Human Pose Estimation with Fields of Parts

Martin Kiefel and Peter V. Gehler

Max Planck Institute for Intelligent Systems, Tübingen Germany

1 Introduction

In this work we consider the challenging problem of human pose estimation from a single image. This task serves as a crucial pre-requisite step to many high level vision applications, for example human action recognition [13], and natural human computer interfaces [22]. Therefore, it is among the most studied problems in the field of computer vision.

The main difficulty of pose estimation is the weak local appearance evidence for every single body part. While heads nowadays can reliably be detected, localization of general body parts such as arms, or legs remain challenging. Several factors complicate detection: fore-shortening and self-occlusion of parts; different clothing and light environments lead to variability in appearance; some parts might just be a few pixels in size which makes it hard to encode them robustly.

Most work focuses on the main dimensions of the pose estimation problem: use of discriminative appearance information ([19, 17, 18, 26, 27, 9, 8] and many more) and stronger models for the spatial body configuration [21, 23, 17]. Examples of better appearance models are the local image conditioned features used in [19], the use of mid-level representations via Poselets [11, 2, 17], or semantic segmentation information to include background evidence [9, 25, 16, 3]. The spatial model of [10] is a tree, a limitation that obviously does not reflect dependencies in the human body, for example color relation between left and right limbs. This has been addressed by introducing loopy versions [23] or regression onto part positions directly [5, 12]. Another dimension is inference efficiency, richer appearance features typically requires more computations, some approaches perform well but are slow. The same is true for changes in the graph, giving up the tree structure usually results in more involved inference techniques. To speed up inference in pose estimation models enabling the use of richer appearance or graph structure methods like cascading [20] or coarse-to-fine search [19] have been proposed.

In this work we propose the Fields of Parts (FoP) model; a re-formulation of the human pose estimation problem. The FoP model offers a different view on all three dimensions – appearance, structure, and inference. It is inspired by the Pictorial Structures (PS) model, but has different semantics which lead to interesting modeling possibilities. The main idea behind this model is simple: the presence or absence of a body part at every possible location, orientation, and scale of a body part is modelled using a binary random variable.

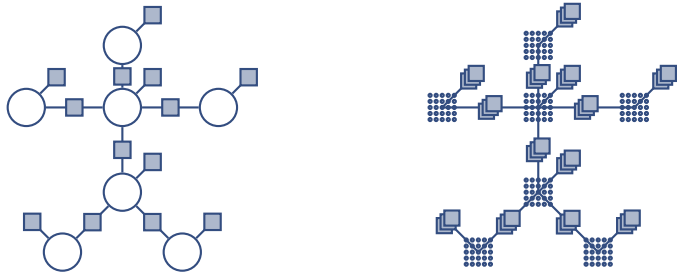

Fig. 1. From Pictorial Structure models (left) to the Fields of Parts model (right). For each body part in the PS model we introduce a field of binary random variables, one for each of its states. When two body parts are connected by a pairwise factor (left) we densely connect the corresponding fields (right). The binary variables 0/1 encode absence or presence of a body part at its location and type (rotation).

The FoP model is built upon advances from three separate domains: efficient inference for segmentation [15], parameter estimation with approximate inference [6, 7], and expressive PS models [27]. We report on modeling, technical, and experimental contributions:

- A reformulation of the human pose estimation problem. This opens up new modelling flexibility and provides a new viewpoint on this well-studied problem.
- An generalization of the inference algorithm from [15]. This makes it possible to use efficient mean field inference in the FoP formulation.
- A new estimator that is tailored to pose prediction using a binary CRF formulation.
- Experimentally, we demonstrate that the FoP model with the same set of parameters as [27] achieves a performance increase of 6.0% on the LSP dataset [14], novel variants improve this even further.

2 Fields of Parts

The flexible body part model of [27] serves as the starting point for our derivation. The authors of [27] propose to model each body part p as a random variable $Y^p = (U, V, T)$ with three values: (U, V) for the position in the image I and $T \in \{1, \ldots, K\}$ a latent type variable. The idea of introducing T is to capture appearance differences of a part due to fore-shortening, rotation, etc, while at the same time increasing the flexibility of the body configuration. We gather all possible states of Y^p in the set $\mathcal{Y}^p$

2.1 Model

We parametrize the problem in the following way: For every part p and every possible state in $\mathcal{Y}^p$ we introduce a binary random variable $X_i^p, i = 1, \ldots, |\mathcal{Y}^p|$. Each such variable represents the presence $X_i^p = 1$ and absence $X_i^p = 0$ of a part at its location, type, and scale in the image. We refer to the collection of variables for a part $X^p = \{X_i^p\}_{i=1,\ldots,|\mathcal{Y}^p|}$ as a *field of parts*. With X we denote the collection of all variables for all parts.

Model	Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	avg
Fields of Parts		83.1	76.5	55.2	29.0	74.8	70.3	63.7	64.7
Yang&Ramanan [27]		80.0	75.2	48.2	28.9	70.4	60.5	53.2	59.5
Yang&Ramanan [27]	(single det.)	79.5	74.9	47.6	28.4	69.9	59.0	51.6	58.7
Pishchulin et al., [18]		**88.0**	**80.6**	**60.4**	**38.2**	**81.8**	**74.9**	**65.4**	**69.9**

Table 1. Comparison of pose estimation results on the LSP dataset. Shown are the APK [27] results (observer-centric annotations [9]).

Given an image I and model parameters θ, we write the energy of a Gibbs distribution as the sum of unary and pairwise terms

$$E(x|I,\theta) = \sum_{p=1}^{P} \sum_{i=1}^{|\mathcal{Y}^p|} \Psi_{\text{unary}}(x_i^p|I,\theta) + \sum_{p\sim p'} \sum_{i=1}^{|\mathcal{Y}^p|} \sum_{j=1}^{|\mathcal{Y}^{p'}|} \Psi_{\text{pairwise}}(x_i^p, x_j^{p'}|I,\theta).$$

Note, that the neighborhood relationship is defined between different fields $p \sim p'$, for example wrist and elbow. Between any two neighbouring fields, all pairs of random variables $(X_i^p, X_j^{p'})$ are connected by a factor node. We illustrate the resulting cyclic CRF graph in Figure 1 for the case of kinematic chain connections $p \sim p'$ and six body parts.

Local appearance of body parts is captured through the *unary factors* Ψ_{unary}. Concretely, we use exactly the same log-linear factors as in [27] in order to make the models comparable: HOG [4] responses $\psi(I)$ and a linear filter θ^p_{unary} of size 5×5 at different scales of the image.

The important piece of the FoP model are the *pairwise connections*. Their form needs to fulfill two requirements: encode a meaningful spatial configuration between neighboring fields, and allow for efficient approximate inference. We are inspired by the observation of [15]. In their work they show that mean field inference in densely connected models with Gaussian pairwise potentials can be implemented as a bilateral filtering. Since for this operation exist highly optimized algorithms [1], the approximate inference is efficient. The pairwise terms in the FoP model have the following form

$$\Psi_{\text{pairwise}}(x_i^p, x_j^{p'}|I,\theta) = \sum_m L_m(x_i^p, x_j^{p'})\ k_m^{p,p'}\left(f_m(i,p;I,\theta), f_m(j,p';I,\theta);\theta\right)$$

$$k_m^{p,p'}(f,f';\theta) = \exp\left(-\frac{1}{2}(f-f'-\mu_m^{p,p'})^T(\Sigma_m^{p,p'})^{-1}(f-f'-\mu_m^{p,p'})\right).$$

The key observation is that this allows to encode the same spatial relation between body part variables X_i^p and $X_i^{p'}$, as the PS model does for Y^p and $Y^{p'}$. This potential is a linear combination of Gaussian kernels k_m weighted by a compatibility matrix L. The Gaussian kernel function k measures the influence of two variables i, j on each other; it has a high value if variables i and j should be in agreement.

To encode the same spatial relationship as PS models we use the 2D positions of the states i as features $f(i,p;I,\theta)$. The influence decreases exponentially depending on the distance of two states i, j and the variance $\Sigma_m^{p,p'}$.

Fig. 2. From left to right: Result from [27], part marginals, stick predictions, for two positive results.

Note that a state i also includes the type/mixture component T. For every part there are as many random variables at the same 2D location as we have mixture components K in the model. For every type/type pair we could use a different offset and variance. Again to enable comparison we implement the choice made in [27].

2.2 Learning and Inference

Exact inference in the FoP model is unfortunately prohibitive due to the loopy structure of the factor graph. We resort to approximate inference, and in particular to a mean field approximation.

We generalizes the results of [15] where there is no part connection relationship $p \sim p'$. In the mean field update step we can exploit the underlying structure of the factor graph to perform bilateral filtering of the two affected neighboring fields.

We use a structured maximum-margin estimator [24] to encourage the model to fit parameters that lead to a low Average Precision of Keypoints (APK).

3 Experiments

We empirically test the proposed method with the standard benchmark dataset of "Leeds Sport Poses" (LSP) [14].

Note that the described FoP model uses *the same* unary potentials and *the same* features for the pairwise potentials as [27]. Also we use the same preprocessing steps: clustering and assignment of the types on the training dataset. Any performance difference of the two methods thus can be attributed solely to the change in model structure, learning objective and inference.

The direct comparison using APK is reported in Table 1, some example detections are depicted in Figure 2. We compare FoP to the PS counterpart and observe that we obtain an improvement for every body part, while being on par on "wrist". The improvement in average APK is 5.2%. For all FoP results we use the top prediction per image only, and have not implemented Non-Maximum-Supression to retrieve multiple detections. The results of [27] when reporting only the top scoring part are also included in the table, in this case we the performance gain is 6.0%. The results increase over all body parts, most prominently on the feet, for example more than 12% on ankles.

References

1. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. Comput. Graph. Forum 29(2), 753–762 (2010)
2. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)
3. Bray, M., Kohli, P., Torr, P.H.: Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In: ECCV (2006)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
5. Dantone, M., Gall, J., Leistner, C., Gool., L.V.: Human pose estimation using body parts dependent joint regressors. In: CVPR (2013)
6. Domke, J.: Parameter learning with truncated message-passing. In: CVPR (2011)
7. Domke, J.: Learning graphical model parameters with approximate marginal inference. PAMI (2013)
8. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC (2009)
9. Eichner, M., Ferrari, V.: Appearance sharing for collective human pose estimation. In: ACCV (2012)
10. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
11. Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. In: CVPR (2013)
12. Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C.: Learning human pose estimation features with convolutional networks. arXiv (2013)
13. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV (2013)
14. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
15. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: NIPS (2011)
16. Ladicky, L., Torr, P.H.S., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: CVPR (2013)
17. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR (2013)
18. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: ICCV (2013)
19. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: CVPR (2010)
20. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: ECCV (2010)
21. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR (2011)
22. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: CVPR (2011)
23. Sun, M., Telaprolu, M., Lee, H., Savarese, S.: An efficient branch-and-bound algorithm for optimal human pose estimation. In: CVPR (2012)
24. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR 6, 1453–1484 (Dec 2005)

25. Vineet, V., Sheasby, G., Warrell, J., Torr, P.H.: Posefield an efficient mean-field based method for joint estimation of human pose, segmentation and depth. In: EMMCVPR (2013)
26. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
27. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI 35 (2013)