improved predictions of elbows and hands on a highly variable dataset. Further work will focus on applying these results within a tracking framework to investigate the applicability to HCI, gesture recognition and other related areas of cognitive vision.

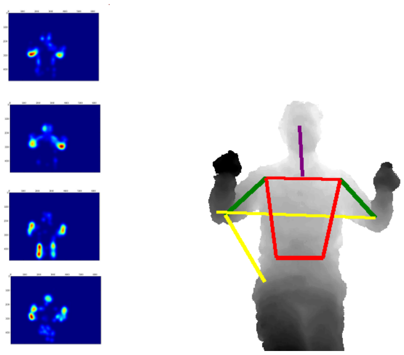

Figure 5. Qualitative example of not using inference to improve predictions.

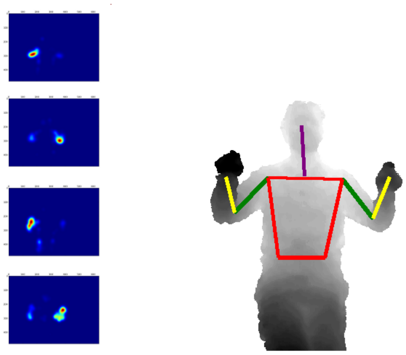

Figure 6. Qualitative example of using inference.

Acknowledgments. This research was supported by the EU project DictaSign (FP 7 ICT project no. 23113).

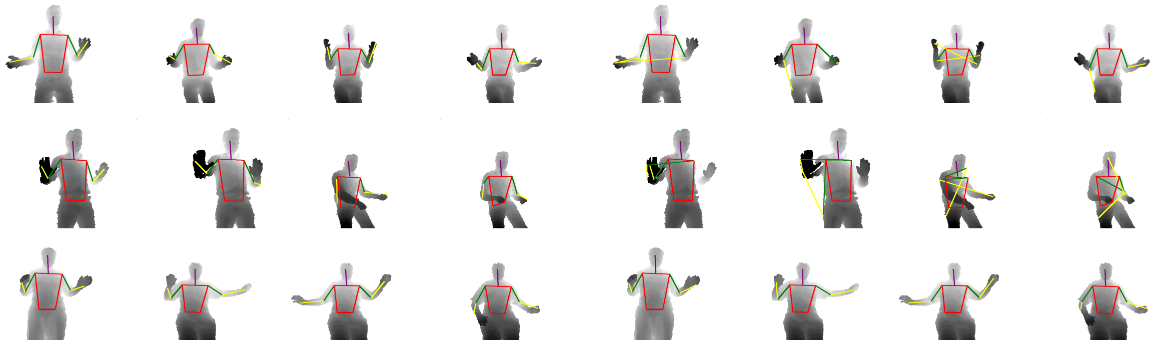

Figure 7. *Left side:* Example predictions using the proposed method with inference. *Right side:* Predictions without inference. This shows where the proposed method works, and also includes a few failure cases.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 2006. 2

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2, 5

[3] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 2010. 2

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 2, 3

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2

[6] L. Breiman. Random forests. *Machine Learning*, 2001. 3

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Chapman and Hall, 1984. 3

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2, 3

[9] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. In *BMVC*, 2009. 2, 5

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

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2

[12] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 2, 4, 5

[13] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 1973. 2

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

[15] M. Kumar, A. Zisserman, and P. Torr. Efficient discriminative learning of parts-based models. In *ICCV*, 2009. 2

[16] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004. 2

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

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

[19] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 4

[20] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. S. Torr. Randomized trees for human pose detection. In *CVPR*, 2008. 2

[21] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010. 2, 5

[22] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV*, 2010. 2

[23] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003. 2

[24] 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. 2, 3

[25] L. Sigal and M. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006. 2

[26] V. K. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV*, 2010. 2

[27] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, 2010. 2

[28] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010. 2

[29] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 2

[30] Y. Zhu, B. Dariush, and K. Fujimura. Controlled human pose estimation from depth image streams. In *CVPR (Workshop)*, 2008. 2

[31] Y. Zhu and K. Fujimura. A bayesian framework for human body pose tracking from depth image sequences. *Sensors*, 2010. 2