Human Pose Estimation with Fields of Parts

Martin Kiefel and Peter Vincent Gehler

Max Planck Institute for Intelligent Systems, Tübingen Germany

Abstract. This paper proposes a new formulation of the human pose estimation problem. We present the *Fields of Parts* model, a binary Conditional Random Field model designed to detect human body parts of articulated people in single images.

The Fields of Parts model is inspired by the idea of Pictorial Structures, it models local appearance and joint spatial configuration of the human body. However the underlying graph structure is entirely different. The idea is simple: we model the presence and absence of a body part at every possible position, orientation, and scale in an image with a binary random variable. This results into a vast number of random variables, however, we show that approximate inference in this model is efficient. Moreover we can encode the very same appearance and spatial structure as in Pictorial Structures models.

This approach allows us to combine ideas from segmentation and pose estimation into a single model. The Fields of Parts model can use evidence from the background, include local color information, and it is connected more densely than a kinematic chain structure. On the challenging Leeds Sports Poses dataset we improve over the Pictorial Structures counterpart by 6.0% in terms of Average Precision of Keypoints.

Keywords: Human Pose Estimation, Efficient Inference.

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 [16], and natural human computer interfaces [28]. 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, legs, or hands 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.

Consequently, the pre-dominant method for this problem are approaches that model both appearance and part configuration jointly. This idea of combining

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 331–346, 2014.

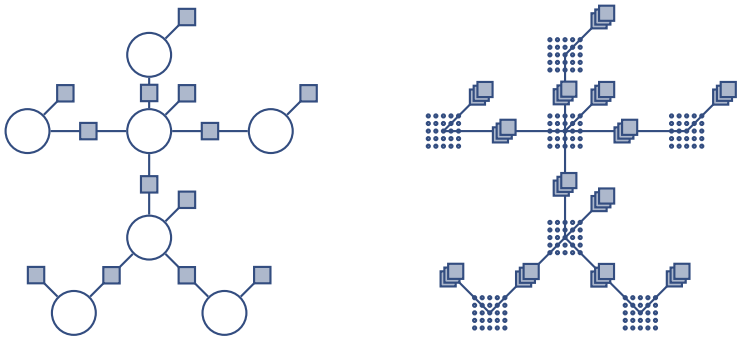

Fig. 1. From Pictorial Structure models (left) to the Fields of Parts model (right). For each body part in the PS model we introduce a field of binary random variables, one for each of its states. When two body parts are connected by a pairwise factor (left) we densely connect the corresponding fields (right), illustrated by the stacked factors. The binary variables 0/1 encode absence or presence of a body part at its location and type (rotation). This is a dense graph and thus contains multiple cycles. This is an illustration with six parts, resp. fields, only.

part appearance evidence with spatial configuration for part relations dates back to [13] and was popularized as a CRF model by [11]. The CRF approach of [11] elegantly expresses pose estimation in a statistical structured prediction problem and introduces with the distance transform an efficient exact inference technique. This model serves as a basis for many variants and thus resulted in significant empirical improvements on increasingly challenging datasets [24,12,17].

Most work focuses on the main dimensions of the pose estimation problem: use of discriminative appearance information ([25,22,23,33,34,10,9] and many more) and stronger models for the spatial body configuration [27,29,22]. Examples of better appearance models are the local image conditioned features used in [25], the use of mid-level representations via Poselets [14,3,22], or semantic segmentation information to include background evidence [10,31,20,4]. The spatial model of [11] 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 [29] or regression onto part positions directly [6,15]. 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 [26] or coarse-to-fine search [25] 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. This results in a huge number of variables, seemingly complicating the matter.

In this paper we show that this model is tractable and present a way to perform efficient marginal inference and more importantly, that this re-parametrization offers new and interesting modeling possibilities. In particular it allows to carry over many ideas from semantic segmentation. We achieve this without the need to explicitly include a segmentation layer or rely on a pose estimation pipeline as a pre-processing step in order to generate body part proposals. The FoP model provides a full interpretation of the image: the presence of a body is explained at every position simultaneously while including evidence from the background without the need for explicit segmentation variables. The graph topography is flexible, we are not bound to a tree structure with restricted potentials in order to use the distance transform. Nevertheless, it does not enforce the detection of a single person in the image anymore. Depending on the application domain this might be advantageous or unwanted.

The marginal inference technique that we propose, namely mean field, is approximate. However, we reason that this is not a severe limitation. We account for the approximation already during training time using Back-Mean-Field learning [7,8]. The inference complexity dependends only linearly on any important dimension of the model: number of part-connections, number of feature dimensions, and size of the image. Furthermore it is amendable to parallelization.

The FoP model is built upon advances from three separate domains: efficient inference for segmentation [18], parameter estimation with approximate inference [7,8], and expressive PS models [34]. 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 (model in Sect. 3.1, discussion in Sect. 3.2).
- An generalization of the inference algorithm from [18]. This makes it possible to use efficient mean field inference in the FoP formulation (Sect. 4.1).
- A new estimator that is tailored to pose prediction using a binary CRF formulation. (Sect. 4.2).
- Experimentally, we demonstrate that the FoP model with the same set of parameters as [34] achieves a performance increase of 6.0% on the LSP dataset, novel variants improve this even further (Sect. 5).

2 Related Work

We adapt the part based formulation from [34] since it offers a good trade-off between flexibility and efficiency. The authors propose to model a body as a collection of body joints, with each body joint being represented as a point in the 2D plane for its position, and a multinomial type variable that accounts for appearance variations. For the FoP model we enumerate all those states and model each one with a binary random variable. A different way to model body part appearance is by a representation as boxes with a center, orientation

and scale, e.g.[2]. The model in [23] combines both the body part and body joint representations into a single joint one. The authors report improved performance, however their proposed method has a runtime of several minutes per image. Other approaches introduce more connections in the factor graph to account for the dependencies of body parts not reflected in the tree structure. One such example is [29] that combines a densely connected model with efficient branch and bound inference.

PS models can be understood as body pose detectors that only model the foreground object while largely ignoring background information. Several authors used segmentation information within their pose estimation model [12,25], this typically complicates the inference process. Therefore, these methods either use sequential algorithms [12] or CRF inference methods with elaborate search based methods [26]. Another way to include background evidence is to explicitly include a separate segmentation layer [20,4,31,32]. Most of these works following this choice have in common that they rely on a separate pose estimation algorithm (e.g.,[2,33]) to retrieve a number of candidate poses. Based on these proposals a CRF structure is then instantiated with factors for segmentation and selector variables for the proposals. Additional CRF layers could be foreground/background segmentation [32], additionally body part segmentation [20], and combination with stereo estimation [31]. Finally, the authors of [10] exploit commonalities in the background appearance within a dataset by fitting a separate color likelihood term to an estimate of background on.

For inference and learning we build upon the advances from [18] that we generalize so it can be used for our purpose. The authors show that mean field inference in densely connected models with Gaussian pairwise potentials reduces to an application of bilateral filtering. The other connection that we draw is to marginal based learning techniques advocated in [7,8]. Domke argues that learning should both take the desired loss function as well as the approximate nature of the inference procedure into account. Our model implements this using Back-Mean-Field learning, also mentioned in [19].

3 Fields of Parts

The flexible body part model of [34] serves as the starting point for our derivation. The authors of [34] 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 $\mathcal{Y}^p$, the entire body is then represented as the concatenation $Y = (Y^1, \ldots, Y^P)$. This PS model defines a Gibbs distribution $p(Y|I, \theta)$, where θ denotes the collection of all model parameters.

In this work we propose a different kind of parametrization. In this section we will introduce the model (Section 3.1) and discuss the gained flexibility that it offers (Section 3.2). The technical contributions on inference (Section 4.1), and

learning (Section 4.2) that enable the use of this parametrization are the topic of Section 4.

3.1 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. The total number of variables per part p is $|\mathcal{Y}^p|$, the total number for all parts $S = \sum_p |\mathcal{Y}^p|$, and thus the state space of X is of size 2^S. We do introduce all variables on different image scales but do not use a super-/sub-script, so as not to clutter the notation. Next, we discuss how to connect the variables in a meaningful way.

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

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

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.

Unary Factors. Local appearance of body parts is captured through the unary factors Ψ_{unary}. In the simplest case this might be a log-linear model

$$\Psi_{\text{unary}}(x_i^p|I,\theta) = \langle \theta_{\text{unary}}^p, \psi_i(I) \rangle.$$

Concretely, we use exactly the same factors as in [34] in order to make the models comparable: HOG [5] responses $\psi(I)$ and a linear filter θ_{unary}^p of size 5×5 at different scales of the image.

Pairwise Factors. 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 [18]. 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

$$\Psi_{\text{pairwise}}(x_i^p, x_j^{p'} | I, \theta) = \sum_m L_m(x_i^p, x_j^{p'})\ k_m^{p,p'}\left(f_m(i,p;I,\theta), f_m(j,p';I,\theta);\theta\right) \tag{2}$$

$$k_m^{p,p'}(f, f';\theta) = \exp\left(-\frac{1}{2}(f - f' - \mu_m^{p,p'})^T(\Sigma_m^{p,p'})^{-1}(f - f' - \mu_m^{p,p'})\right). \tag{3}$$

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 $Y^{p'}$. To see this, let us take a closer look at Eq. (2). This potential is a linear combination of Gaussian kernels k_m weighted by a compatibility matrix L. Remember that all random variables are binary, thus L is of size 2×2.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 $f(i, p; I, \theta)$. Consider two variables $X_i^p, X_j^{p'}$, and their 2D image positions. The two states with maximal influence on each other are those whose 2D position are offset by exactly $\mu_m^{p,p'}$. This influence decreases exponentially depending on the distance of two states i, j and the variance $\Sigma_m^{p,p'}$.

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 [34], namely that the offset only depends on one of the two types (in [34] the child type determines the offset and variance). In summary the same kind of flexible body part configuration is represented in the FoP model. A minor difference is that here, we use Gaussian potentials, whereas in the PS model the spatial term is log-linear (without the exp in Eq. (3)).

3.2 Discussion

The parametrization of the FoP model allows to carry over ideas from semantic segmentation into the pose estimation problem.

It is important to note, that the Gaussian pairwise terms are more general than using only positional information. In fact we can use any features $f(i, p; I, \theta) \in \mathbb{R}^D$ to modulate the influence of two states on each other. For example, we can encode color by appending RGB values to the image locations, resulting in a bilateral kernel. This is in contrast to PS models [11,34,2] where extra local image evidence can not easily be included. The reason is inference time, in order to use the distance transform, the features have to lie on a grid, and for example RGB values do not. Without this restricted form of the features, the general sum product algorithm scales quadratically in the number of states.

We exploit this new possibilities in three different ways: including color information, using foreground/background segmentation of a person, and connecting the CRF more densely.[1]

Additionally to the between-fields connections, we also connect the variables within a single field p using as a pairwise factor

$$\Psi_{\text{pairwise}}(x_i^p, x_j^p | I, \theta) = L(x_i^p, x_j^p) k^p \left(f(i, p; I, \theta), f(j, p; I, \theta); \theta \right). \tag{4}$$

We set $L(x, x') = \delta_{x=0 \text{ and } x'=0}$ and use as features $f(i, p; I, \theta)$ the 2D position and RGB color in a 3×3 neighborhood around the position of i. This potential affects variables X_i^p, X_j^p that are near each other in the image *and similar* in color. For example a variable may be certain that it does not represent a certain body part, a patch in the sky that is blue and is smooth. The term of (4) is "encouraging" all other blue patches in the image (it is densely connected) to also be in state 0. In effect this propagates color background information in the image over the random variables. This is the same type of a bilateral kernel as used in segmentation methods [18,31], in this case it aids prediction of body parts without explicitly reason over segmentation.

Fields of Parts - Segmentation. As a second example, we include a segmentation prediction as extra image evidence into the pairwise terms. The decision tree implementation of [21] and its features are used to train a person/background classifier on the training images. From ground truth bounding box annotations we construct 0/1 segmentation masks for training. The final decision tree yields a score in $d_{u,v} \in [0, 1]$ for every position (u, v) in the image, namely, the fraction of person-pixels in the corresponding leaf. We then append this score to the spatial features to all states i at the corresponding position. This again results in a bilateral kernel and allows for propagation of information to be different inside or outside of the predicted segmentation.

Fields of Parts - Loopy. The CRF of the FoP model is a loopy graph already. In the upcoming section we will show that the inference complexity depends only linear on the number of field-field $p \sim p'$ connections. This allows us to connect the fields more densely than rather along the kinematic chain with only a modest increase in computational complexity. In this variant (*Fields of Parts - Loopy*) we introduce 10 more connections between parts that contain spatial information about each other, like left and right hip, etc.

Future Work. We mention some additional possibilities that we plan to investigate in future work. Beyond standard RGB, different texture and color information can be encoded in f. An interesting example is the mid-level representation used in [22]. The authors condition the pairwise terms of a PS model *globally* on reponses of a poselet detector [3] and report impressive performance

[1] The precise details of the variants are included in the appendix.

gains. With the FoP model this type of evidence can be included *locally*. A connection strength between variables can be modulated given that they are in mutual agreement with a poselet response at a corresponding position.

Another route is to combine the FoP model with the different body parameterization as a collection of sticks/card-boards. For example a "field of sticks" can be fused into the model in the very same way the body part fields are connected.

3.3 Comparison to Pictorial Structures

There are two main differences between the FoP model and the PS model concerning the semantic of their outputs. PS models explain the foreground, they represent a conditional distribution $p(y|I,\theta)$ over all possible body configurations. In contrast the FoP model explains the entire image $p(X|I,\theta)$, i.e. foreground and background. Hence, the FoP model is not just a relaxation of the PS model in the sense that we allow multiple detections for one part. Consider for this again Eq. (2). If at least one of the arguments x_i^p, $x_j^{p'}$ is assigned the label 0 for background then a non-trivial term is added to the total energy. Contrast this to the energy for the PS model where no such term exists[2]. This is much more in spirit of works that try to combine segmentation information into the pose estimation problem [4,10,31,20] but with the crucial difference that the FoP model is designed for pose estimation. It does not require a separate algorithm to generate part proposals, nor is an explicit segmentation layer needed.

Second, consider the case of multiple, including no persons in an image. What would the optimal distribution be? With no person in the image the best a PS model can do is to achieve a uniform distribution over the body poses, it has no notion of absent body parts. In the case of multiple persons the distribution becomes multi-modal. Consequently, the probability mass has to be distributed over different persons and thus the scores will have to decrease. A similar effect will happen if the image size is increased. This can be undesirable depending on the application, the score/probability of a body pose should not depend on the number of people in the image or its size. Therefore a detection step is a crucial pre-requisite for the PS model.

4 Learning and Inference

In this section we present the technical extension of [18] that enables efficient inference (Sect. 4.1) in this model. We then present an estimator tailored to the pose prediction problem with this binary CRF (Sect. 4.2).

4.1 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. With mean field the intractable distribution is replaced by a factorizing approximation Q, usually by the product of

[2] The comparison of the LP-formulations of the two models in the appendix shows another perspective of this.

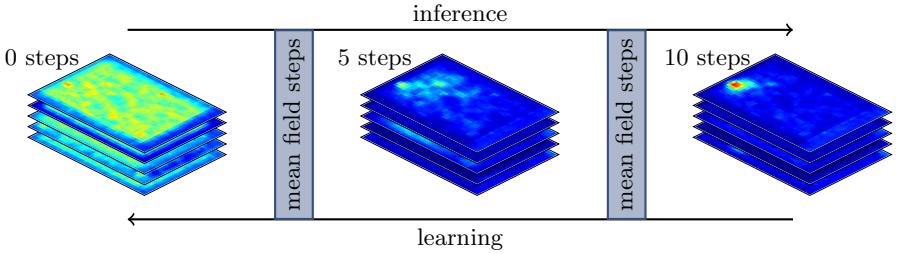

Fig. 2. Evolution of part fields over the filter steps of the mean field updates. For parameter estimation the FoP model builds the gradient w.r.t. the model parameters θ by backpropagating it through the filter steps of the mean field updates.

its marginals $Q(x|I,\theta) = \prod_i Q(x_i^p|I,\theta)$, that are then fit to yield a low KL divergence with the target distribution. Every binary state variable X_i^p gets its approximating probability distribution $Q(x_i^p)$. Note that by finding the factorizing distribution Q we gain all included state marginals of the X_i^p.

The authors of [18] have shown that the mean-field update equations in discrete CRF models with Gaussian pairwise potentials can be implemented by means of bilateral filtering. In the FoP model the mean field update equations can be derived to[3]

$$Q(x_i^p|I,\theta) \propto \exp(-\Psi_{\text{unary}}(x_i^p|I,\theta) - \sum_{p\sim p'} \sum_{l'\in\{0,1\}} \sum_m L_m(x_i^p, l') \sum_{j=1}^{|\mathcal{Y}^{p'}|} k_m^{p,p'}(f(i,p;I,\theta), f(j,p';I,\theta);\theta)\ Q(x_j^{p'} = l'|I,\theta)). \tag{5}$$

This generalizes the results of [18] where there is no part connection relationship $p \sim p'$. In the update step Eq. (5) we can exploit the underlying structure of the factor graph to perform bilateral filtering of the two affected neighboring fields. There are two filtering operations – from p to p' and back – for every field connection $p \sim p'$. The full update algorithm is described in Algorithm 1.

As noted by the authors of [18] this block update scheme is not guaranteed to converge. In practice we have not seen any convergence problems for our model.

To come by the expensive operation of calculating the message from one part field p to another part field p', we also make use of an acceleration technique of the permutohedral lattice [1]. This reduces the computational cost to be linear in the number of states of the two involved fields in contrast to the quadratic cost in the number of states in a naive implementation. We loosen the probabilistic interpretation of the mean field update and allow the compability matrix $L^{p,p'}$ to differ for the messages passed from p to p' and vice-versa.

For images that contain a single person only we report, for each field separately, the state that is most probable to be of value 1,

$$\hat{i}^p = \underset{i\in\mathcal{Y}^p}{\operatorname{argmax}}\, Q(x_i^p = 1|I,\theta). \tag{6}$$

[3] See appendix.

Algorithm 1. Mean field update in the Fields of Parts model

$Q(x_i^p = l) \leftarrow \text{normalize}(-\Psi_{\text{unary}}(x_i^p = l|I, \theta))$
for n iterations **do**
  $\tilde{Q}(x_i^p = l) \leftarrow 0, \forall i, p$ ▷ Initialize all messages
  for $p \sim p'$ **do**
    ▷ Message passing from part p' to p
    $\hat{Q}_{i,m}(l) \leftarrow \sum_{j=1}^{|\mathcal{Y}^{p'}|} k_m^{p,p'}(f(i,p;I,\theta), f(j,p';I,\theta);\theta)\ Q(x_j^{p'} = l)$
    ▷ Compability transform and accumulation of messages
    $\tilde{Q}(x_i^p = l) \leftarrow \tilde{Q}(x_i^p = l) + \sum_{l' \in \{0,1\}} \sum_m L_m(l,l') \hat{Q}_{i,m}(l')$
  end for
  $Q(x_i^p = l) \leftarrow \text{normalize}(\exp(-\Psi_{\text{unary}}(x_i^p = l|I,\theta) - \tilde{Q}(x_i^p = l)))$
end for

Nevertheless, there is no reason not to use a different prediction rule, e.g. in the case of multiple persons in one image. The complexity of the inference algorithm scales very favorably, namely linear in every dimension: number of mean field iterations, number of Gaussian kernels m, linear in the number of pairwise features D, linear in the number of part-part connections $p \sim p'$. Furthermore the model is amendable to easy parallelizations, e.g. by calculating the messages sent by the part fields in parallel. In our current CPU implementation the model requires about $6s$ for inference on a single level in an image of size 100×200.

4.2 Parameter Estimation

Part annotations are available as 2D positions (u, v) of the separate body parts which needs to be translated into the binary CRF formulation. Using K types for part p, the FoP model contains K random variable that represent the position (u, v), one for each type. It is desirable to find parameters θ that yield a high probability for at least one of those variables being in state 1. Here we construct an max-margin objective that is tailored to pose estimation: the predicted state $\hat{i}^p$ (Eq. 6) should be at the correct image position. There is no loss for background states in pose estimation, and thus they are not included in the objective.

Prediction Loss. In practice the performance of body pose models is measured using loss functions that ideally represent the desired output of the systems. For the parametrization of body parts as 2D positions the Average Precision of Keypoints (APK) measure is natural, [34] refers to it as the "golden standard". A prediction is considered correct if it falls inside a small region of the annotated point. To be precise, for a given part at the annotated location i_*, the loss for a prediction $\hat{i}$ is defined to be

$$\Delta^p(i_*, \hat{i}) = I(\|i_* - \hat{i}\| > \alpha \max(h, w)), \tag{7}$$

where I stands for the indicator function. The loss depends on the size of the object to be found (namely height h and width w) and a threshold α to restrict

the region where we count a part as detected. The authors of [34] choose α to be equal to 0.1 on full body pose estimation tasks.

Objective Function. We use a structured maximum-margin estimator [30] to encourage the model to fit parameters that lead to a low loss Δ^p. Similar to the loss we decompose the optimization problem along the parts

$$\text{minimize}_{\theta,\xi^p \geq 0} \sum_p \ell(\xi^p) + C(\theta) \tag{8}$$

$$\text{sb.t. } s^p_{i_*} - s^p_i \geq \Delta^p(i_*, i) - \xi^p \ \forall p, \forall i \in \mathcal{Y}^p \tag{9}$$

$$s^p_i := \sigma^{-1}(Q(x^p_i = 1)|\theta). \tag{10}$$

Equation (9) demands a margin of $\Delta^p(i_*, i)$ between the score of the annotated state i_* and every other state i. The score s^p_i is the result of an inverse sigmoid function[4] applied to the probability of the positive state of a state variable X^p_i. We allow the constraint to be violated by the slack variable ξ^p. The objective (8) consists of a Hingle-loss ℓ and a regularizer C to prevent over-fitting to training data. In our experiments we set C to be the squared norm of the parameter vector θ and weight the result with 0.001. We did not change this value over the course of the experiments.

Optimization. We can rewrite Eqns. (8)+(9) equivalently as an unconstrained optimization problem

$$\text{minimize}_\theta \sum_p \ell(\max(0, -s^p_{i_*} + \max_{i \in \mathcal{Y}^p}(s_i + \Delta^p(i_*, i))) + C(\theta). \tag{11}$$

Every evaluation of the unconstrained objective contains solutions to a loss-augmented inference problem of the APK proxy loss. This problem decomposes over parts and the offending state is the maximum in each loss-augmented field. This objective is piecewise differentiable and we resort to stochastic sub-gradient methods. We apply ADADELTA [35], with decay parameter 0.95 and $\epsilon = 10^{-8}$.

In an implementation only a finite number of mean field iterations are executed, some termination criterion has to be applied. In our experiments we chose a fixed number of 10 iterations to calculate the marginals $Q(X^p_i)$ from Algorithm 1. Performance does not depend on any convergence that may occur when the inference is run longer. When optimizing (11) we take this into account by computing the gradient of the marginals w.r.t. parameters by back-propagating the objective Eq. (11) through the mean field updates as illustrated in Figure 2. This is an application of the Back-Mean-Field idea of [8], a procedure advocated for learning with approximate inference when predicting with marginal inference.

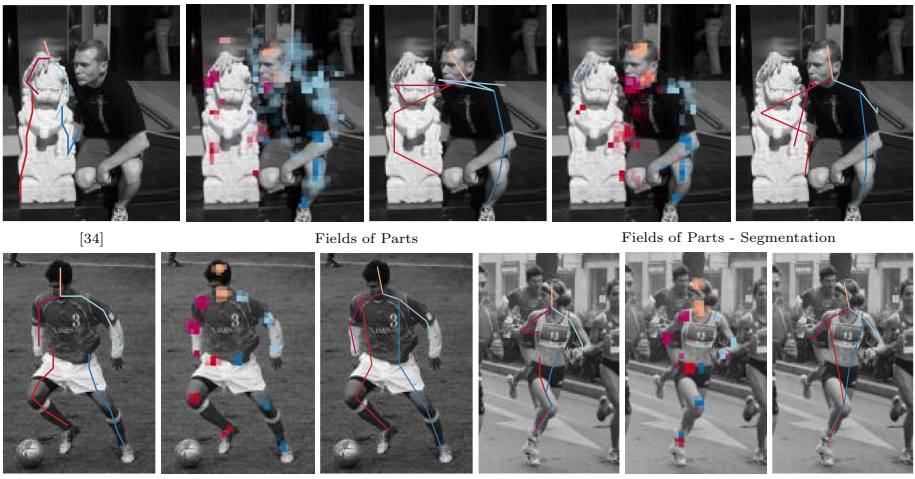

Fig. 3. Top row, from left to right: Result from [34], a visualization of the marginal inference result of the base model, inference result with added segmentation information. The part marginals are considerably sharpened by using the additional features in the pairwise connections. **Bottom row from left to right:** Result from [34], part marginals, stick predictions, for two positive results.

5 Experiments

We empirically test the proposed method with the standard benchmark dataset of "Leeds Sport Poses" (LSP) [17]. This dataset consists of 1000 training and 1000 test images of people performing various sports activities and is challenging due to strong body pose articulation.

5.1 Comparison to Pictorial Structures

The idea of reparametrization the body pose problem can in principle be applied to many PS variants. Here we chose the model [34], and thus it serves as the PS "counterpart" we compare against[5]. Note that the described FoP model uses *exactly the same* unary potentials and *exactly the same* features for the pairwise potentials. Also we use the same pre-processing steps: clustering and assignment of the types on the training dataset. Both models have almost identical number of parameters, a total of about $130k$ most of them unary parameters θ_{unary}. 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 3. First we compare FoP to the PS counterpart and observe that we obtain an improvement for every body part, while being on par

[4] This maximizes the margin with respect to the ratio between the two 1 probabilities and the two 0 probabilities; see appendix.

[5] We thank the authors of [34] for making the code (version 1.3) publicly available.

Table 1. Comparison of pose estimation results on the LSP dataset. Shown are the APK results (observer-centric annotations [10]).

Model	Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	avg
Fields of Parts	Unary only	44.7	28.7	2.1	3.3	5.6	5.8	25.8	16.6
Fields of Parts		83.1	76.5	55.2	29.0	74.8	70.3	63.7	64.7
Fields of Parts	Bilateral	83.3	77.0	56.2	30.9	76.1	71.2	64.5	65.6
Fields of Parts	Segmentation	84.9	77.7	56.9	29.7	78.1	71.9	65.2	66.4
Fields of Parts	Loopy	83.0	76.2	55.7	29.0	77.7	72.0	64.3	65.4
Yang&Ramanan	[34]	80.0	75.2	48.2	28.9	70.4	60.5	53.2	59.5
Yang&Ramanan	[34] (single det.)	79.5	74.9	47.6	28.4	69.9	59.0	51.6	58.7
Pishchulin et al.,	[23]	**88.0**	**80.6**	**60.4**	**38.2**	**81.8**	**74.9**	**65.4**	**69.9**

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 [34] 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.

When comparing the extensions (Bilateral, Segmentation, Loopy) against the FoP model we observe a modest but consistent improvement. Again results increase across all parts. Since all models are trained in the very same way this effect can only be due to the image conditioning terms and extra connections that we introduced.

5.2 Comparison with State-of-the-Art

We also compare using the Percentage of Correct Parts (PCP) measure to [34] and recent results from the literature. The numbers are shown in Table 2. The FoP model performs better than the PS models [2,34].

Interestingly, when comparing the differences between [34] and the FoP models we observe that a higher APK number is not directly translating into higher PCP scores. Especially on the arms, the APK criterion with a threshold of $\alpha = 0.1$ that was used during training, appears not to be indicative of PCP performance. The FoP model makes more points correct in terms of APK and we conjecture that switching to a parametrization based on sticks, the model will improve results on the PCP loss.

Methods that make use of richer appearance information (Poselets [22], Poselets and extra DPM detectors for every body part [23], assumptions about the background color distribution [10]) achieve higher results in terms of PCP. We are encouraged by the result of [10] and believe that adapting their color background model should result in similar performance gains, especially, since they extend [34] by an additional unary factor. The methods of [22,23] make use of mid-level representations for bodies. We already discussed a possibility to adapt

Table 2. Pose estimation results using the PCP criterion on the LSP dataset. We compare our method against the current top performing methods in the literature.

Model	Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
Fields of Parts		82.2	71.8	66.5	52.0	27.7	76.8	59.5
Fields of Parts	Bilateral	83.4	72.8	67.0	52.2	28.0	77.0	60.0
Fields of Parts	Segmentation	84.4	74.4	67.1	53.3	27.4	78.4	60.7
Fields of Parts	Loopy	81.8	73.7	66.9	52.0	26.8	77.3	59.8
Yang&Ramanan	[34]	81.0	67.4	63.9	51.0	31.8	77.3	58.6
Andriluka et al.,	[2]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
Pishchulin et al.,	[22]	87.5	75.7	68.0	54.2	33.9	78.1	62.9
Pishchulin et al.,	[23]	**88.7**	**78.8**	**73.4**	**61.5**	**44.9**	**85.6**	**69.2**
Eichner&Ferrari	[10]	86.2	74.3	69.3	56.5	37.4	80.1	64.3

and extend their approach to a locally conditioned term in Sect. 3.2. However their current implementation runs at several minutes per frame and thus would negatively affect inference time.

6 Conclusion

We have introduced the FoP model, a binary CRF formulation for human pose estimation. Despite being different in structure, it allows to encode a similar spatial dependency structure as done in PS. Further, it permits extensions with more general image conditioned part connections. We have shown two applications of this, by including color and segmentation information as extra features. We have demonstrated how to perform inference and learning in this model through a technical extension of [18], and a max-margin estimator for parameter learning. Because inference complexity depends linearly on almost all relevant model dimensions we also implemented a variant with denser connections than just along the kinematic chain. Experimentally, we validated that the FoP model outperforms [34] on equal ground.

The important new dimension of the proposed parametrization is that it opens up connections to image segmentation. We have discussed several interesting extensions of this model in Section 3.2: image conditioned part configurations, combination with cardboard models, changes in graph topology, etc. Extensions to an explicit person and/or body part segmentation can be easily included, especially, because the inference needs not to be changed.

An interesting aspect of the FoP model is that it explains the image locally at every position; it is not affected by image size, number of persons in the image, or their size. This output semantic differs drastically compared to the PS model. In the future we plan to investigate further along this direction, our goal is to remove the sequential process of current pose estimation pipelines into a single process that performs joint detection and pose estimation of multiple people.

Acknowledgment. The authors would like to thank Leonid Pishchulin for the helpful discussions. MK's work is supported by a grant from Microsoft Research Ltd.

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. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
3. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)
4. Bray, M., Kohli, P., Torr, P.: poseCut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 642–655. Springer, Heidelberg (2006)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
6. Dantone, M., Gall, J., Leistner, C., Gool., L.V.: Human pose estimation using body parts dependent joint regressors. In: CVPR (2013)
7. Domke, J.: Parameter learning with truncated message-passing. In: CVPR (2011)
8. Domke, J.: Learning graphical model parameters with approximate marginal inference. PAMI (2013)
9. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC (2009)
10. Eichner, M., Ferrari, V.: Appearance sharing for collective human pose estimation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 138–151. Springer, Heidelberg (2013)
11. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
12. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
13. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. (1973)
14. Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. In: CVPR (2013)
15. Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C.: Learning human pose estimation features with convolutional networks. arXiv (2013)
16. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV (2013)
17. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
18. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: NIPS (2011)
19. Krähenbühl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: ICML (2013)
20. Ladicky, L., Torr, P.H.S., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: CVPR (2013)

21. Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P.: Decision tree fields. In: ICCV (2011)
22. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR (2013)
23. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: ICCV (2013)
24. Ramanan, D.: Learning to parse images of articulated objects. In: NIPS (2006)
25. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: CVPR (2010)
26. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 406–420. Springer, Heidelberg (2010)
27. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR (2011)
28. 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)
29. Sun, M., Telaprolu, M., Lee, H., Savarese, S.: An efficient branch-and-bound algorithm for optimal human pose estimation. In: CVPR (2012)
30. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR 6, 1453–1484 (2005), http://dl.acm.org/citation.cfm?id=1046920.1088722
31. Vineet, V., Sheasby, G., Warrell, J., Torr, P.H.S.: PoseField: An efficient mean-field based method for joint estimation of human pose, segmentation, and depth. In: Heyden, A., Kahl, F., Olsson, C., Oskarsson, M., Tai, X.-C. (eds.) EMMCVPR 2013. LNCS, vol. 8081, pp. 180–194. Springer, Heidelberg (2013)
32. Wang, H., Koller, D.: Multi-level inference by relaxed dual decomposition for human pose segmentation. In: CVPR, pp. 2433–2440 (2011)
33. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
34. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI 35 (2013)
35. Zeiler, M.: Adadelta: An adaptive learning rate method (December 2012)