Human Context: Modeling human-human interactions for monocular 3D pose estimation

Mykhaylo Andriluka[†] Leonid Sigal[‡]

[†]Max Planck Institute for Informatics, Saarbrücken, Germany
[‡]Disney Research, Pittsburgh, USA

Abstract. Automatic recovery of 3d pose of multiple interacting subjects from unconstrained monocular image sequence is a challenging and largely unaddressed problem. We observe, however, that by tacking the interactions explicitly into account, treating individual subjects as mutual "context" for one another, performance on this challenging problem can be improved. Building on this observation, in this paper we develop an approach that first jointly estimates 2d poses of people using multi-person extension of the pictorial structures model and then lifts them to 3d. We illustrate effectiveness of our method on a new dataset of dancing couples and challenging videos from dance competitions.

1 Introduction and Related Work

Human pose estimation and tracking have been a focal point of research in vision for well over 20 years. Despite much progress, most research focuses on estimation of pose for single well separated subjects. Occlusions and part-person ambiguities, that arise when two people are in close proximity to one another, make the problem of pose inference for interacting people a very challenging task. We argue that the knowledge of pose for one person, involved in an interaction, can help in resolving the pose ambiguities for the other, and vice versa; in other words, two people involved in an interaction (e.g., dance, handshake) can serve as mutual context for one another.

Recent tracking-by-detection (TbD) methods [1] have shown impressive results in real world scenarios, but with a few exceptions [2], are restricted to individual people performing simple cyclic activities (e.g., walking). Despite successes, TbD methods ignore *all* contextual information provided by the scene, objects and other people in it. As a result, in close interactions, independent pose estimates, ignorant of one another, compete and significantly degrade the overall performance. In contrast, [3, 4] argue that context is an important cue for resolving pose ambiguities, and show that human-object context improves pose estimation. Inspired by these recent advances, in this paper, we advocate the use of human-to-human context to facilitate 3d pose inference of multiple interacting people. This differs from prior work where interactions are either ignored [1, 5] or are only considered in the form of partial occlusions [6].

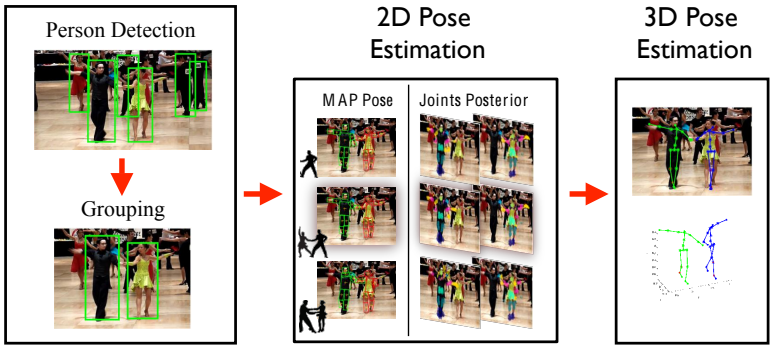


Fig. 1. Complete overview for the proposed method illustrated through an actual result obtained with our framework on a real competition dance sequence.

Contributions: Our key contribution is an automatic framework for estimating 3d pose of interacting people performing complex activities from monocular observations. In developing this hierarchical framework we incorporate and analyze the role of modeling interactions, in the form of human-human context, at all levels. We introduce a novel multi-aspect flexible pictorial structure (MaFPS) model to facilitate joint 2d inference over pairs of people. The aspects encode the modes of interaction and result in non-tree structured model for which we introduce efficient approximate inference. We show results on challenging monocular sequences that contain dancing couples. The couples in our dataset (to be made publicly available) appear on cluttered background and are engaged in considerably more challenging and diverse set of motions than is typical of state-of-the-art methods (e.g., walking people in street scenes [1]).

2 Method

To achieve our goal of estimating 3d poses of interacting people in images of realistic complexity we leverage the recent progress in people detection, tracking and pose estimation. We rely on layered hierarchical framework that combines *bottom-up* and *top-down* information. At a high level, the approach can be expressed as a generative model for 3d human pose that combines rich *bottom-up* likelihood with a *top-down* prior:

$$p(Y_1, Y_2|I) \propto p(I|Y_1, Y_2)p(Y_1, Y_2), \tag{1}$$

where I is a set of image observations and $Y_i = \{y_i, d_i\}$ correspond to the parameters of 3d pose, y_i, and camera parameters required to project the pose into the image, d_i, for the i-th person. The inference amounts to searching for a *maximum a posteriori* (MAP) estimate of Y_1 and Y_2 with respect to the model in Eq. (1). In our model we incorporate interactions at different levels and take them into account both in the prior and likelihood terms.

The 3d pose prior, $p(Y_1, Y_2)$, captures the activity-induced correlations between poses of the two subjects and also models the relative orientation and position of the subjects with respect to one another. We rely on the Gaussian Process Dynamical Model (GPDM) [7] and learn parameters of the prior model from the motion capture data. The use of GPDM also allows us to learn the model of dynamics for stitching individual 3d poses together when tracking.

To avoid depth and observation ambiguities, typical in monocular inference, we define a rich likelihood model, $p(I|Y_1, Y_2)$, that encodes consistency between the projected 3d poses and 2d posteriors over body part locations. Characterizing 2d posteriors, and hence the likelihood, involves inference over 2d pose of the body that takes into account spatial configuration of parts and discriminatively learned part appearances. For further robustness and temporal consistency of 2d pose estimates, we condition the 2d model on position and scale of person detections.

Formally, we introduce a set of auxiliary variables $L_i = \{L_i^t\}$ which correspond to 2d configuration of the body and $D_i = \{D_i^t\}$ which correspond to position and scale of the i-th subject in each frame of the sequence; with t being the frame index. We make a first-order Markov assumption over D_i and assume conditional independence of 2d poses L_i given positions of people in each frame so that:

$$p(L_1, L_2, D_1, D_2|I) = \prod_t p(L_1^t, L_2^t|D_1^t, D_2^t, I) \\ p(D_1^t, D_2^t|I)p(D_1^t, D_2^t|D_1^{t-1}, D_2^{t-1}). \quad (2)$$

The posterior, $p(L_1^t, L_2^t|D_1^t, D_2^t, I)$, on the right-hand side of the equation corresponds to the joint multi-aspect flexible pictorial structure (MaFPS) model for the two interacting subjects, which we describe in detail in Sec. 2.2.

To properly account for uncertainty in the 2d pose estimates we define the likelihood in Eq. (1) by evaluating the projection of the 3d pose under the posterior distribution given by Eq. (2). We define the likelihood of the pose sequence as:

$$p(I|Y_1, Y_2) = \prod_{t,n} p_{L1,n}(\pi_n(Y_1^t))p_{L2,n}(\pi_n(Y_2^t)), \quad (3)$$

where $p_{L1,n}$ denotes the marginal posterior distribution of the n-th body part of the configuration L_1 and $\pi_n(Y_1^t)$ corresponds to the projection of the n-th part into the image.

In order to obtain a MAP estimate for the posterior in Eq. (1) we adopt a multi-stage approach in which we first infer auxiliary variables D_i and L_i and then infer the 3d poses using local optimization, while keeping the auxiliary variables fixed. To further simplify inference we make an observation that in most sequences person detection and our tracking and grouping procedure are reliable enough to allow us to infer D_i first by obtaining modes of $p(D_1^1, D_2^1)p(D_1^1, D_2^1|I)\prod_{t=2}^{T} p(D_1^t, D_2^t|I)p(D_1^t, D_2^t|D_1^{t-1}, D_2^{t-1})$ before inferring posterior over L_i conditioned on D_i.

2.1 Person Detections and Grouping

As a first step towards inferring 3d poses of people we proceed with recovering positions of potentially interacting people and tracking them over time. This corresponds to estimating the values of the variables D_1 and D_2 in Eq. (2).

We employ the tracking-by-detection approach described in [1] and find tracks of people by connecting hypothesis obtained with the person detector [8]. We then identify pairs of tracks that maintain close proximity to one another over all frames, and use them as estimates for D_1 and D_2.

Denoting the set of people detection hypothesis in frame t by h^t and a track corresponding to a sequence of selected[1] hypothesis over T frames by $h_\alpha = \{h^t_{\alpha_t}; t = 1, \ldots, T\}$, we are looking for two such tracks that are both consistent over time (with respect to position and scale) and at the same time are likely to correspond to the interacting persons. In this work we use spatial proximity of detections as the main cue for interaction and focus of finding two tracks that maintain close proximity to one another over all frames. Ideally, we would like to jointly estimate the assignment of hypothesis to both tracks. However, we found that the following greedy procedure works well in practice. We first identify tracks of individual people by optimizing the following objective with Viterbi-decoding[2]:

$$p(h_\alpha) = p(h^1_{\alpha_1}) \prod_{t=2}^{T} p(h^t_{\alpha_t}) p(h^t_{\alpha_t}, h^t_{\alpha_{t-1}}), \tag{4}$$

where the unary terms correspond to the confidence score of the person detector and the binary terms are zero-mean Gaussian with respect to the relative distance in position and scale.

Given a set of single person tracks we associate two tracks using the closest distance. We define the distance between two tracks h_{α_1} and h_{α_2} as the average distance between corresponding track hypothesis:

$$D(h_{\alpha_1}, h_{\alpha_2}) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \|x(h_{\alpha_{1,t}}) - x(h_{\alpha_{1,t}})\|, \tag{5}$$

where $x(\cdot)$ is the centroid of detection and t_1 and t_2 are the first and last frame contained in both tracks. We only link two tracks with distance less than a predefined threshold and link tracks with the smallest distance if multiple tracks are sufficiently close to each other. Finally we merge all the tracks that were assigned to each other into disjoint groups and independently infer poses of people for each such group.

[1] The index of selected hypothesis at frame t is denoted by α_t.

[2] We extract multiple tracks by stopping the Viterbi inference once at least one of the transition probabilities between frames falls below a predefined threshold and repeat the optimization using hypothesis have not been assigned to a track yet. Following the initial optimization we filter all the tracks that are too short or have too low score.

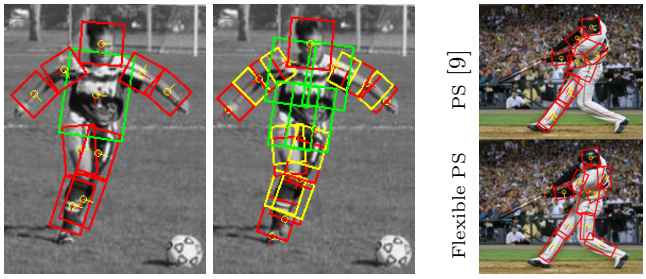


Fig. 2. Flexible body model: Traditional 10-part PS model [9] on left, and the proposed 22-part flexible PS model in the middle; half-limbs for all body parts, that are allowed to slide with respect to one another, are illustrate in red and yellow; torso is modeled using 4 flexible parts in green.

2.2 2D Pose Estimation

In our approach we define the likelihood of 3d body pose using estimates of the 2d projections of the body joints. This allows us to leverage the recent results in 2d pose estimation and rely on the discriminativly trained body part detectors and robust local appearance features [9–11].

Basic pictorial structures model: We denote the 2d configuration of subject i in frame t by by $L_i^t = (l_{i0}^t, \ldots, l_{iN}^t)$, where $l_{ij}^t = (x_{ij}^t, \theta_{ij}^t, s_{ij}^t)$ correspond to the image position, rotation and scale of the j-th body part; $N = 10$ denotes the total number of parts, which are traditionally chosen to correspond to torso, head, lower and upper legs, forearms and upper arms [12, 11, 9]. Given the image evidence I^t the posterior over 2d body configuration is proportional to the product of likelihood and prior terms: $p(L_i^t|I) \propto p(I^t|L_i^t)p(L_i^t)$.

In the absence of interactions, one can rely on the tree-structured pictorial structures model to represent the posterior over L_i^t. In addition, we assume that the likelihood term factorizes into the product of individual part likelihoods $p(I^t|l_i^t)$, which we model using boosted body part detectors as in [9]. Since representing joint posterior over 2d configurations is intractable, for 3d likelihood in Eq. (3), we approximate this posterior as a product over posterior marginals of individual body parts: $p(L_i^t|I) \approx \prod_n p(l_{in}^t|I)$.

Flexible pictorial structure model: The use of traditional 10-part PS model commonly employed in the literature [3, 9, 11, 13] presents a number of challenges when applied in our setting. Focusing on individual people, traditional PS model: (i) does not properly model foreshortening of parts because parts are represented with rigid rectangles of fixed size and (ii) is not effective in inference of poses across variety of camera views. To address (ii) view-based [1] and multi-pose [3] models have been introduced. These amount to a collection of PS models trained with view-specific or pose-specific spatial priors. In our case, where we are after jointly modeling multiple people, such mixture models will result in an exponential number of PS models. Instead, following [14, 15] we propose a

more flexible extension that is able to deal with both (i) foreshortening and (ii) diverse camera views using one coherent PS model.

In our model we represent human body with an extended 22-part pictorial structures model shown on Fig. 2. In this model each of the body limbs is represented with two parts (half limbs) attached to the ends of the limb. These parts are allowed to slide along the part axis with respect to each other, capturing the foreshortening of the limb. In addition, we model the torso with 4 parts attached to shoulders and hips so that the model is capable of representing various torso orientations by shifting this parts with respect to each other. The 4 torso parts are connected in a star-shape pattern. In Fig. 2 (right) shows example of body configuration inferred with our 22-part model on the "People" dataset from [11]. Note, that the model could properly adjust to the orientation of the torso which also resulted in better estimate of the other body parts (for more see Fig. 3).

Conditioning on person detection: One of the challenges of recovering pose of multiple people using pictorial structures is *double-counting*. The double-counting problem, in this context, refers to the fact that since the two subjects are conditionally independent[3] the model is likely to find location of the two subjects one on top of another situated on the most salient person in the image. While we use posterior for the 3d pose likelihood, weak modes that appear on the less salient subject still cause problems. To address the double-counting issue one can use: (1) repulsive potentials that penalize substantial overlap between parts [16], or (2) resort to pixel-wise occlusion reasoning by introducing and marginalizing over image layers [17].

We take a different approach, that stems from an observation that our person detection and grouping works remarkably well in separating out interacting couples. To ensure that body-part hypothesis of both subjects are well represented and have substantial probability mass in the posteriors we condition 2d pose inference process on the estimated positions of both people given by D_1^t and D_2^t. This bares some similarity to the progressive search approach of [18]. We assume that positions of body parts are conditionally independent of D_i^t given the position of the root node l_{i0}^t, so that conditioning the model on D_i^t corresponds to replacing the uniform prior on position of root part $p(l_{i0}^t)$ with conditional distribution $p(l_{i0}^t|D_i^t)$, which we assume to be Gaussian centered on the image position given by D_i^t.

Multi-person pictorial structure model: Our joint model incorporates interactions as a form of constraints on positions of the body parts of the interacting people. Clearly, depending on the type of the interaction, positions of different body parts of people involved will be dependent on each other. For example, during the waltz arms of both subjects are typically close together, whereas during the crossover motion in cha-cha partners will only hold one hand. In order to accommodate these modes of interaction we introduce an interaction aspect variable a^t which will specify the mode of interaction for the frame t.

[3] Same is true for models with weak conditional dependence between parts as those imposed by interaction aspects.

Given the interaction aspect, the joint posterior distribution over configurations of interacting people is given by

$$p(L_1^t, L_2^t | I^t, a^t) \propto p(I^t | L_1^t) p(I^t | L_2^t) p(L_1^t, L_2^t | a^t), \tag{6}$$

where we have assumed independence of the appearance of both people allowing us to factorize the joint likelihood into the product of likelihoods of each person. The joint prior on configurations of people is given by

$$p(L_1^t, L_2^t | a^t) = \prod_{i=1}^{2} p(L_i^t) \prod_{(n,m) \in W} p(l_{1n}^t, l_{2m}^t)^{a_{nm}^t}, \tag{7}$$

where $p(L_i^t)$ is a tree structured prior, W is a set of edges between interacting parts and a_{nm}^t is a binary variable that turns the corresponding potential on and off depending on the type of interaction. The interaction potential are given by $p(l_{1n}^t, l_{2m}^t) = \mathcal{N}(x_{1n}^t - x_{2m}^t | \mu_{nm}, \Sigma_{nm})$, and specify the preferred relative distance between the positions of the interacting parts.

We model 4 aspects that correspond to hand holding; these include: (1) no hand holding, (2) left hand of one subject holding right hand of the other subject, (3) left hand of one subject holding left hand of the other subject, and (4) two-hand hold. These interaction aspects are motivated by our target application of looking at dancing couples. That said, we want to emphasize that our joint 2d pose estimation model is general and applicable to most interactions.

Inference in the multi-person model: Modeling dependencies between subjects comes at a cost of more expensive inference. In tree-structured model inference can be made efficient with the use of convolution [12]. The dependencies between subjects can introduce loops (as is the case with the two-hand hold that we model) which makes exact inference prohibitively expensive. In order to make the inference tractable, we rely on the following two-stage procedure. In the first stage we ignore interaction factors and perform the inference in the tree-structured model only. We then sample a fixed number[4] of positions for each body part of each subject from the marginal of the tree-structured model, and repeat the inference with the full model using the sampled positions as a state-space. This inference procedure relates to branch and bound method proposed by Tian et al. [19].

2.3 3D Pose Estimation

To estimate 3d poses of interacting people we rely on a prior model that incorporates three types of dependencies between subjects: dependencies between body pose, relative orientation and position between subjects. To capture these dependencies we rely on a joint Gaussian process dynamical model (GPDM) [7] trained using motion capture sequence of couples dancing. We train one GPDM model for each dance move, performed by 3 to 4 couples.

[4] In our implementation we sample positions 2000 times for each part and remove the repeating samples.

Joint 3d prior model: Typically, GPDM is used to learn a latent model of motion for a single subject. In our case, we are interested in learning a joint model over two interacting people. To do so, we express our training samples as $Y = (Y_1, Y_2, Y_\delta, Y_{\theta_1}, Y_{\theta_{1\to 2}})$, where Y_1 and Y_2 are 3d poses of the two subjects, Y_δ is relative position of subject 2 with respect to 1, Y_{θ_1} is the root orientation of first subject in a canonical frame of reference and $Y_{\theta_{1\to 2}}$ is the orientation of subject 2 with respect to 1. For convenience, we collect all training samples in our dataset $\mathcal{D}$ into $\mathbf{Y} = \{Y \in \mathcal{D}\}$. We learn a joint GPDM model by minimizing negative log of posterior $p(\mathbf{X}, \bar{\alpha}, \bar{\beta}|\mathbf{Y})$ with respect to latent positions $\mathbf{X} \in \mathbb{R}^{d\times|\mathcal{D}|}$ and hyperparameters $\bar{\alpha}$ and $\bar{\beta}$ [5].

MAP pose inference: In our approach the 3d pose estimation corresponds to finding the most likely values for Y_1 and Y_2 and the parameters of their projection into the image, Q, given the set of image observations, I, and the GPDM prior model learned above – $\mathcal{M}_{GPDM} = (\mathbf{X}, \mathbf{Y}, \bar{\alpha}, \bar{\beta})$.

The projection parameters are given by $Q = \{r^t, \delta^t, \gamma_1^t, \gamma_t^2, \phi^t\}$, where r^t is the position of the first person in frame t, the γ_1^t and γ_t^2 are the scales of first and second person, ϕ^t is the absolute rotation of the canonical reference frame for the couple (with respect to which Y_{θ_1} is defined) and δ^t is the deviation in the image position of the second person with respect to the position predicted by the projection of Y_δ into the image. Note, δ^t allows us to deviate from the GPDM prior in order to generalize across closer and more distant embraces that we were not able to explicitly model using few mocap sequence samples.

Assuming there is negligible uncertainty in the reconstruction mapping [20], the 3d pose of both subjects in the canonical space, given a latent pose X, is given by the mean of the Gaussian process: $\mu_Y(X) = \mu + \mathbf{Y}\mathbf{K}_Y^{-1}\mathbf{k}_Y(X)$ where $\mathbf{K}_Y^{-1}$ is the inverse of a kernel matrix, and $\mathbf{k}_Y(X)$ is a kernel vector computed between training points and the latent position X. With this observation the likelihood in Eq. (1) can be expressed directly as a function of latent position X and projection parameters Q. With slight abuse of notation, we can hence re-write Eq. (1) as:

$$p(\mu_Y(X), Q|I) \propto p(I|\mu_Y(X), Q)p(Y|X, \mathcal{M}_{GPDM}) \tag{8}$$

where $p(Y|X, \mathcal{M}_{GPDM}) = \frac{d}{2}\ln\sigma^2(X) + \frac{1}{2}\|X\|^2$ and $\sigma^2(X)$ is a covariance of a GPDM model defined as $\sigma^2(X) = k_Y(X, X) - \mathbf{k}_Y(X)^T\mathbf{K}_Y^{-1}\mathbf{k}_Y(X)$.

We approach the inference by directly optimizing Eq. (8) with respect to X and Q using gradient-based continuous optimization (scaled conjugate gradients). In order to define the gradients of the likelihood function Eq. (3) we represent the posteriors of the 2d configurations L_i using the kernel density estimate given by $p_{L_1,n}(l) = \sum_k w_k exp(\|l - l_{nk}\|) + \epsilon_0$, where l_{nk} are the samples from the posterior of part n used in the second stage of the inference procedure described in Sec. 2.2 and w_k are the value of posterior distribution for this sample; $\epsilon_0 = 0.02$ is a uniform outlier probability to make the likelihood robust.

[5] For details of learning we refer the reader to [7].

Method	Torso	Upper leg	Lower leg	Upper arm	Forearm	Head	Total
Cond. PS [9] (10 parts)	96.1	88.2 89.5	72.4 64.5	26.3 23.7	18.4 13.2	72.4	56.4
Flexible Cond. PS (22 parts)	**100.0**	**92.1 96.1**	75.0 **89.5**	**46.1** 42.1	32.9 25.0	**96.1**	69.4
MaFPS (joint model, 44 parts)	**100.0**	**92.1 96.1**	**76.3 89.5**	**46.1 47.4**	**39.5 27.6**	**96.1**	**71.0**

Table 1. Quantitative performance of 2d pose estimation: three models are compared in all cases conditioned on the person detections. Numbers indicate percentage of body parts correctly found by the model.

Implementation details: We found that good initialization is important for quick convergence of the optimization. In order to obtain a set of good initial hypothesis we initialize the projection parameters Q from the estimates of people positions given by D_i and select a set of candidate poses from the training set with the highest likelihood. We also found that the converges can be significantly sped up by first optimizing the projection parameters Q while keeping the latent positions X constant, and then jointly optimizing the pose and projection parameters.

3 Experiments

We conduct experiments to validate modeling choices and show the role of encoding interactions at all levels.

Dataset: We collected a dataset of dancing couples consisting of 4 amateur couples performing latin ballroom, mainly *cha-cha*. We first recorded mocap of the subjects for training of GPDM models in mocap suites and video in their natural clothes using 2 synchronized video cameras. Video recordings were then annotated with 3d poses by clicking on joints in two views and optimizing 3d pose to match 2d projections using a continuous optimization. We annotated every 3-rd frame of the selected 3 sequences (corresponding to 3 common cha-cha moves[6] 60 frames each) from 2 couples. Our dataset hence comprises of 120 annotated 3d frames. We use a subset of 40 frames from two of the cameras with their 2d annotations for 2d experiments.

Learning: We learn appearance for part detectors and spatial priors for PS model from a public dataset of [13] to avoid overfitting to 8 people present in our dataset.

Error: We conduct experiments on two levels with 2d pose and 3d pose estimation. For experiments in 2d we use a well established percentage of parts correct (PCP) metric introduced in [18]. In 3d we measure error using average Euclidean distance between joints. In all cases we report error average over both interacting subjects.

Person detection: Person detector had a nearly perfect 100% true positive rate (with few deviations in scale) on our dataset of 60 frames/seq. $\times$ 3 seq. $\times$

[6] This included new yorker, underarm turn and basic enclosed. We train one GPDM for each across couples.

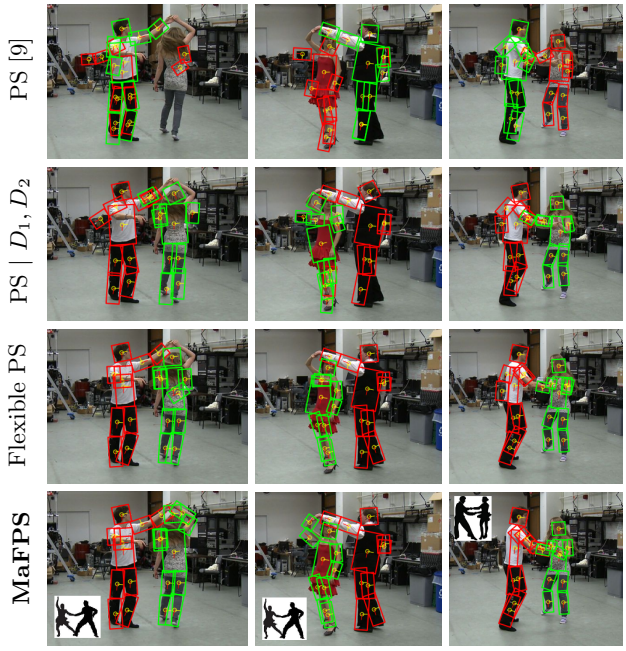
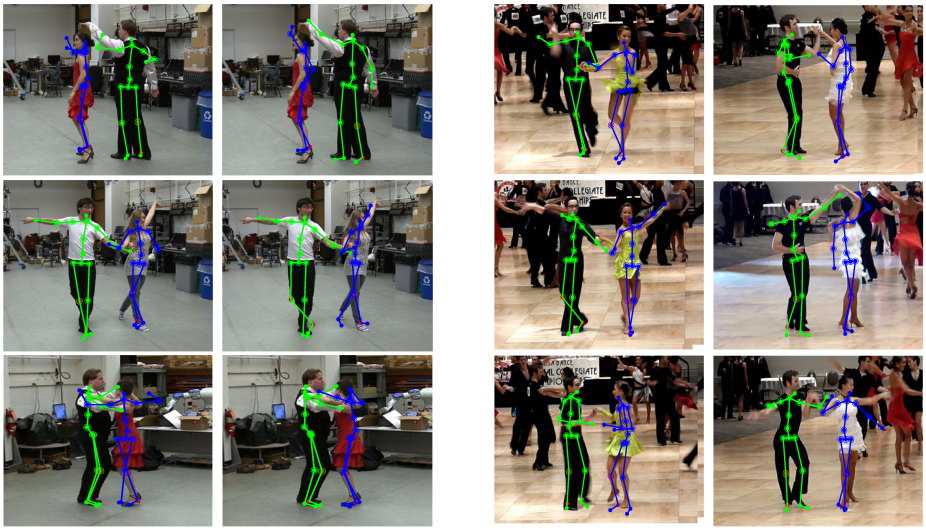


Fig. 3. 2D pose estimation: Comparison of results obtained with extensions proposed in Sec. 2.2 on our dance dataset; corresponding quantitative results are in Table 1. Top row: traditional cardboard PS [9] model – note the over-counting issue; second row: [9] model but conditioned on the person detection; third row: conditional model but with flexible parts; last row: our complete MaFPS model with interaction aspects – note the difference in the arms. Aspects chosen for the image are illustrated by inlay icons.

2 couples $\times$ 2 views = 720 images. The false positive rate was on average $2-3$ detections per frame, but all false positives were removed during grouping. In the interest of space we forgo more formal evaluation.

2d pose estimation: We first evaluate extensions we make to the pictorial structures (PS) model in Sec. 2.1 to support 2d inference of interacting people. We illustrate typical behavior of the 4 models we consider and visual improvements that we gain by conditioning on person detection, adding flexibility and interaction aspects to the model in Fig. 3. Quantitatively, based on results in Table 1, we gain 23% by adding flexibility to parts and 2.3% on average by adding aspects. It is worth noting that even though the improvements that result from modeling interaction aspects are marginal on average, they are large for the parts that are affected by the interaction (e.g., up to 20% improvement for the forearm). We only report quantitative results for models conditioned on detections; unconditioned PS model [9] lost one of the persons completely in more then half of the frames due to double-counting (see Fig. 3 top left).



(a) (b)

Fig. 4. Results of our method: (a) Several examples of 3D poses estimated by our approach. Compared are the independent 3d prior model learned for each person individually (left) and our joint 3d prior (right), both based on the same 2d pose estimates. (b) 3D poses estimated from challenging monocular dance competition videos; notice motion blur, variations in appearance, clothing, lighting, and pose.

3d pose estimation: We also compare the role of modeling 3d pose prior jointly over the two subjects. Estimating poses of each subject independently we achieve average joint error of 25 cm. Joint model that includes interactions between subjects improves this result to 19 cm. Qualitative results are illustrated in Fig. 4 (a). The joint prior model is particularly instrumental in determining the proper global view of the subjects (see Fig. 4 (a) second and third row), and resolving depth ambiguities (see Fig. 4 (a) first row).

Real sequences: Lastly, we illustrate performance on real image sequence obtained at dance competition in Fig. 4 (b). Note, that dancers in these sequence were more professional then those in our dataset. Despite this and variations in performance style, we are still able to recover meaningful 3d motions in this extremely challenging scenario.

4 Conclusions

We explore the role of human-human context in estimating 3d pose of dancing couples from monocular video. To model interactions, we introduce a novel layered model that leverages latest advances in person detection, 2d pose estimation, and latent variable models for encoding 3d priors over pose trajectories. In the process, we introduce extensions to the traditional PS model that are

able to account for aspects of human-human interactions and better deal with foreshortening and changing viewpoint. We illustrate performance on very challenging monocular images that contain couples performing dance motions. These sequences go beyond what has been shown in state-of-the-art. In the future, we intend to look at further extending the current model to also take into account occlusions among the interacting subjects.

References

1. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR. (2010)
2. Pellegrini, S., Edd, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behaviour for multi-target tracking. In: ICCV. (2009)
3. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)
4. Kjellstrm, H., Kragic, D., Black, M.J.: Tracking people interacting with objects. In: CVPR. (2010)
5. Ionescu, C., Bo, L., Sminchisescu, C.: Structured svm for visual localization and continuous state estimation. In: ICCV. (2009)
6. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: ECCV. (2010)
7. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. PAMI **30** (2008)
8. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010)
9. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009)
10. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC. (2009)
11. Ramanan, D.: Learning to parse images of articulated objects. In: NIPS. (2006)
12. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision (2005)
13. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. (2010)
14. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR. (2011)
15. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
16. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: CVPR. (2009)
17. Sigal, L., Black, M.J.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: CVPR. (2006)
18. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)
19. Tian, T.P., Sclaroff, S.: Fast globally optimal 2d human derection with loopy graph models. In: CVPR. (2010)
20. Urtasun, R., Fleet, D., Fua, P.: 3d people tracking with gaussian process dynamical models. In: CVPR. (2006)