Holistic Human Pose Estimation with Regression Forests

Vasileios Belagiannis, Christian Amann, Nassir Navab, Slobodan Ilic

Computer Aided Medical Procedures, Technische Universität München, Germany
{belagian, christian.amann, navab, slobodan.ilic}@in.tum.de

Abstract. In this work, we address the problem of human pose estimation in still images by proposing a holistic model for learning the appearance of the human body from image patches. These patches, which are randomly chosen, are used for extracting features and training a regression forest. During training, a mapping between image features and human poses, defined by joint offsets, is learned; while during prediction, the body joints are estimated with an efficient mode-seeking algorithm. In comparison to other holistic approaches, we can recover body poses from occlusion or noisy data. We demonstrate the power of our method in two publicly available datasets and propose a third one. Finally, we achieve state-of-the-art results in comparison to other approaches.

1 Introduction

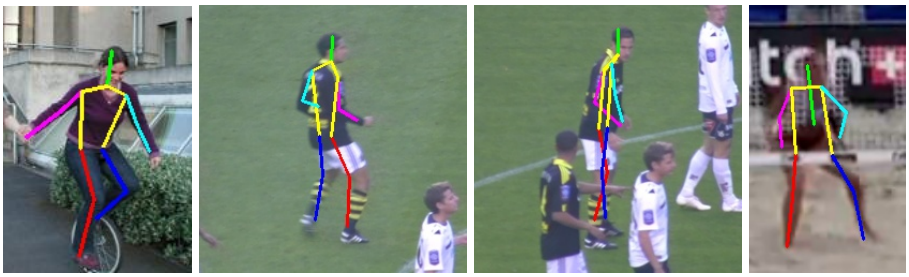
Human pose estimation from single images is a fundamental problem in Computer Vision [1]. It has a wide range of potential applications such as surveillance, health care and human computer interaction. Real life applications involve a huge amount of human appearance variations. Furthermore, out of studio environments usually include dynamic background and clutter. To address these challenges, most of the recent work relies on modelling the human body from an ensemble of parts [2, 3].

There are two main categories of approaches in human pose estimation: holistic and part-based. In both categories, the human pose is defined in terms of a body skeleton which is composed of a number of connected joints. On one hand, the part-based approaches synthesise the body skeleton from a set of parts. The most acknowledged model of this category is pictorial structures [4–6]. Currently, most of the state-of-the-art approaches for human pose estimation rely on pictorial structures [7, 8, 2, 3]. Those approaches have delivered promising results on standard evaluation datasets, but they build on complex appearance and body prior models.

On the other hand, the holistic approaches predict directly the body skeleton by learning a mapping between image features and skeletons [9–12]. These approaches usually face problems with occlusion or noise because they require complete data. They also generalize up to the level at which unknown poses start to appear. However, Random Forests [13] have been proven to generalize well with unknown poses [14, 15].

In this work, we address the problem of human pose estimation in still images, by building on the holistic idea. We propose to learn the appearance of the human body from image patches. These patches, which are randomly chosen from a bounding box around the person, are used for extracting HOG features and training a regression forest [13]. During training, we learn a mapping between image features and human poses, defined by joint offsets. During prediction, we can recover the human pose even under occlusion or from noisy data (Figure 1). Moreover, we propose an efficient algorithm for estimating the mode of the joint density function from the aggregated leaf samples.

In the experimental section, we demonstrate that a holistic approach is not limited to complete data for performing accurate human pose estimation. To show this, we evaluate our model on two publicly available datasets which include self-occlusion, large appearance and pose variations. In addition, we propose a new challenging dataset which is different from the existing datasets because of its low resolution and noisy data. We have compared our method with the state-of-the-art approaches and achieved better or similar results.



(a) (b) (c) (d)

Fig. 1: **Human Poses**: Qualitative results of our algorithm on different datasets. We can recover human poses with large appearance and motion variations. Furthermore, our method handles (b)-(c) self-occlusion or (d) noisy input data.

2 Related Work

There is a tremendous amount of approaches that tackle the problem of human pose estimation from still images [1]. We follow the categorization of the methods into holistic and part-based and review only the most related work.

Part-based approach. Starting from the part-based methods, pictorial structures models have become the current state-of-the-art in human pose estimation in the last decade. They have been introduced in the 70s [6], but got a lot of attention much later [4, 5]. In the pictorial structures models, the human body

is decomposed into a set of body parts, prior on the human pose. The goal is to infer the most plausible body configuration given the image likelihoods, usually estimated by body part detectors, and a prior. One idea for improving the model is by using better appearance models [16–18]. This has also been done by using Random Forests for body part classification [19] or regression [20]. Shape-based body parts generally achieved better performance [21]. The other direction of improvement is to introduce richer priors using a mixture of models [8, 3] or fully connected graphical models [22]. Recently, the idea of modelling the body part templates jointly has been also explored [20, 23]. In [20], two layers of Random Forests capture the information between different body parts, while in [23] the parts are sharing similar shape. Both directions of improving pictorial structures have resulted in strong local appearance and prior models. However, part-based models, such as pictorial structures, fail to capture the whole anatomy of the human body. Morever, they have evolved by building on computationally expensive and complex models.

Holistic approach. Unlike part-based methods, the holistic approaches rely on learning and predicting the joint positions of the human skeleton at once. They usually rely on learning a mapping between image features and human poses. Mapping exemplars to human poses, in particular, became the standard way on holistic pose estimation [24, 10, 25]. The disadvantage of the exemplar-based approaches is the necessity for accurate matching of the whole body. To solve this problem, classification [9], regression [12] and segmentation-based [26] methods have been proposed. However, these methods can be sensitive to noisy input and cannot generalise to unknown poses. In order to cope with these problems, holistic approaches have relied on Random Forests [14, 11, 15]. In the depth domain, Random Forests have been used for classification [15] and regression [14]. In both cases, a holistic model has been proposed for classifying the body joints [15] or predicting their position [14] in the 3D space. In the image domain, Random Forests have been introduced for human body pose classification [11].

Finally, the combination of holistic and part-based methods has been explored by introducing the concept of Poselets [27] in the pictorial structures framework [2, 28]. These approaches have proposed an intermediate representation but they still do not capture the whole anatomy of the human body.

In our work, we adapt the idea of regression forests to the image domain and learn to map image features to 2D human poses. To the best of our knowledge, we are the first ones who apply a regression forest to image data for estimating the body joints at once. The big advantage of our method in comparison to other holistic approaches is our ability to cope with incomplete data.

3 Method

Random Forests have become very popular for human pose estimation from depth data [29, 14, 15]. In this work, we build on a regression forest for extracting the human pose from image data. Below, we explain the basic principles of a regression forest and the way we apply it to our problem.

3.1 Regression forest

A regression forest is an ensemble of regression trees T that estimates continuous output. The goal of training a regression forest is to learn a mapping between image patches and the parameter space. In our paradigm, the parameter space $\mathbb{R}^{2\times N}$ consists of a set of N joints in the 2D space. The body skeleton is defined by the joints and the image patches are estimated using HOG features [30].

In the training phase, a pool of randomly extracted image patches P with associated skeleton joint offsets serves as input to each tree. The patches are extracted from random positions within a bounding box that localises the human. Then, a tree is built from a set of nodes which include binary split functions. Each node encloses a split function θ which is defined on the values of the HOG features of the patch. The HOG feature vector of the image patch is extracted as in [31]. The binary split function determines if a p sample image patch will go to the left P_l or right P_r subset of samples. In particular, the split function is a threshold on one dimension of the HOG feature vector. Among the dimensions of the HOG feature vector, the threshold that gained the best split defines the split function:

$$\theta^* = \arg\max_{\theta} g(\theta) \tag{1}$$

where $g(\theta)$ corresponds to the information gain. The information gain measures how well the split function divides the training data into two subsets P_l and P_r. Thus, the criterion for choosing the split function is to maximize the information gain $g(\theta)$ by optimally splitting the input training image patches of the current node. The information gain can be formulated as:

$$g(\theta) = H(P) - \sum_{i\in\{l,r\}} \frac{|P_i(\theta)|}{|P|} H(P_i(\theta)) \tag{2}$$

where $H(P)$ is the entropy. For estimating the entropy, the sum-of-squares-differences is used:

$$H(P) = \sum_{p\in P} \sum_{j} \left\| \mathbf{v}_{p,j} - \boldsymbol{\mu}_j \right\|_2^2 \tag{3}$$

where the vector $\mathbf{v}_{p,j}$ includes the offsets for each joint j from the image patch centre and $\boldsymbol{\mu}_j$ denotes the mean for each joint offset. In order to estimate the mean $\boldsymbol{\mu}_j$, we introduce a threshold ρ to consider only joints that are close to the sampled patch, similar to [14]. Finally, the tree grows until it reaches the maximum depth, the minimum number of samples per leaf or the information gain for the node drops below a threshold. The same process is a repeated for all the trees of the forest. Finally, we store the offsets of all body joints in the leaves.

3.2 Forest Parameters

In order to correctly train the regression forest, there is a number of parameters that has to be determined for the training data.

Image Patches: The size of all image patches is predefined during training and prediction. Thus, all the HOG feature vectors have the same size. We discretize the image gradients into 9 bins and follow the implementation from [31].

Scale Invariance: The training persons in different training images are apparently of different sizes, but they are all localized by a bounding box. We scale all the data with respect to the height of the bounding box which usually corresponds to the height of the person. This allows us to capture pose variations of different humans using a common scale. Since we assume a localized person, we scale at the prediction phase as well.

Threshold ρ: We argue that a split function has a more local than a global role. For that reason, samples having large offsets are penalized by a threshold. We set it experimentally to 0.8 of the human bounding box height and exclude the joints that are outside this radius.

3.3 Prediction

In the prediction phase, the human is localised with a bounding box which is also rescaled. Similar to training, random pixel positions are used as input to our algorithm. An image patch is extracted for each random position and the HOG feature vector is then estimated. In each tree, the split functions direct, left or right, the input image patch until it reaches the leaf in which we have stored the vectors that predict the joint positions. Thus, the next step is to aggregate the votes of the leaves of the different trees of the forest.

For a certain joint, finding the most probable location of the joint corresponds to estimating the mode of the density function. The most common algorithm for estimating the mode is Mean Shift [32]. However, Mean Shift is a computationally expensive algorithm and requires a significant amount of time to converge, given a plethora of samples at the leaves. To overcome this limitation, we propose the *dense-window* algorithm which is a greedy approach for estimating the mode of a density function from samples. The *dense-window* algorithm relies on a sliding window search in which convergence is deterministic. It only depends on the step of the sliding window and scales linearly with the number of the samples.

To enable fast estimation, the *dense-window* algorithm discretizes all the 2D predictions for every joint on a grid such that every grid cell stores the number of predictions that lie within this cell. The runtime is linear to the number of joint predictions s. Then, an integral matrix is generated for each cell in order to accumulate its votes. All the cells together form an integral image. Now, the window containing the maximum number of points can be found by sliding the window over the integral image. This can be done in $O(m^2)$ time where m is the resolution of the grid. We set experimentally the sliding window to 0.1 of the person's bounding box height and the grid resolution to 100x100 pixels. The complexity of this algorithm is $O(s + m^2)$ which is much faster than $O(Ts^2)$ of Mean Shift, where T is the number of iterations.

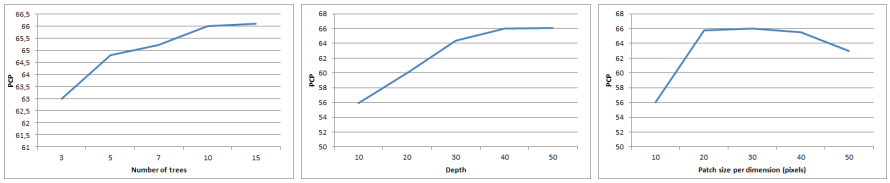
4 Experiments

The current state-of-the-art on human pose estimation, from still images, relies on part-based models [16, 2, 3]. Through our experimental evaluation, we stress that holistic human pose estimation leads to high performance as well. In this section, we analyse our model, evaluate on three datasets and compare it with the state-of-the-art approaches.

First, we present the results for estimating the parameters of the regression forest. We perform all the experiments only on the training images of the Image Parse [3] dataset to avoid parameter over-fitting. Then, we compare our method with an approach which relies on body part classification forests on the KTH Football dataset [19]. In order to show the power of our model in comparison to part-based methods, we evaluate on the Image Parse dataset. Finally, we propose the new and very challenging Volleyball dataset which has very noisy and low resolution data. We evaluate our approach on it and compare with a part-based method [3]. For all the experiments, we use the PCP evaluation score [17].

4.1 System parameters

We first choose the parameters of the regression forest by evaluating on the Image Parse dataset [3]. We mainly focus on determining the number and depth of the trees, as well as the size of the window of the image patch. Figure 2 presents the results.



(a) Number of trees (b) Depth of trees (c) Image patch size

Fig. 2: **Forest parameters**: We have estimated the parameters of the regression forest on the training dataset of the Image Parse dataset [3]. The number and the depth of trees are explored, as well as the size of the image patch.

Based on the results of the Figure 2, we have chosen to use 15 trees with a depth of 40. The trees are very deep due to the high variation in terms of appearance and motion of the human poses. The patch size is set to 30 pixels per dimension.

Finally, we have evatluated the prediction step with the Mean Shift and *dense-window* algorithm and we ended up with almost identical results. In particular with the the *dense-window* algorithm, we achieved PCP 67.1 while with the Mean Shift algorithm PCP 67.0.

4.2 Football dataset

In this experiment, we compare our method with the part-based method which relies on classification forests [19]. In this work the forest classifies each pixel in the image as a specific body joint. Afterwards, a body prior model (i.e. pictorial structures) helps to improve the final result. The results are summarized in Table 1. For the method of Yang and Ramanan [3], we have compiled and test their code that is available online.

	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs	Avg.
Our method	0.86	**0.98**	0.88	0.57	0.92	0.80	0.84
Yang&Ramanan [3]	0.84	0.98	0.86	0.55	0.89	0.73	0.80
Kazemi et al. [19]	0.94	0.96	0.90	0.69	0.94	0.84	0.87
Kazemi et al. [19] + Prior	**0.96**	0.98	**0.93**	**0.71**	**0.97**	**0.88**	**0.89**

Table 1: **KTH Football**: PCP evaluation results for different body parts.

For most of the body parts, we achieve similar results with the classification forest of [19]. In our formulation, we do not rely on a body prior model for smoothing the results. In Figure 3 some of our results on the KTH football dataset are presented.

4.3 Image Parse dataset

The Image Parse dataset [3] is one of the most standard datasets for human pose estimation from images. It includes images of humans with different appearance and pose (Figure 4). In Table 2, we present our results and compare with several part-based approaches.

	Torso	Upper Legs	Lower Legs	Upper Arms	Lower Arms	Head	Avg.
Our method	88.8	**80.9**	**72.8**	58.2	27.5	74.1	67.1
Andriluka et al.[4]	86.3	66.3	60.0	54.6	35.6	72.7	59.2
Yang&Ramanan [3]	82.9	69.0	63.9	55.1	35.4	**77.6**	60.7
Pischulin et al. [2]	**92.2**	74.6	63.7	54.9	39.8	70.7	62.9
Pischulin et al. [33] + [2]	90.7	80.0	70.0	59.3	37.1	77.6	66.1
Johnson&Everingham [8]	87.6	74.7	67.1	**67.3**	**45.8**	76.8	**67.4**

Table 2: **Image Parse**: PCP evaluation results for different body parts.

Our method achieves similar results to the other approaches with the great difference that we use smaller amount of training data. We have used the set of 100 train images for our regression forest. This is significantly lower in contrast to Pischulin et al. ([2],[33]), where they train with 1000 images. Similarly, Johnson and Everingham [8] train with 10000 images. The reason for achieving similar results is that Random Forests can generalise to unknown poses. The only case where we have lower performance is at the lower arms due to the blurry input.

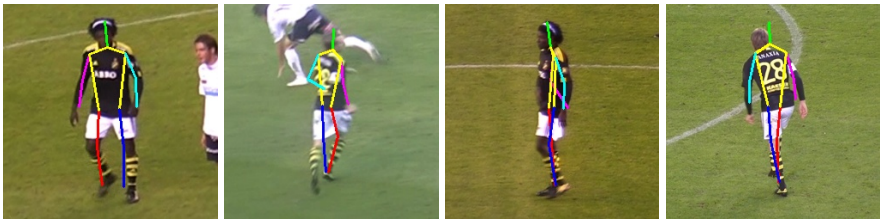


Fig. 3: **KTH Football**: Qualitative results of our algorithm on some samples. The main feature of the dataset is the motion variation.

4.4 Volleyball dataset

We propose the Volleyball dataset [1] for 2D human pose estimation. The dataset is composed of 800 training image of men and 205 testing images of women playing volleyball. We have used two different volleyball matches to create the dataset. The main feature of this dataset is the low quality and noisy image data. In Figure 5, we demonstrate some samples of the Volleyball dataset with the inferred pose. Evaluating on this type of data, we would like to highlight that our holistic model can cope with incomplete data.

	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs	Avg.
Our method	**97.5**	**81.4**	**54.4**	19.3	**65.1**	**81.2**	**63.8**
Yang&Ramanan[3]	76.1	80.5	40.7	**33.7**	52.4	70.5	59.0

Table 3: **Volleyball**: PCP evaluation results for different body parts.

We have evaluated our method on the Volleyball dataset using the PCP evaluation score. In order to compare with another approach, we have trained and tested the code of Yang and Ramanan [3]. The results are summarized in Table 3. We perform better for most of the body parts but we have achieved worse results for the lower arms. This happens because the lower arms are often fully occluded and then the forest predicts an average pose.

5 Conclusion

We have presented a holistic model for human pose estimation from 2D images. The model has been built on Random Forests and image patches. We have demonstrated that our formulation delivers state-of-the-art results by evaluating on two datasets and comparing with other approaches. We have also introduced a new challenging dataset which main feature is the noise and the low quality of image data. In all datasets, we have showed that our holistic approach can perform well and equally compete with the most recent part-based approaches.

[1] `http://campar.in.tum.de/Chair/SingleHumanPose`

Fig. 4: **Image Parse**: Qualitative results of our algorithm on some samples. The dataset has large appearance variation.

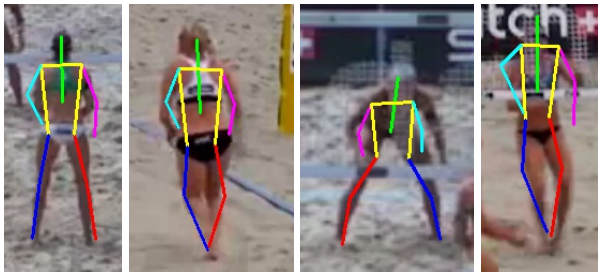


Fig. 5: **Volleyball**: Qualitative results of our algorithm on some samples. This is a new challenging dataset with low resolution and noisy images.

References

1. Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L.: Visual Analysis of Humans. Springer (2011)
2. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR, IEEE (2013) 588–595
3. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR, IEEE (2011) 1385–1392
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR, IEEE (2009) 1014–1021
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV **61**(1) (2005) 55–79
6. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers **22**(1) (1973) 67–92
7. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: CVPR, IEEE (2014)
8. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR, IEEE (2011) 1465–1472
9. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. TPAMI **28**(1) (2006) 44–58
10. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. In: ECCV. Springer (2002) 666–680

11. Rogez, G., Rihan, J., Ramalingam, S., Orrite, C., Torr, P.H.: Randomized trees for human pose detection. In: CVPR, IEEE (2008) 1–8
12. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity-independent human pose inference. In: CVPR, IEEE (2008) 1–8
13. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
14. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: ICCV, IEEE (2011) 415–422
15. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1) (2013) 116–124
16. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. IJCV **99**(3) (2012) 259–280
17. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. (2009)
18. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: ECCV. Springer (2010) 406–420
19. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: BMVC. (2013)
20. Dantone, M., Gall, J., Leistner, C., Van Gool, L.: Human pose estimation using body parts dependent joint regressors. In: CVPR, IEEE (2013) 3041–3048
21. Zuffi, S., Freifeld, O., Black, M.J.: From pictorial structures to deformable structures. In: CVPR, IEEE (2012) 3546–3553
22. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. IJCV **87**(1-2) (2010) 93–117
23. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV, IEEE (2011) 723–730
24. Gavrila, D.M.: A bayesian, exemplar-based approach to hierarchical shape matching. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(8) (2007) 1408–1421
25. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV, IEEE (2003) 750–757
26. Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: ICCV, IEEE (2011) 2220–2227
27. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV, IEEE (2009) 1365–1372
28. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR, IEEE (2011) 1705–1712
29. Criminisi, A., Shotton, J.: Decision forests for computer vision and medical image analysis. Springer (2013)
30. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. Volume 1., IEEE (2005) 886–893
31. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9) (2010) 1627–1645
32. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. TPAMI **24**(5) (2002) 603–619
33. Pishchulin, L., Jain, A., Andriluka, M., Thormahlen, T., Schiele, B.: Articulated people detection and pose estimation: Reshaping the future. In: CVPR, IEEE (2012) 3178–3185