We Are Family: Joint Pose Estimation of Multiple Persons

Marcin Eichner and Vittorio Ferrari

Computer Vision Laboratory, ETH Zurich, Switzerland
{eichner,ferrari}@vision.ee.ethz.ch

Abstract. We present a novel multi-person pose estimation framework, which extends pictorial structures (PS) to explicitly model interactions between people and to estimate their poses jointly. Interactions are modeled as occlusions between people. First, we propose an occlusion probability predictor, based on the location of persons automatically detected in the image, and incorporate the predictions as occlusion priors into our multi-person PS model. Moreover, our model includes an inter-people exclusion penalty, preventing body parts from different people from occupying the same image region. Thanks to these elements, our model has a global view of the scene, resulting in better pose estimates in group photos, where several persons stand nearby and occlude each other. In a comprehensive evaluation on a new, challenging group photo datasets we demonstrate the benefits of our multi-person model over a state-of-the-art single-person pose estimator which treats each person independently.

1 Introduction

Look at the photo in Figure 1a. A group of people poses for a souvenir picture. The majority of body parts of the people in the back row are occluded by the people in the front row. Many photos of this kind can be found in personal photo collections or on community sites like Flickr or Facebook.

Unfortunately, even state-of-the-art 2D articulated human pose estimation (HPE) algorithms [1–3] typically fail on such photos (Fig. 1b). These failures are due to treating every person independently, disregarding their interactions. As HPE algorithms lack a global view on the scene, they cannot handle occlusions between people.

In this paper we propose a pose estimation approach which explicitly models interactions between people and estimates their poses *jointly*. Our model handles occlusions between people and prevents body parts of neighboring persons from covering the same image region (Fig. 1c).

Our main contributions are: (i) an algorithm to predict the probability that a body part of a person is occluded given only the locations of all persons in the image (without knowing their poses); (ii) a novel model extending pictorial structures to jointly estimate the pose of multiple persons. This model incorporates the occlusion predictor as well as mutual exclusion terms preventing body parts from different people in the same image region. We also give an efficient inference technique for this model; (iii) a new dataset of group photos fully annotated with a labeling of which body parts are visible/occluded and with the location of visible parts.

Fig. 1: **Group photo scenario.** (a) example image; (b) result of independent pose estimation [1]; (c) result of our joint multi-person pose estimation.

We demonstrate experimentally on the new group photo dataset that (i) the occlusion predictor performs well and better than various baselines, including an occlusion prior probability estimated from a training set; (ii) the whole joint multi-person algorithm considerably outperforms a state-of-the-art single-person estimator [1]. Our source code is available at [4].

Related Works. In this work we explore interactions between people to improve HPE in group photos. In this section we briefly review recent works on relevant topics.

Recovering articulated body poses is a challenging task. We build on Pictorial Structures [5], a popular paradigm for single-person HPE in still images [2,3,5,6] (sec. 3.1).

As a part of our multi-person model we look at occlusions. In articulated HPE some previous works model self-occlusions [7–10]. Here instead we consider occlusions between people. Modeling interactions between people is at the core of our work. They were exploited before by multi-person trackers in the tracking-by-detection paradigm [11–14] (e.g. in [13] as space-time constraints preventing multiple people from occupying the same 3D space at the same time). In [14] the authors learn the behavior of people in crowded urban scenes and predict the path of a pedestrian given the location and direction of others. All these trackers [11–14] handle occlusions at the level of entire persons, who are considered as atomic units (and not at the level of body parts).

To the best of our knowledge, we are the first to propose a joint multi-person occlusion-sensitive model for articulated HPE, where interactions between people are modeled at the level of body parts.

2 We Are Family - Scenario and Dataset

A typical group photo contains several people standing nearby and occluding each others' body parts. We argue that for such photos a joint multi-person reasoning is beneficial over estimating the pose of each person independently.

To investigate this claim we collected a new dataset of group photos, e.g. classmates, sport teams and music bands. We collected the images from Google-Images and Flickr using queries composed of words like "group", "team", "people", "band" and "family". The resulting dataset has 525 images with 6 people each on average. People appear upright in near-frontal poses and often occlude one another (Fig. 1a). They sometimes are even lined up in a few rows, which results in many occlusions. Across different images people appear at a variety of scales and illumination conditions.

The six upper-body parts have been annotated (head, torso, upper and lower arms). For each person, the annotation includes an occlusion vector $\mathcal{H}$ indicating which body parts are visible/occluded, and a line segment for each visible body part. Moreover, the

depth order of the people in each image is also annotated, so we know who is in front of who. We plan to release this new dataset freely on-line.

3 Multi-Person Pictorial Structure Model (MPS)

We first introduce the pictorial structure (PS) framework [5] for HPE of single persons (sec. 3.1), then we describe a naive extension to multiple persons and discuss its shortcomings (sec. 3.2) and finally we sketch our novel joint multi-person method (sec. 3.3).

3.1 Single-Person Pictorial Structures (1PS)

PS Model. In the PS framework [5], a person's body parts are nodes tied together in a Conditional Random Field [15]. Parts l_i are rectangular image patches and their position is parametrized by location (x, y), orientation θ, scale s, and sometimes foreshortening [5, 16]. This parametrization constitutes the state-space of the nodes in the PS. The posterior of a configuration of parts $L = \{l_i\}$ given an image I is

$$P(L \mid I, \Theta) \propto \exp\left(\sum_{(i,j)\in E} \Psi(l_i, l_j) + \sum_i \Phi(l_i \mid I, \Theta)\right) \tag{1}$$

The pairwise potential $\Psi(l_i, l_j)$ is a prior on the relative position of two parts. It embeds kinematic constraints (e.g. the upper arms must be attached to the torso) and, in a few works, also other relations such as self-occlusion constraints [7] or the coordination between parts [17] (e.g. the balance between arms and legs during walking).

In many works the model structure E is a tree [2, 3, 5, 6], which enables exact inference, though some works explored more complex topologies [2, 7, 16, 17].

The unary potential $\Phi(l_i \mid I, \Theta)$ corresponds to the local image evidence for a part in a particular position (likelihood) and it depends on appearance models Θ describing how parts look like.

Appearance Models. The success of PS depends critically on having good appearance models Θ, which limit the image positions likely to contain a part. In an effort to operate on a single image, with unknown part appearances, Ramanan [6] proposes *image parsing*. In this approach Θ are improved iteratively, by adding person specific appearance models computed from previously estimated pose, where the first pose is obtained using only generic edge models as unary potentials.

Person Detection. As in several recent HPE methods [1, 12, 18], we use a generic person detector to determine the approximate location and scale of the persons in an image. This was shown to be useful for estimating pose in uncontrolled, cluttered images [2], as it reduces the state-space of the PS nodes by fixing the scale and reducing the (x, y) search region to an enlarged area around the detection window. In this paper, the set of detection windows $\mathcal{D}$ also determines the set of people $\mathcal{P}$ in the image.

In [1] authors also use the initial detection to obtain person-specific appearance models Θ from a single image (as an alternative to [6]). They propose to compute Θ using part specific *segmentation priors*, learned wrt D, and then improve Θ using an *appearance transfer* mechanism that exploits between part appearance dependencies.

3.2 Naive Multi-Person Algorithm

Given an image with a set of people $\mathcal{P}$, a simple algorithm for multi-person HPE could be: (1) estimate the pose of the first person using an off-the-shelf algorithm, e.g. [1, 3]; (2) remove the support of the estimated pose from the image, e.g. erase pixels covered by the pose; (3) repeat (1)-(2) for the next person.

We call *front-to-back order (FtB)* Z the sequence in which people are processed by the algorithm. Since the true depth ordering is unknown, one must run the algorithm for $|\mathcal{P}|!$ different orders and then pick the best one according to some criterion, e.g. the product of eq. (1) over people.

There are three problems with the naive algorithm: (i) it is infeasible to run due to the factorial complexity in the number of people $|\mathcal{P}|$; (ii) such a greedy algorithm doesn't have a global view on the scene, so people interactions are limited to removing image evidence; (iii) typical HPE algorithms [1, 3] don't handle occlusions and always try to find an image position for all body parts (except [7] for self-occlusions). Therefore, even if the naive algorithm ran over all the orders, it would not find out which parts are occluded. Removing image evidence in step (2) might even lead to double-counting, e.g. when both arms of a person are assigned to the same image region.

3.3 Our Approach to Multi-Person Pose Estimation (MPS)

We propose a joint multi-person Pictorial Structures model which explicitly takes complex interactions between people into account:

$$P(\mathcal{L} \mid I, \Theta, Z) \propto \exp\left(\sum_{p\in\mathcal{P}}\sum_{(i,j)\in E}\Psi(l_i^p, l_j^p) + \sum_{p\in\mathcal{P}}\sum_{i}\Phi(l_i^p \mid I, \Theta, Z) + \sum_{(p,q)\in\mathcal{X}}\sum_{i}\sum_{j}a_{ij}\omega(l_i^p, l_j^q)\right) \tag{2}$$

where the first term is a kinematic constraint as in the 1PS model (eq. (1)), but with additional summations over people $p \in \mathcal{P}$.

Interactions between people are at the core of our work and play an important role in two terms of our joint model. First, the unary potential Φ is updated to depend on the FtB order Z and to include an occlusion state (sec. 5.3). The probability of the occlusion state is estimated specific to each person and body part before running PS inference, as a function of the location of all other persons in the image as given by the detection windows $\mathcal{D}$. Moreover, in sec. 4 we propose techniques to strongly reduce the number of FtB orders that the algorithm has to try out, also based on the spatial distribution of detection windows. The second point where people interactions are modeled is the new inter-people exclusion term ω, which prohibits body parts from different persons (p, q) to occupy the same region (sec. 7), $\mathcal{X}$ is the set of interacting people (sec. 4).

In section 6 we show how to perform efficient approximate inference in our MPS model. Finally, sec. 9 presents a quantitative evaluation of the occlusion predictor and a comparison of our joint MPS model to 1PS.

4 Reducing the Number of Front-to-Back Orders

We propose exact and approximate ways to reduce the number of FtB orders $\mathcal{Z}$.

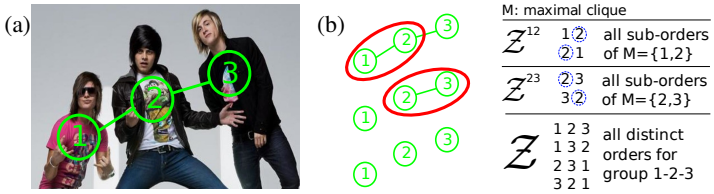

Fig. 2: **FtB Orders.** (a) Group example (b) Calculating distinct FtB orders, red: maximal cliques found, blue: position of the common node in each FtB order

4.1 Exact Reductions

A person can influence another person only if they are within a certain proximity. We define two persons to be *interacting* if they are closer than their arm extents (more precisely, if their enlarged detection windows overlap (Fig. 2a)). An image can then be represented as an interaction graph $\mathcal{X}$ with nodes corresponding to people and edges corresponding to interactions (Fig. 2a).

Group Independence. We define *groups of people* as connected components $\mathbb{G}$ in the interaction graph $\mathcal{X}$. The first reduction is found by observing that any two persons from different groups cannot interact (i.e. groups are independent). Hence, pose estimation can be run on each group independently, and so the effective total number of orders is reduced from $|\mathcal{P}|!$ to $\sum_{\mathcal{G}\in\mathbb{G}} |\mathcal{G}|!$

Order Equivalence. Within a group, different FtB orders might lead to the same pose estimation result. This is the case for orders 132 and 312 in the graph 1-2-3, as there is no direct interaction between nodes 1 and 3 (Fig. 2a). We say that the two orders are *equivalent*. If person 2 is in the back then the order between 1 and 3 has no influence on pose estimation results. Analogously, orders 213 & 231 are also equivalent. Hence, there are only 4 distinct FtB orders instead of $3! = 6$ in the graph 1-2-3.

This intuition is formalized in the following algorithm to find all distinct FtB orders in a group $\mathcal{G}$: (1) find the maximal clique M of $\mathcal{G}$; (2) keep a record which nodes are in M and then remove all intra-clique edges from $\mathcal{G}$; (3) compute sub-orders of the maximal clique M as all permutations of its nodes ($\mathcal{Z}^{12}$ and $\mathcal{Z}^{23}$ in Fig. 2b); (4) repeat until there are no edges in $\mathcal{G}$; (5) compute the distinct orders of $\mathcal{G}$ as all permutations between sub-orders of all maximal cliques $\mathcal{M}_\mathcal{G}$ found, concatenated at the position of the common node ($\mathcal{Z}$ in Fig. 2b).

Group independence and order equivalence reduce the number of orders to:

$$\sum_{\mathcal{G}\in\mathbb{G}} \prod_{M\in\mathcal{M}_\mathcal{G}} |M|! \tag{3}$$

where $\mathcal{M}_\mathcal{G}$ is the set of maximal cliques found in group $\mathcal{G}$.

4.2 Approximate Reductions

As the vast majority of group photos are taken parallel to the ground plane, the people appearing higher in the image are typically further away from the camera. Moreover if a person appears larger then another, then it is likely closer to the camera. We propose two heuristics based on the spatial arrangement of the detected persons $\mathcal{P}$ that capture these observations. Both estimate which person in a pair (p, q) is in front of the other:

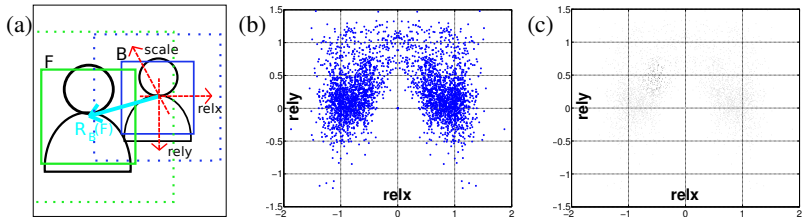

Fig. 3: **Inter People Occlusion Probability.** (a): detection windows (solid), enlarged detection windows (dashed), (b) Distribution of relative locations of occluders f wrt the occluded person b over the training set, (c) relative contribution of the training points for test point [-0.5,0.5,1]

Relative Size. If p is more than 2.5 times bigger than q, then p is in front of q.
Relative Position. If the center of p is higher than the top of q, then p is behind q.

5 Occlusion Probability (OP)

The visibility of the body parts of a person depends on her position with respect to other persons and wrt to the image border. We take both aspects into account by defining two types of occlusion probabilities. One type defines the probability that a part of a person is occluded, given the locations of all other persons and their FtB order (sec. 5.1). The other type defines the probability that a part is not visible, given the location of the person wrt to the image borders (sec. 5.2). We combine both probabilities specific to each person p and body part i into a single occlusion prediction $\mathcal{O}_i^p$ (details in sec. 6), which is then used to set the energy of a new occlusion state in our MPS model (sec. 3.3).

5.1 Inter People Occlusion Probability (POP)

When two people are standing nearby it is likely that one is occluding some body parts of the other. The inter people occlusion probability (POP) is defined between a pair of interacting persons. One person f is considered to be in the front (occluder) and the other b in the back (according to the FtB order). POP tells how likely a body part l of b is occluded given the relative location $\mathcal{R}_b(f)$ of f in b's coordinate frame (i.e. $\mathcal{R}_b(f) = [(x_b - x_f)/w_b, (y_b - y_f)/h_b, h_f/h_b]$, with x, y, w, h the center, width, and height of a window (Fig. 3a).

Learning. We model POP as a non-parametric distribution $P(l_i^b = \text{o} \mid f, \mathcal{T})$ where l_i^b is part i of the back person and $\mathcal{T}$ is a set of training person pairs. For each pair (f, b), the training data is $\mathcal{R}_b(f)$, defined as above, and the ground-truth occlusion vector $\mathcal{H}^b$ of the back person b (which is annotated in our dataset (sec. 2)) (Fig. 3b).

To take into account the uncertainty of the detector around the true position of a person, we run it on the training images and then associate detection windows to the annotated persons (as in [1]). This gives the person windows used for training.

Every pair of interacting persons (f, b) leads to a training sample $(\mathcal{R}_b(f), \mathcal{H}^b)$ (two persons interact if their enlarged windows overlap, sec. 4.1). We determine which is f using the true FtB order, which is annotated in our dataset (sec. 2).

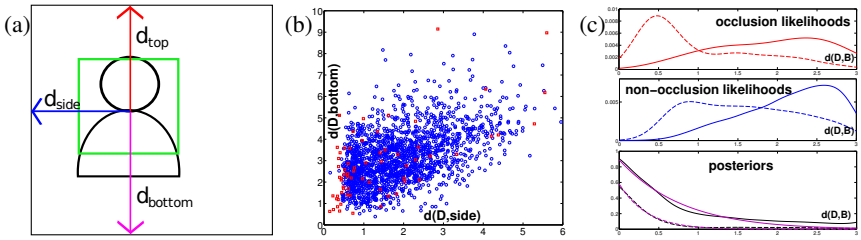

Fig. 4: **Border occlusion probability.** (a) Distances to border types. (b) Distribution of $d(D, B)$ wrt the bottom (y-axis) and side (x-axis) borders over the training set $\mathcal{T}_D$. Red dots: windows of persons with occluded upper arm. Blue dots: not occluded. (c) Top: example of occlusion likelihoods $P(d(D, B) \mid l_i = \text{o})$ for upper arms wrt to side and bottom borders (dashed and solid curves respectively). Middle: as top but for non-occlusion likelihoods $P(d(D, B) \mid l_i \neq \text{o})$. Bottom: as top but for posterior distributions $P(l_i = \text{o} \mid d(D, B))$ (in black) and their parametric approximations $Y_\zeta(x)$ (in magenta) cropped to the range of the posteriors.

Test Time. At test time, we compute the probability that a body part i of a new person p is occluded by a person q in front of her:

$$P(l_i^p = \text{o} \,|\, q, \mathcal{T}) = \sum_{(f,b)\in\mathcal{T}} \alpha^{qpfb} \mathcal{H}_i^b \quad \text{with} \quad \alpha^{qpfb} = \frac{\mathcal{N}(\|\mathcal{R}_p(q) - \mathcal{R}_b(f)\| \mid 0, \sigma)}{\sum_{(d,c)\in\mathcal{T}} \mathcal{N}(\|\mathcal{R}_p(q) - \mathcal{R}_c(d)\| \mid 0, \sigma)} \quad (4)$$

The weights α^{qpfb} are set according to normalized Gaussian-weighted Euclidean distances between the relative location of the test pair (q, p) and those of the training pairs $\mathcal{T}$ (Fig. 3c). The resulting POP value is always in $[0, 1]$.

For a given FtB order Z, if person p is behind more than one occluder then the POP probability of her part i is:

$$P(l_i^p = \text{o} \mid Z, \mathcal{T}) = \max_{f\in\mathcal{F}_Z^p} P(l_i^p = \text{o} \mid f, \mathcal{T}) \quad (5)$$

with $\mathcal{F}_Z^p$ the set of occluders of p in FtB order Z.

FtB orders for POP. Only the *immediate* neighborhood $\mathcal{V}^p$ of person p in the interaction graph has an influence of her POP values. Therefore, the FtB orders for calculating POP are node-specific (as opposed to the FtB orders for pose estimation, which are group-specific (sec. 4.1)). Since $\mathcal{V}^p$ has a star-like topology, all its maximum cliques $\mathcal{M}$ have size 2, so the number of FtB orders affecting POP values for person p is $|\mathcal{Z}^p| = 2^{|\mathcal{V}^p|}$, typically much smaller than the number of FtB orders in her group.

5.2 Border Occlusion Probability (BOP)

Some parts of a person might not be visible due to her proximity to an image border. We model here the probability of a part being occluded given the location of the person wrt to image borders $\mathcal{B} = \{\text{top}, \text{bottom}, \text{side}\}$.

We define BOP as $P(l_i^p = \text{o} \mid d(D^p, B), \zeta_i^B)$ the probability that part i of person p is not visible given the normalized distance $d(D^p, B)$ of her detection window D^p to a border B (Fig. 4a). ζ_i^B are the parameters of the distribution.

Learning. To learn BOP we use our group photo dataset to collect training detection windows $\mathcal{T}_D$ and associate them to ground-truth occlusion vectors $\mathcal{T}_\mathcal{H}$ (as in sec. 5.1). For each type of body part i (e.g. right upper arm) and type of border B, we construct the occlusion $P(d(D,B) \mid l_i = \text{o})$ and non-occlusion $P(d(D,B) \mid l_i \neq \text{o})$ likelihoods as non-parametric kernel density estimates on the training data $D \in \mathcal{T}_D$ (Fig. 4b). The Bayesian posterior of l_i being occluded given the distance to B is (Fig. 4):

$$P(l_i{=}\text{o} \mid d(D,B)){=}\frac{P(d(D,B)\,|\,l_i{=}\text{o})P(l_i{=}\text{o})}{P(d(D,B)\,|\,l_i{=}\text{o})P(l_i{=}\text{o}){+}P(d(D,B)\,|\,l_i{\neq}\text{o})(1-P(l_i{=}\text{o}))} \quad (6)$$

where $P(l_i = \text{o})$ is a prior calculated as the frequency of occlusion of part i over the training set. We approximate the non-parametric posterior estimates $P(l_i = \text{o} \mid d(D,B))$ with a parametric function $Y_\zeta(x) = c\mathcal{N}(x \mid \mu, \sigma)$, fitted in the least square sense. This makes BOP more compact and does not restrict the image size at test time. As Fig. 4c shows, the fitted functions are very close to the non-parametric posteriors.

Test Time. At test time, we compute the probability that a body part i of a new person p is not visible wrt to each border type $B \in \mathcal{B}$ and then select the maximum:

$$P(l_i^p = \text{o}\,|\,D^p, \mathcal{B}, \zeta_i^\mathcal{B}){=}\max_{B\in\mathcal{B}} P(l_i^p = \text{o}\,|\,d(D^p,B), \zeta_i^B){=}\max_{B\in\mathcal{B}} Y_{\zeta_i^B}(d(D^p,B)) \quad (7)$$

where $\zeta_i^B = \{c, \mu, \sigma\}_i^B$ are the parameters of the posterior approximation $Y_\zeta(x)$ for border type B and part type i.

5.3 Incorporating Occlusion in the MPS Model

To handle occlusions in our MPS model (eq. (2)) we add an occlusion state to the state-space of the nodes (body parts). This results in an additional entry in the unary appearance likelihood $\Phi(l_i \mid I, \Theta, Z)$ and an additional row and column in the pairwise kinematic prior $\Psi(l_i, l_j)$.

We consider the head as the root of the kinematic tree and set the pairwise term so that if a node is occluded, then all its children must be occluded as well. We consider the head to be always visible and give it no occlusion state.

We use the occlusion prediction $\mathcal{O}_i^p$ to set the energy of the occlusion state in the extended MPS model (and the energies of the corresponding row/columns in the pairwise term). Therefore, the MPS model sees $\mathcal{O}_i^p$ as a prior for a part to be occluded.

6 Inference

To find the optimal configuration of body parts in our joint MPS model (eq. (2)) we must minimize its energy also over FtB orders $\mathcal{Z}$. This is infeasible due to the factorial number of orders in the number of persons and the relatively high cost of pose estimation for a person. The techniques we propose in sec. 4 bring us closer to the goal, as they greatly reduce the number of orders to be considered. Yet, it remains inconveniently expensive to find the exact global optimum. Therefore, we show here how to perform efficient approximate optimization of eq. (2) over $\mathcal{L}$. Notice that the optimization is done only once as all FtB orders are marginalized out while computing POP (sec. 5.1).

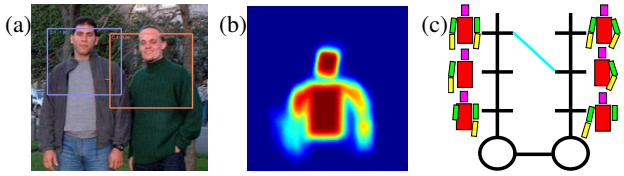

Fig. 5: **Inference.** (a) an inference example, (b) a stack of samples drawn from the joint probability of configuration of the left person, (c) puppet state-space graphical model (eq. (8)), the lowest energy configuration according to the joint model is marked by the cyan line.

Person-level model. The key idea is to rewrite eq. (2) on a coarser level, where a node is a person rather than a body part:

$$P(\mathcal{L} \mid I, \Theta) \propto \exp\left(\sum_{u \in \mathcal{U}} \sum_{p \in \mathcal{P}} u(L^p \mid I, \Theta) + \sum_{(p,q) \in \mathcal{X}} \Omega(L^p, L^q)\right) \tag{8}$$

where $\mathcal{U}$ is the set of unary terms related to one person and Ω is the inter-person exclusion term (as ω in eq. (2) but now defined on the person level). A single state for a node in eq. (2) was a particular location of a body part of a person, whereas in eq. (8) it is a spatial configuration of all body parts of a person - a *puppet* (Fig. 5c). All terms of eq. (2) relating to one person become unary terms u in eq. (8) (also the pairwise kinematic prior $\Psi(l_i, l_j)$ between parts). The set of model edges corresponds to the interaction graph $\mathcal{X}$. The exclusion term Ω is detailed in the next section, as it can be computed very efficiently by exploiting the properties of the puppet-space and of the inference algorithm below.

Efficient inference. This remodeling of the problem enables to take advantage of two important facts: (i) the occlusion probabilities POP/BOP depend on the output from the person detector only; (ii) the number of FtB orders that affects the occlusion probabilities is much smaller than the number of FtB orders affecting pose estimation (sec. 5.1). Based on these facts, we design the following approximate inference on the joint MPS model:

(1) *Compute the occlusion probability $\mathcal{O}_i^p$ for every part of every person* by combining POP (eq. (5)) and BOP (eq. (7)). As at test time the FtB order is not given, we marginalize it out when computing POP (it does not affect BOP anyway):

$$\mathcal{O}_i^p = \max\left\{P(l_i^p = \text{o} \mid D^p, \mathcal{B}, \zeta_i^{\mathcal{B}}), \quad \tfrac{1}{|\mathcal{Z}^p|}\sum\nolimits_{Z \in \mathcal{Z}^p} P(l_i^p = \text{o} \mid Z, \mathcal{T})\right\} \tag{9}$$

(2) *Sample a small set of candidate puppets $\mathcal{S}^p$ for every person*. We reduce the joint model to contain only the image likelihood Φ and kinematic prior Ψ terms for one person p and plug $\mathcal{O}^p$ in the occlusion states as described in sec. 5.3. Then we sample 1000 puppets according to the posterior of the reduced model - *a proposal distribution* (Fig. 5b). When implemented efficiently, this sampling has computational complexity similar to finding the best puppet in the 1PS model eq. (1).

(3) *Optimize the joint model*. We setup the state-space of each person p in the joint model (eq. (8)) to contain only the sampled puppets $\mathcal{S}^p$ (Fig. 5c). As the interaction graph may contain loops, we use TRW-S [19] to run inference in this model. The computation time of this operation is negligible.

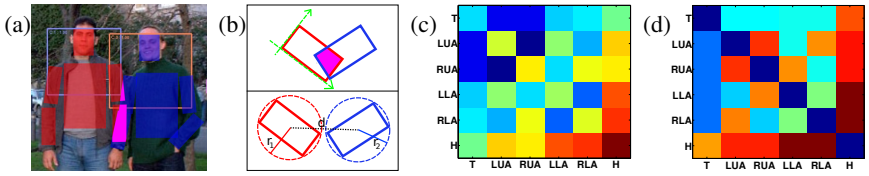

Fig. 6: **(a)-(c) Exclusion between people.** (a) Blue and red overlays show likely body configuration for a single person HPE algorithm [1], magenta depicts image areas covered by both persons. (b) top: fast IoU between two rectangles using Sutherland-Hodgman clipping algorithm [20], bottom: bound on intersection of two rectangles, (c) limb-pairs contributions into the overall between people exclusion Ω. **(d) Anti double-counting.** Limb-pair contributions (sec. 8)

Additional terms. The final model eq. (8) contains additional unary terms in $\mathcal{U}$, defined on individual persons, designed to compensate flaws of the original PS formulation [5]. We explain them in more detail in sec. 8.

7 Exclusion Between People Ω, ω

We explain here the inter-people exclusion term Ω, which penalizes configurations where different people have body parts in the same image region (Fig. 6a).

We define it as $\Omega(L^p, L^q) = \sum_i \sum_j a_{ij}\omega(l_i^p, l_j^q)$ where $\omega(l_i^p, l_j^q)$ is the exclusion defined on per body part level and a_{ij} are per limb pair (i, j) weights (eq. (2)); $\omega(l_i^p, l_j^q)$ is defined as $\log(1 - \text{IoU}(l_i^p, l_j^q))$ where $\text{IoU}(l_i^p, l_j^q) \in [0, 1]$ is the area of intersection-over-union between body parts l_i^p, l_j^q of two persons p, q and a body part is approximated by a rectangle of constant aspect-ratio (Fig. 6a).

The inference approach of sec. 6 must compute the exclusion term between all pairs of body parts between all pairs of sampled puppets between all pairs of interacting people. If implemented directly this requires $|\mathcal{S}|^2 * |L|^2 * |\mathcal{X}|$ IoU computations, where $|\mathcal{S}|$ is the number of puppet samples, $|L|$ the number of body parts, and $|\mathcal{X}|$ the number of edges in the interaction graph $\mathcal{X}$. Although one IoU can be computed efficiently based on the Sutherland-Hodgman clipping algorithm [20] (Fig. 6b top), doing it for all pairs is very expensive.

We can drastically reduce the number of IoU computations without doing any approximation by observing that two rectangles i, j can have non-zero IoU only if the distance $d(c_i, c_j)$ between their centers is smaller then the sum of the radii r_i, r_j of their circumscribing circles (Fig. 6b bottom). Therefore, we compute this bound for all pairs of rectangles and then only compute IoU for the small minority with $d(r_i, r_j) < r_i + r_j$. The cost of computing the bound is negligible compared to the cost of IoU.

Learning weights a_{ij}. In our model (eq. (2)), the exclusion terms between different pairs of body parts (i, j) between two persons are combined in a sum weighted by a_{ij}. We learn these weights from the training set as follows. For each pair of parts (i, j), we compute the average IoU m_{ij} between all pairs of interacting ground-truth stickmen (to avoid a bias, only pairs of not occluded parts contribute to the average). We then set $a_{ij} = 1 - m_{ij}$. As the arm of a person can be partially in front of her neighbor's torso and yet both are visible, we want to penalize this situation little. Instead, we want

to exclude that the arm of a person can overlap with the head of another. The learned weights follow these intuitions, and give a high weight for head-arm overlaps but lower weight to arm-torso overlaps (Fig. 6c).

8 Additional Single-Person Cues

We include in our joint model (eq. (8)) additional terms defined on individual persons, designed to compensate shortcomings of a plain PS model (sec. 3.1).

Anti Double-Counting Γ. The original PS formulation (eq. (1)) lacks a mechanism for discouraging two parts of the same person from explaining the same image region. This *double-counting* typically happens between left/right arms or legs. Several methods were proposed to tackle this problem including non-tree models [2, 17] and sequential image likelihood removal [16].

Interestingly, we can easily incorporate anti double-counting penalties in our model simply by adding a term analog to the inter-people exclusion ω, but now between pairs of body parts of the same person. This is more principled than manually adding dependencies between parts incline to double-counting [2], as it enables to learn a weight for every part pair (in the same way as for ω, sec. 7). The learned weights nicely lead to the desired behavior, e.g. high weights between all combinations of upper and lower arms, but low weights between arms and torso, resulting in a model that does not penalize configurations where the arms are in front of the torso (Fig. 6d).

Foreground-Fill Λ Foreground-fill Λ encourages configurations of body parts colored differently then the background, similar to [21, 22]. It gives intermediate energies when body parts are occluded.

Symmetric Arm Appearance Υ. Symmetric Arm Appearance Υ encourages configurations where the left and right upper arms (as well as the left and right lower arms) have similar appearance. If one in a the pair is occluded, then it gives an intermediate energy to both.

9 Experiments and Conclusions

We present a comprehensive evaluation of (i) the algorithm's complexity drop when using the FtB orders reductions (sec. 4); (ii) the ability of the method to predict which body parts are occluded (sec. 5); (iii) the pose estimation accuracy of our joint MPS model, compared to a state-of-the-art 1PS estimator [1] (sec. 6).

We split the group photo dataset into training (first 350 images) and testing (remaining 175 images). We use the training set to learn all the parameters. The test set is used for evaluating both the occlusion predictor and pose estimation performance.

Automatic parameter setting In order to incorporate the occlusion probabilities in the MPS model (sec. 5.3) we need just two parameters (the scaling for the unary energy of the occlusion state and the real-to-occlusion state transition energy in the pairwise terms). We search over a grid of values and retain those maximizing the performance of the HPE algorithm (i.e. the Percentage of Correct body Parts) on the training set (sec 9.(iii)). The optimal weights between the various terms of MPS (eq. (8)) are learned

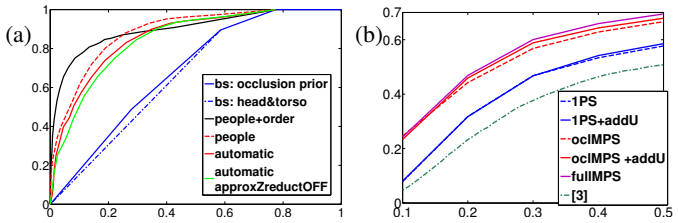

Fig. 7: **Evaluation.** (a) ROC curves for binary occlusion classification (y-axis = true-positive rate, x-axis = false-positive rate). **Baselines:** *occlusion prior* - constant OP for each part set to the part's frequency of occlusion over the training set; *head&torso* - head and torso always visible and all other parts always occluded. **Modes for our method:** *people+order* - ground-truth person detections and order Z given; *people* - only ground-truth detections given; *automatic* - true test scenario with nothing given; *automatic approxZreductOFF* - as *automatic* but without using the heuristics for reducing the number of FtB orders.
(b) PCP curves for pose estimation: *1PS* - [1]+ [3], *1PS+addU* - [1]+ [3] + additional single person terms $\Gamma\Lambda\Upsilon$ (sec. 8); *oclMPS* - the lowest energy puppet sampled from the *proposal distribution* of our MPS model including occlusion probabilities (sec. 6); *oclMPS+addU* - *oclMPS* with $\Gamma\Lambda\Upsilon$; *fullMPS* - *oclMPS* with $\Gamma\Lambda\Upsilon$ and the inter-people exclusion term Ω (the full multi person model).

using a constraint generation algorithm inspired by [23], again to maximize PCP. In the complete model, we train both types of parameters jointly (i.e. by running the constrain generation algorithm at every point on the grid). All other parameters of our model are learned as described in the respective sections 5.1, 5.2, 7, 8.

Person Detector. Since our approach relies on a person detector, we need one yielding high detection rates and low false positive rates, also on images where people are only visible from the waist up (Fig. 8). For this we combine a face detector [24] with the upper and full human body models in the detection framework of [25]. This detector achieves 86% detection-rate at 0.5 false positives per image on our group photo dataset.

(i) FtB Orders Reduction. Without any of the FtB orders reductions proposed in section 4, the median number of required pose estimations per image over the entire dataset is 600. When utilizing the exact reductions (sec. 4.1) this decreases to 80, and with also approximate reductions to 48 (sec. 4.2).

(ii) Occlusion Prediction (OP). Given a test image, we compute occlusion probabilities using eq. (9). This estimates a probability of occlusion $\mathcal{O}_i^p$ for each person p and body part i in the image. We evaluate the quality of this estimation by using it to classify body parts as occluded or not-occluded. For this, we draw ROC curves by thresholding the probability at increasing values in $[0, 1]$ (Fig. 7a).

Fig. 7a shows the performance of our method in 3 modes, differing in the amount of information given to the algorithm, and a few intuitive baselines to compare against. Our OP predictor in all modes clearly outperforms even the strongest baseline (*occlusion prior*). The influence of the order marginalization (eq (9)) on the prediction quality is visible by comparing *people+order* to *people*. This approximation only causes a modest performance drop. The influence of using our automatic (and imperfect) person detector can be seen by comparing *people* to *automatic*. The performance of the occlusion predictor decreases only marginally compared to using ground-truth detections.

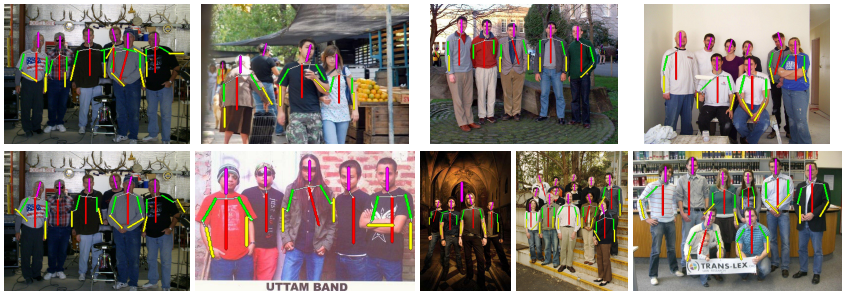

Fig. 8: **Results.** First column: top - results of single person model *1PS+addU*, bottom - full multi-person model *fullMPS*. Other columns: more results returned by our full model.

Finally, comparing *automatic* and *automatic-approxZreductOFF* demonstrates that the heuristics for reducing the number of FtB orders (sec. 4.2) do help the OP predictor. The good performance of the *automatic* mode shows that our predictor can reliably estimate the occlusion probabilities of persons' body parts given just their (automatically detected) image locations.

(iii) Pose Estimation. We evaluate the impact of our joint MPS model, which explicitly models interactions between people, on pose estimation performance. For each body part of every person, our method returns a line segment or deems the part as occluded. We evaluate performance using the framework of [1] (on-line) modified to account for occlusions. The performance is measured by average PCP (Percentage of Correctly estimated body Parts) over all persons correctly localized by the person detector. An estimated part is considered correct if its segment endpoints lie within a fraction of the length (pcp-threshold) of the ground-truth segment from their annotated location. An occluded body part is considered correct only if it is also occluded in the ground-truth. Fig. 7b shows PCP performance for pcp-threshold in $[0.1, 0.5]$.

We compare to, and based our model on, the single-person HPE of [1] with added the excellent body part models of [3] (*1PS*). This achieves sharper part posterior marginals than [1] alone, which is beneficial to our sampling procedure (sec. 6). We also compare to the complete HPE of [3] using the code released by the authors[1], initialized from the same person detections as 1PS and MPS. As Fig. 7b shows, extending *1PS* into our MPS by incorporating the occlusion probability prediction brings a substantial gain of 10% PCP (*1PS* vs *oclMPS*). Further adding the additional unary cues improves performance by another 2% (*oclMPS+addU*). Adding also the inter-people exclusion term Ω brings another 2% improvement (*fullMPS*). This shows that all components presented in this paper are valuable for good performance in group photos. Overall, our full multi-person model improves over *1PS* by 15% (at pcp-threshold = 0.2). Note how already our *1PS* outperforms [3] on this dataset. Fig. 8 shows some qualitative results (illustrations for the entire test set are available at [4]).

Conclusions We presented a novel multi-person pose estimation framework that explicitly models interactions between people and estimates their poses jointly. Both occlusion probability and pose estimation evaluations confirm our claims that joint multi-

[1] We thank Andriluka and Schiele for help in evaluating their approach on our dataset.

person pose estimation in the group photo scenario is beneficial over estimating the pose of every person independently.

References

1. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC. (2009)
2. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: CVPR. (2009)
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009)
4. : (http://www.vision.ee.ethz.ch/∼calvin)
5. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV **61** (2005)
6. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. (2006)
7. Sigal, L., Black, M.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: CVPR. Volume 2. (2006) 2041–2048
8. Lan, X., Huttenlocher, D.P.: A unified spatio-temporal articulated model for tracking. In: CVPR. Volume 1. (2004) 722–729
9. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. PAMI **28** (2006) 44–58
10. Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. In: ECCV. (2008)
11. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. IJCV **75** (2007) 247–266
12. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR. (2008)
13. Ess, A., Leibe, B., Schindler, K., Gool, L.V.: Robust multi-person tracking from a mobile platform. PAMI **31(10)** (2009) 1831–1846
14. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV. (2009)
15. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML. (2001)
16. Buehler, P., Everinghan, M., Huttenlocher, D., Zisserman, A.: Long term arm and hand tracking for continuous sign language tv broadcasts. In: BMVC. (2008)
17. Lan, X., Huttenlocher, D.: Beyond trees: Common-factor models for 2D human pose recovery. In: ICCV. Volume 1. (2005)
18. Gammeter, S., Ess, A., Jaeggli, T., Schindler, K., Van Gool, L.: Articulated multi-body tracking under egomotion. In: ECCV. (2008)
19. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI **28** (2006) 1568–1583
20. Sutherland, I., Hodgman, G.: Re-entrant polygon clipping. In: Communications of the ACM. (1974)
21. Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Long term arm and hand tracking for continuous sign language TV broadcasts. In: BMVC. (2008)
22. Jiang, H.: Human pose estimation using consistent max-covering. In: ICCV. (2009)
23. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR **6** (2005) 1453–1484
24. Froba, B., Ernst, A.: Face detection with the modified census transform. In: IEEE international conference on automatic face and gesture recognition. (2004)
25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2009) in press.