Human Pose Co-Estimation and Applications

Marcin Eichner and Vittorio Ferrari

Abstract—Most existing techniques for articulated human pose estimation consider each person independently. Here we tackle the problem in a new setting, coined *Human Pose Co-estimation* (PCE), where multiple persons are in a common, but unknown pose. The task of PCE is to estimate their poses jointly and to produce prototypes characterizing the shared pose. Since the poses of the individual persons should be similar to the prototype, PCE has less freedom compared to estimating each pose independently, which simplifies the problem. We demonstrate our PCE technique on two applications. The first is estimating pose of people performing the same activity synchronously, such as during aerobic, cheerleading and dancing in a group. We show that PCE improves pose estimation accuracy over estimating each person independently. The second application is learning prototype poses characterizing a pose class directly from an image search engine queried by the class name (e.g. 'lotus pose'). We show that PCE leads to better pose estimation in such images, and it learns meaningful prototypes which can be used as priors for pose estimation in novel images.

Index Terms—human pose estimation, articulated objects, multiple image correspondence, object detection

1 Introduction

Articulated human pose estimation (HPE) is the task of recovering the spatial configuration of the body parts of a person from images. Recovering the 2D body layout in a monocular setting [1, 6–8, 14, 18, 19, 22, 23, 27], without any prior knowledge of person or background appearance, is very challenging, especially from uncontrolled still images such as amateur photographs. Background clutter, high diversity in clothing appearance and poor image quality often cause HPE algorithms to fail. Despite recent progress of HPE on uncontrolled images [1, 6, 14, 23, 27, 30], the problem is still far from solved.

Many previous works estimate pose independently for each person. In this paper instead we tackle the problem in a new setting, coined *Human Pose Co-estimation* (PCE): given images with persons in a common, but unknown pose, estimate their pose *jointly over all persons*. PCE simultaneously estimates a prototype pose, characterizing the shared pose, and one pose specific to each person (fig. 1). The key idea is that the poses of the individual persons should all be *similar* to the prototype. As PCE reasons jointly over multiple persons under this similarity constraint, it effectively has less freedom than estimating pose independently for each person. This implicitly simplifies the problem and leads to better results. The similarity constraint indirectly promotes a collaboration between multiple pose estimation problems, where easier problems guide pose estimation on harder ones, for which single-person HPE would fail (e.g. when one person stands against a cleaner background than another person).

The main contribution of this paper is a technique to perform pose co-estimation. We propose several PCE models of varying level of complexity accompanied with efficient inference algorithms (sec. 3). We demonstrate our technique experimentally on two scenarios where PCE occurs naturally: (i) estimating poses of people performing the same activity

• *M. Eichner and V. Ferrari are with the Computer Vision Lab at ETH Zurich*

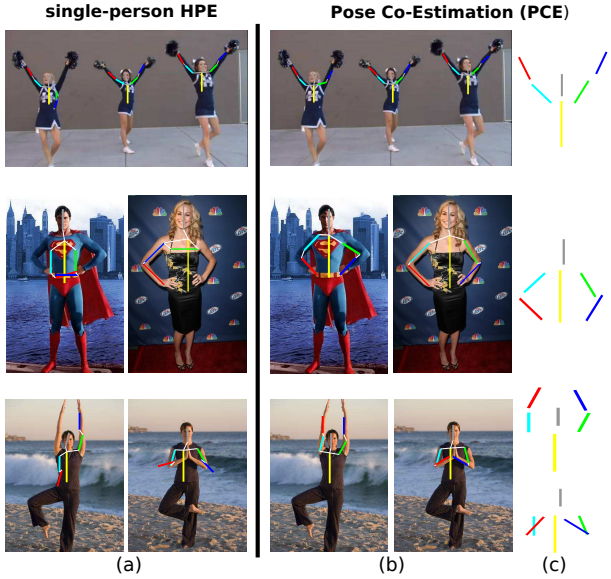


Fig. 1: **Pose Co-Estimation.** *Poses estimated by the PS model independently (a) and by our PCE technique jointly over multiple persons (b); (c) pose prototypes delivered by PCE; (top) synchronized activity scenario; (middle+bottom) learning from the web scenario for 'hips' and 'yoga tree' poses. The more complex 'yoga tree' pose is composed of 2 prototypes (discovered fully automatically).*

synchronously, such as when dancing in a group, doing aerobic or cheerleading routines (sec. 5). We show that PCE improves pose estimation accuracy compared to standard HPE techniques run independently on each person; (ii) weakly supervised learning of *pose classes* from the web (sec. 6), i.e. regions of pose space corresponding to semantically distinct configurations of body parts, such as the lotus pose. We collect images of a pose class by querying an image search engine with the name of the class, and then apply PCE to estimate the pose of all persons and to learn a prototype characterizing the pose class. Again, we show that PCE leads to better pose estimation in this scenario. Furthermore, in sec. 6.6 we use the prototypes as a *prior* in the traditional task of single-person HPE on novel images. The idea is to represent an *activity*

(e.g. *yoga*) as a collection of pose classes (e.g. *tree, lotus, warrior two*). We first show that using prototypes from several yoga poses improves pose estimation in novel yoga images. We then generalize this idea and demonstrate that using a complex prior composed of many prototypes from diverse activities improves results on two standard benchmark datasets containing a variety of poses (Buffy and ETHZ Pascal [6]).

In order to properly evaluate our techniques in the two scenarios above, we introduce a new dataset of 357 images of synchronized activities, and another dataset of 521 images of pose classes collected from the web. We release both datasets on http://www.vision.ee.ethz.ch/~calvin/.

2 Related Work

We build our pose co-estimation technique on Pictorial Structures [8], a popular paradigm for single-person HPE in still images [1, 6, 8, 19, 23]. We review it in detail below. In general, most previous works do HPE for each person independently (including works based on other paradigms than Pictorial Structures [17, 18]). One exception is [7], which models the occlusion interactions between nearby people in an image. Works on 3D pose reconstruction from 2D images [2, 15, 26, 28] often consider multiple views of a person taken at the same time. However, the initial 2D poses are either manually annotated [28] or estimated independently for every image [2, 15, 26]. To the best of our knowledge, we are the first to jointly estimate the pose of multiple persons under the assumption that they share a similar, but unknown, pose.

The idea of automatically learning pose classes from image search engines from the web is analog to previous works on learning face [3] and object classes [10, 24]. The term 'pose class' was used before in [12] in the context of retrieving poses similar to a query from a large database. However, they did not learn pose class models nor use them to improve pose estimation on novel images. Our notion of pose class can be seen as a partitioning of the continuous pose space into discrete units (classes). This is related to [2, 14], which partitioned large sets of *manually annotated* stickmen into viewpoint clusters. To the best of our knowledge, this paper is the first to demonstrate *weakly supervised* learning of *semantic* pose classes (i.e. from images, not from manual annotations).

2.1 Pictorial Structures (PS)

In the PS framework [8], a person's body parts are nodes tied together in a Conditional Random Field [20]. Parts l_p are rectangular image patches and their position is parametrized by location (x, y), orientation θ, scale s, and sometimes foreshortening [5, 8]. This constitutes the state-space of the nodes. The posterior of a configuration of parts $L = \{l_p | p = 1..B\}$ given an image I is

$$P(L \mid I) \propto \exp -\left(\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(l_p, l_q) + \sum_p \Phi_p(I|l_p) \right) \quad (1)$$

Inference in the PS model yields the MAP configuration $L^* = \arg\max_L P(L \mid I)$ [1, 8] or the posterior marginal distribution for each part [19].

In many works the model structure $\mathcal{E}$ is a tree [1, 8, 12, 19], which enables exact inference, though some works explored more complex topologies [12, 16, 25] or mixtures of trees [13, 14]. The pairwise potential $\Psi_{pq}(l_p, l_q)$ is a prior on the relative position of two parts. Usually, it embeds kinematic constraints (e.g. the upper arms must be attached to the torso) and sometimes more complex relations as parts coordination [16] or self-occlusion constraints [25]. The unary potential $\Phi_p(I|l_p)$ corresponds to the local image evidence for a part in a particular position (likelihood) and it depends on appearance models describing how parts look like. The success of PS depends strongly on having good appearance models, which limit the image positions likely to contain a part. Among the best performing models we find generic models based on gradients [1] and super-pixels [23], as well as person-specific models derived automatically from the image [6, 19].

2.2 Normalized PS state-space

As in many recent works [1, 6, 11, 12, 23], we first detect persons using a state-of-the-art detector (sec. 4). We then reduce the state-space of the PS nodes to the scale of the detection and to a range of locations around it. This makes pose estimation computationally more efficient and removes part of the background clutter. The procedure is particularly advantageous in our multi-person setting as it normalizes the state-spaces of all persons to a common reference frame.

Note how there are also recent works which perform HPE directly, without first detecting people [14, 19, 27]. However, these typically find only one person per image and assume her scale is known [14, 19, 27]. The very recent work of [30] is a notable exception to this trend.

3 Pose Co-Estimation Model (PCE)

Let $\mathcal{I} = \{I^n | n = 1..N\}$ be a set of image regions, each containing a person in a pose from an *unknown* pose class shared by all images. Pose co-estimation (PCE) is the task of estimating all poses *jointly over* $\mathcal{I}$. The desired output is one local pose per image region $\mathcal{L} = \{L^n | n = 1..N\}$, and one or a few *prototype poses* characterizing the pose class (fig. 1b). Note how the input regions could come from the same image, e.g. several people performing a synchronized activity (fig. 1top), or from many images, e.g. the images returned by Google Images by the name of a pose class (sec. 6.1). Our PCE model does not differentiate between these situations and simply inputs an unordered set of image regions returned by a person detector (sec. 2.2). To simplify the explanations, in the following we use the word 'image' for an image region.

In this section we introduce several PCE models, ordered by increasing level of complexity. We highlight their advantages and shortcomings and explain how to perform inference in each of them. In the simplest model, the pose class has a single prototype M (fig. 1(top)) and the local poses L^n are all identical to each other and to the prototype (sec. 3.1). In the second model, the local poses $\mathcal{L}$ are all different variants of the prototype (sec. 3.2). Finally, the most complex model allows for a multi-modal pose class by having multiple prototypes

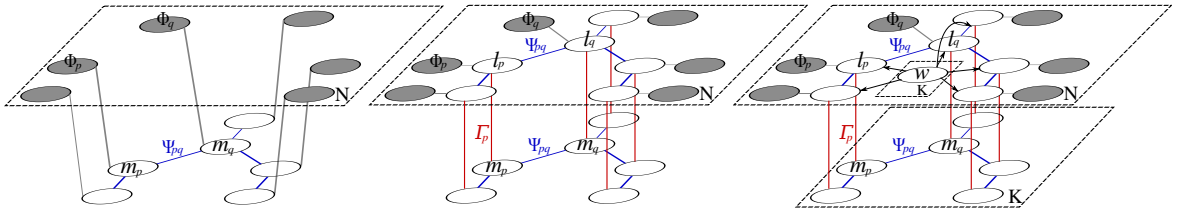


Fig. 2: **Pose Co-Estimation Models.** *(a) Direct: a single prototype pose M identical to all local poses; (b) Hierarchical: local poses are variations of the prototype; (c) Multi-modal Hierarchical (MmH): multiple prototype poses. For clarity we omit superscripts and differentiate by color the two kinds of pairwise potentials, i.e. kinematic constraints Ψ (blue) and prototype compatibility constraints Γ (red).*

$\mathcal{M} = \{M^k | k = 1..K\}$ (sec. 3.3). For example, the "tree pose" can be carried out by joining hands over the chest or above the head (fig. 1(bottom)).

3.1 Direct Model

In the direct model there is a single prototype M and all local poses are equal to it: $\mathcal{L} = \{L^n = M | n = 1..N\}$.

Model. The direct model reduces PCE to finding a single configuration of body parts (prototype) $M = \{m_p | p = 1..B\}$ that directly explains all images $\mathcal{I}$ (fig. 2a):

$$P(M \mid \mathcal{I}) \propto \exp - \Big(\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(m_p, m_q) + \sum_n \sum_p \Phi_p(I^n | m_p) \Big) \quad (2)$$

It extends the PS model (1) over multiple images (here L is replaced by M to indicate that we estimate the prototype).

Inference. When $\mathcal{E}$ has a tree structure, inference in the direct model is globally optimal, as for the PS model (1). The main difference between (1) and (2) is that likelihoods Φ are accumulated over all images in $\mathcal{I}$.

3.2 Hierarchical Model

By using a single, fixed configuration of body parts $L^n = M$ to explain every image I^n in $\mathcal{I}$, the direct model might perform poorly, as there is some variability between instances of a pose class. The hierarchical model extends the direct model to allow more flexibility. It models the local poses $\mathcal{L}$ in the images $\mathcal{I}$ as *similar* to the prototype M. This enables to estimate poses *adopted to each image* (fig. 2b).

Model. The hierarchical model is defined over both M and $\mathcal{L}$ simultaneously:

$$P(\mathcal{L}, M \mid \mathcal{I}) \propto \exp - \Big(\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(m_p, m_q) + \sum_n \Big[\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(l_p^n, l_q^n) + \sum_p \Phi_p(I^n | l_p^n) + \sum_p \Gamma_p(l_p^n, m_p) \Big] \Big) \quad (3)$$

All local poses L^n and the prototype M must obey the kinematic constraints Ψ. The similarity between M and each L^n is encouraged by part-specific compatibility constraints $\Gamma_p(l_p^n, m_p) = -\log \mathcal{N}(l_p^n - m_p | 0, \Sigma_p)$, where $\mathcal{N}$ is a zero-mean Gaussian with a diagonal covariance matrix Σ_p. Sec. 6.3 explains how we learn Σ_p from training data.

Indirectly, Γ introduces dependencies between the local poses in different images. We call this model *hierarchical*, as similarity between the local poses is enforced indirectly through the prototype (fig. 2b). Note that for $\Sigma_p \to 0$ the hierarchical model reduces to the direct model (2).

Inference. As the hierarchical model is loopy, even if $\mathcal{E}$ is a tree, globally optimal inference is impractical. We propose an approximate inference scheme composed of two stages:

Stage 1 (find M^*). Ignore from (3) the kinematic constraints $\Psi_{pq}(l_p^n, l_q^n)$ of the local configurations of parts $\mathcal{L}$. Optimize

$$\min_M \Big(\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(m_p, m_q) + \sum_n \Big[\sum_p \Phi_p(I^n | l_p^n) + \sum_p \Gamma_p(l_p^n, m_p) \Big] \Big) \quad (4)$$

As this reduced model is a very large tree, we run an exact inference to obtain the prototype M^*. We find the optimal M^* by passing messages forward to the root and then back to all m_p. As we are not interested in $\mathcal{L}^*$, we do not need to complete the backward pass.

Stage 2 (find $\mathcal{L}^*$). Keeping M fixed to M^*, optimize (3) over $\mathcal{L}$. As every m_p is fixed, Γ_p acts as an additional unary potential independently on each local part l_p^n. Moreover the kinematic potentials $\Psi_{pq}(m_p, m_q)$ of the prototype are constant and have no effect on the optimization. Exact inference over $\mathcal{L}$ in this reduced model is very efficient, as the unclamped variables decompose into N independent trees, one per image. Note, how the configurations $\mathcal{L}^*$ obey now the kinematic constraints.

3.3 Multi-modal Hierarchical Model (MmH)

The hierarchical model delivers a single prototype M and a local pose L^n for each image I^n, which is suitable for a uni-modal pose class (fig. 1(top)). The MmH model extends the hierarchical model to handle a multi-modal pose class (e.g. the "tree pose" in fig. 1(bottom)).

Model. To account for a K-modal pose class shared by N images in $\mathcal{I}$, the model estimates K prototypes $\mathcal{M} = \{M^k | k = 1..K\}$. We couple $\mathcal{L}$ with $\mathcal{M}$ through image-to-mode soft-assignment variables $\mathcal{W} = \{w^{nk} \in [0,1] \mid n = 1..N, k = 1..K\}$. Each w^{nk} indicates to which degree I^n belongs to mode k. The MmH model (fig. 2c) is:

$$P(\mathcal{L}, \mathcal{M}, \mathcal{W} \mid \mathcal{I}) \propto \exp - \Big(\sum_k \sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(m_p^k, m_q^k) + \sum_n \sum_k w^{nk} \Big[\sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(l_p^n, l_q^n) + \sum_p \Phi_p(I^n | l_p^n) + \sum_p \Gamma_p(l_p^n, m_p^k) \Big] \Big) \quad (5)$$

$$\text{with} \quad \sum_k w^{nk} = 1 \quad \text{, for all } n \in \{1..N\}$$

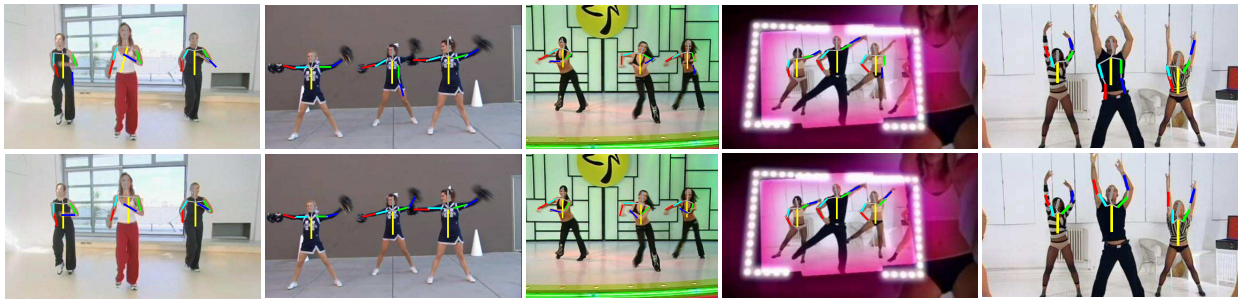


Fig. 3: **Results for Synchronized Activities.** *Top row: independent pose estimation results (comboPS); Bottom row: pose co-estimation results using our Hierarchical PCE model. PCE typically corrects the pose of a person, by enforcing similarity across persons.*

When $K = 1$, the MmH model reduces to the hierarchical model (3). As $\mathcal{L}$ and $\mathcal{M}$ depend on $\mathcal{W}$ and vice-versa, exact inference in the MmH model is prohibitively expensive.

Inference. We extend the approximate inference scheme of the hierarchical model (sec. 3.2) to MmH. We iteratively alternate between the following two stages.

Stage 1 (find $\mathcal{M}^*$). Ignore from (5) the kinematic constraints $\Psi_{pq}(l_p^n, l_q^n)$ of the local configurations of parts $\mathcal{L}$. Optimize

$$\min_{\mathcal{M},\mathcal{W}} \Big(\sum_k \sum_{(p,q)\in\mathcal{E}} \Psi_{pq}(m_p^k, m_q^k) + \sum_n \sum_k w^{nk} \Big[\sum_p \Phi_p(I^n|l_p^n) + \sum_p \Gamma_p(l_p^n, m_p^k) \Big] \Big) \quad (6)$$

$$\text{where} \quad \sum_k w^{nk} = 1 \quad \text{for} \quad n \in \{1..N\}$$

In the multi-modal case ($K > 1$) even the reduced model (6) cannot be optimized exactly, as $\mathcal{M}$ and $\mathcal{W}$ induce circular dependencies (fig. 2c). Hence, we propose a EM-like algorithm that optimizes iteratively over $\mathcal{M}$ or $\mathcal{W}$, while keeping the other fixed. In the first step the prototypes $\mathcal{M}$ are fixed and (6) becomes a linear program (LP) in w^{nk}. It can be solved exactly by solving N independent LPs as there is no interaction between images at the level of $\mathcal{W}$. The LP for image I^n is

$$\min_{w^n} \sum_k w^{nk} f^{nk} \quad (7)$$

$$s.t. \sum_k w^{nk} = 1, \quad \text{with} \quad f^{nk} = \sum_p \Phi_p(I^n|l_p^n) + \sum_p \Gamma_p(l_p^n, m_p^k)$$

where the vector $w^n = [w^{n1}, ..., w^{nK}]$ collects the assignments of image I^n to each prototype. Note f^{nk} does not depend on w^{nk} and thus is a constant coefficient in the LP. The optimal solution of any LP lies in a vertex of the polyhedron defined by the constraints [4]. Hence, the optimal solution to (7) is a hard-assignment of I^n to one prototype

$$w^{*nk} = \begin{cases} 1 & \text{for } k = \arg\min_i(f^{ni}) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

In the second step of the EM-like algorithm we optimize (6) over $\mathcal{M}$ while keeping $\mathcal{W}$ fixed to $\mathcal{W}^*$ found in the first step. Since $\mathcal{W}^*$ consist of hard-assignments, optimizing $\mathcal{M}$ is equivalent to finding K independent prototypes, each by optimizing a tree model defined over a subset of the images. Therefore, both steps of the EM-like algorithm to optimize (6) can be solved exactly and efficiently. We iterate it until convergence, i.e. no changes in $\mathcal{W}^*$. In practice, it takes 3 iterations on average. We describe below how to initialize $\mathcal{W}$.

Stage 2 (find $\mathcal{L}^*$). Keeping $\mathcal{M}$ and $\mathcal{W}$ fixed to $\mathcal{M}^*$ and $\mathcal{W}^*$ found by stage 1, optimize (5) over $\mathcal{L}$. By taking into account that hard-assignments $\mathcal{W}^*$ are obtained in stage 1 and that $\mathcal{M}$ is fixed, finding the optimal $\mathcal{L}^*$ here reduces to K independent problems of the same kind as stage 2 in sec. 3.2. There is one such problem per mode k, defined over the subset of images selected by w^{nk} for all n.

Initializing the assignments $\mathcal{W}$. We explain how to initialize the image-to-mode assignments $\mathcal{W}$ required by our inference scheme. The key idea is to run the PS model (1) on each image independently, and then cluster images based on the similarity of the resulting pose estimates. We measure the similarity of the posterior marginals of (1) (PM), rather than the MAP. Even when the MAP is incorrect, often the correct pose has a high probability. In general, the PM offer a more stable representation of pose than the MAP, especially for difficult images. We initialize $\mathcal{W}$ by agglomerative clustering [21] on the dissimilarity matrix between all pairs of images. We use the dissimilarity measure between two PMs proposed by [12], which compares the distributions of relative locations and relative orientations between body parts. We stop agglomerative clustering just before all images are in a single cluster. We then keep the clusters containing more than 15% of the images and reassign the remaining images to their most similar cluster. This outputs the number of modes K (prototypes) and the initial assignments of images to modes $\mathcal{W}$.

In summary, our method discovers *fully automatically* the number of prototypes K and the prototypes $\mathcal{M}$ themselves. Nothing but the image set $\mathcal{I}$ is given as input.

4 Technical Details

Appearance models Φ. We use two types of appearance models: (i) the generic model of [1], based on shape-contexts computed on image gradients and trained discriminatively with Ada-Boost; (ii) we use the algorithm of [6] to generate person-specific appearance models for each image I^n. We use implementations released on-line by the respective authors[1] [2].

Kinematic constraints Ψ. As in [19]: (i) for the relative position (x, y) we use a truncated cost giving 0 close to the joint location and infinity elsewhere; (ii) for the relative orientation θ we use a non-parametric distribution.

1. Pictorial Structures Revisited, www.d2.mpi-inf.mpg.de/People/andriluka/
2. 2D articulated human pose estimation, www.vision.ee.ethz.ch/~calvin/

Person Detector. We use a publicly available upper-body detector[3], based on the part-based model of [9] complemented with a face detector [29] for improved coverage.

5 Experiments - Synchronized Activities

In synchronized activities multiple persons take on a similar pose at the same time, such as during aerobic exercise, cheerleading, or group dancing (fig. 3). Here we perform PCE over all persons in an image. The input set of image regions $\mathcal{I}$ is obtained from the person detector. As we expect the poses in one image to have a single mode, we use the Hierarchical PCE model (sec. 3.2). However, the poses of different persons do not have to be exactly synchronized or performed in exactly the same way (fig. 3). The compatibility term Γ (eq. (3)) only requires them to be similar.

Dataset. We collected from the Internet 357 images with a total of 1128 persons. Each image contains multiple persons in roughly synchronized poses of cheerleading, aerobic and dancing. We annotate every person with a stickman, i.e. a line segment per body part. These annotations are used only for evaluation, they are not given to the algorithm. The dataset covers upright persons with a great variety of arms poses, covering the space of possible configurations quite uniformly. The stickmen distribution over the whole dataset is shown in fig. 4 using the visualization proposed by [27].

Measure. Our PCE technique estimates a stickman for each detection window (i.e. for each image region in the input set $\mathcal{I}$). We evaluate performance using the same measure as [1, 6, 11, 14, 23, 30], i.e. the percentage of correct body parts (PCP). An estimated body part is correct if its segment endpoints lie within a fraction of the length (pcp-threshold) of the ground-truth segment from their annotated location. The pcp-threshold controls how accurate the estimated sticks must be to be counted as correct. PCP is evaluated only on correct detections, which cover 84.2% of all ground-truth stickmen.

Competitors. We compare to the PS model (1) applied independently to each image, using the same appearance models and kinematic constraints as for PCE (sec. 4). Note how this is a strong baseline, as it combines the successful elements of two recent methods [1, 6] and it was shown to outperform both in [7].We refer to this model as *comboPS*. Moreover, we also compare to [1, 6, 23] using the implementations provided by the authors [1] [4]. All methods are given *the same detection windows* [3] as preprocessing.

Results. The PCP curves in fig. 4 show that PCE brings an improvement over comboPS, despite having only 3 persons on average per image. As we show in sec. 6.4, the improvement is greater when more persons are estimated at the same time.

All single-person HPE techniques we compare to [1, 6, 23] have been pre-trained by the respective authors on the Buffy Stickmen dataset [5], which has a substantially different pose distribution than our synchronic activity dataset, with many more samples with arms below the head than above [27]. The

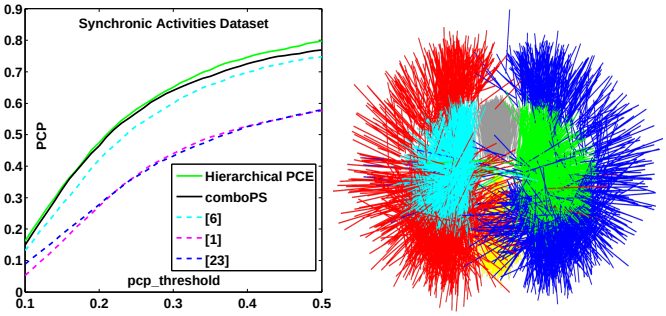


Fig. 4: **Synchronized Activities Dataset.** *(a) Hierarchical PCE performs better than independent pose estimators; (b) scatter plot depicting pose variability over the dataset, inspired by [27] (lower arms in red and blue, upper arms in cyan and green, head in gray, torso in yellow).*

modest performance of [1] could be explained by this "arms-down" pose prior. Unexpectedly, [23] also performs modestly. This could be due to the way the search-space is cropped by the authors of [23] to be just above the head [4], preventing any pose with "arms up" to be correctly estimated. Instead, [6] has a weak, nearly uniform pose prior which is not fit to Buffy and enables it to perform quite well on the synchronized activities dataset. Finally, note how comboPS improves over both its components [1, 6], confirming what reported in [7] on other datasets, and underlying the importance of having good generic appearance models [1].

6 Experiments - Learning Pose Prototypes

Here we apply PCE for learning pose class prototypes directly from images collected from Google Images when queried with the name of a pose class (e.g. 'lotus pose'). We start by summarizing our automatic processing pipeline (sec. 6.1). Next, we assess our PCE technique on two datasets of pose classes (sec. 6.2) and describe the experimental setup and parameter learning in sec. 6.3. We then quantitatively evaluate PCE in two ways: (i) its impact on pose estimation accuracy (sec. 6.4), (ii) the quality of the produced prototypes (sec. 6.5). Finally, we apply the prototypes as a prior for HPE on novel images of an activity-specific dataset (yoga) and two generic standard datasets containing various poses (sec. 6.6).

6.1 Pipeline

We summarize here our pipeline for automatically learning pose classes from the Web: (**1**) query Google Images with a pose class name and collect the top 50 returned images; (**2**) run the person detector [3] on every image; (**3**) apply an *image prefiltering*: remove images without any detection or with more than two (as in this case we do not know who is in the pose class); (**4**) the highest scoring detection in each remaining image defines the image set $\mathcal{I}$; (**5**) normalize the state-space of every detection in $\mathcal{I}$ (sec. 2.2). (**6**) input $\mathcal{I}$ into one of our PCE algorithms (sec. 3). Importantly, the image set $\mathcal{I}$ is the only input given to the PCE algorithms. It is obtained fully automatically, given only the name of a pose class.

3. CALVIN upper-body detector, www.vision.ee.ethz.ch/~calvin/
4. Cascaded Pictorial Structures, www.vision.grasp.upenn.edu/video/
5. www.robots.ox.ac.uk/~vgg/data/stickmen/

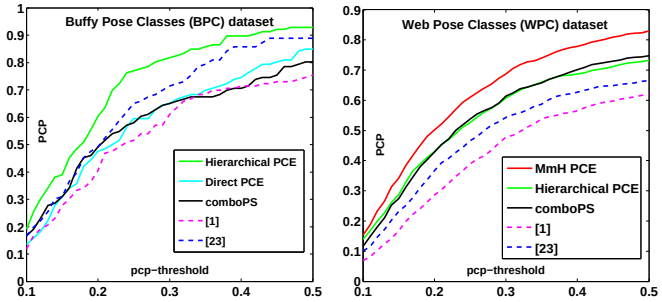


Fig. 5: **Quantitative evaluation of pose estimation.** *PCP curves.*

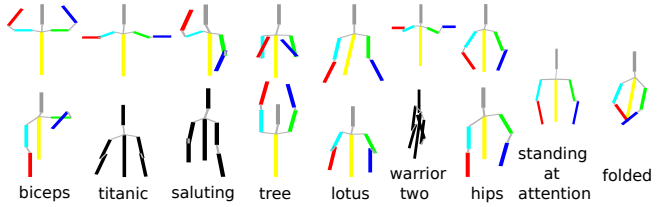


Fig. 7: **Prototype poses learned from the Web** *(WPC dataset). We show all prototypes learned automatically by our MmH model. We mark in black incorrect prototypes learned from noise images.*

6.2 Datasets

We investigate the PCE performance on two datasets: (i) Buffy Pose Classes [12] [6] ; (ii) a new dataset of pose classes collected from the web.

Buffy Pose Classes (BPC). It contains 7 images for each of 3 uni-modal classes (hips, folded, rest) from "Buffy: The Vampire Slayer" (called 'query images' in the dataset release). For evaluation, we annotate them with stickmen, i.e. a line segment per body part. These images enable testing PCE in a controlled scenario, where classes are uni-modal and all images in a set belong to the class (no noise).

Web Pose Classes (WPC). We collect this new dataset by querying Google-Images with the names of 9 pose classes: *biceps pose, folded arms, hips pose, titanic pose, standing at attention, saluting, lotus pose, tree pose, warrior two.*

For each person we mark whether it belongs to the query pose class. If it does, then we annotate it with its pose mode and a stickman. These annotations are used only for evaluation. In this dataset, 12% of the images do not contain any person visible at least from the waist up. Moreover, another 14% contain persons not in the query pose class, for a total of 26% noise images. Applying image prefiltering reduces this to 21% (sec. 6.1). This dataset is much harder than BPC, because of the noise, the lower image quality, and the higher variability in persons, clothing, backgrounds and poses (fig. 6).

Deriving ground-truth prototypes. Based on the annotated stickmen, we derive a ground-truth prototype pose for each mode of every pose class. For each body part of a prototype, we compute the mean (x, y, θ) stick over all images in the mode. We use these ground-truth prototypes to quantitatively evaluate the prototypes produced by PCE (sec. 6.5).

6.3 Experimental setup and parameter learning

We perform experiments in a leave-one-out scheme, i.e. we evaluate one pose class at a time and train on the others. We learn the two groups of parameters below.

Compatibility Γ_p. For each mode of each training pose class, we estimate the covariance matrix of the (x, y, θ) coordinates of the ground-truth stickmen (a separate covariance per body part p). We set the covariance Σ_p of the compatibility potential Γ_p to the average over all modes and classes.

Generating appearance models. We follow the procedure of [6] to train the parameters of the algorithm used to generate person-specific appearance models (sec. 4).

6. www.robots.ox.ac.uk/~vgg/data/buffy_pose_classes/

6.4 Evaluation of pose estimation

Measure. For each pose class we report the average PCP (sec. 5) over the images containing an instance of the class (positive images). This is a subset of the images input to PCE, due to noise images remaining after prefiltering (sec. 6.1). Nevertheless, note how the images returned by Google-Images are automatically filtered and all output, including noise, is fed to PCE with no manual intervention. We report the average PCP over all positive images.

Results. We evaluate on two datasets BPC and WPC (fig. 5). BPC contains uni-modal poses and no noise. On this dataset, our Hierarchical PCE model outperforms comboPS as well as [1, 23] over the entire PCP curve, reaching 93% at the lower accuracy end (pcp-threshold 0.5). As a side note, the Direct PCE model performs worse, as it estimates exactly the same pose in every image. This confirms the importance of local adaptation to the image in the Hierarchical model.

WPC is a harder dataset as it contains multi-modal poses, noise images and higher variability than BPC. This is reflected in the 5-20% lower PCP of comboPS, [1] and [23]. Hierarchical PCE now performs only on par with comboPS, suggesting that accounting for multi-modality is necessary here. Our full model (MmH PCE) allows for multiple prototypes and improves over comboPS and Hierarchical PCE by about 10% consistently over the entire PCP curve. Moreover, it also performs about 18% better than both [1, 23] and achieves 83% PCP at the lower accuracy end.

These experiments demonstrate the benefits of performing pose estimation *jointly* over all persons in a pose class set. Qualitative results on WPC are shown in fig. 6.

6.5 Evaluation of the prototypes

In addition to better pose estimation on the input images, PCE also estimates one or more prototypes characterizing the pose class. Here we quantitatively evaluate these prototypes.

Measure. For evaluation we count how many body parts of the estimated prototype are close to the ground-truth one (at pcp-threshold 0.5). We report the average over all modes in all pose classes, weighted by the number of images in a mode, obtaining a PCP measure.

Results. As a baseline, we compare to averaging the stickmen output by comboPS over the image set $\mathcal{I}$. On the easier, unimodal BCP dataset our Hierarchical PCE method obtains a perfect PCP of 100%, compared to 88.9% for the baseline. On the harder, multi-modal WPC dataset, our MmH method outperforms the baseline by 14% PCP (91.8% vs 77.8%). Fig. 7 shows the prototypes estimated by MmH automatically

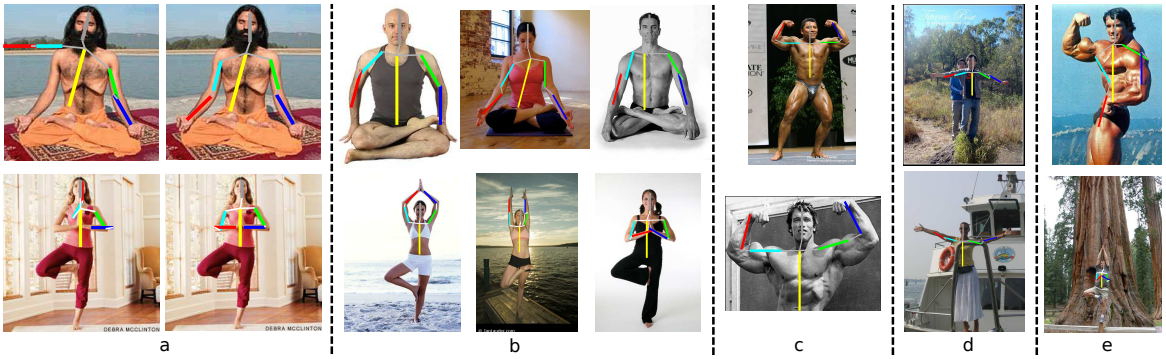


Fig. 6: **Qualitative results on the Web Pose Classes dataset.** *(a) head-to-head comparison between comboPS (left) and our new MmH PCE (right). (b) 3 additional results for lotus (top) and tree (bottom) poses. Note how MmH correctly discovered that these pose classes are bi-modal; (c)-(d) sample results for biceps and titanic. Note how PCE adapts the estimated prototypes (fig. 7) to the individual images. (e) failure cases due to incorrect image-to-mode assignment.*

on WPC (note how their number is also found automatically by MmH, sec. 3.3). These results confirm that the proposed MmH model learns prototypes even in the challenging case of multi-modal poses, which cannot be achieved by HPE on each image independently. Note how PCE can estimate prototypes for arbitrary pose classes, as long as instances of the class recur frequently in input image set $\mathcal{I}$.

6.6 Applications to HPE on new images

Here we apply the prototypes learned by PCE as a prior to improve single-person HPE on novel images. We evaluate this on 3 datasets, one specific to the yoga activity and two generic. No image in these test datasets was seen before by the system when learning the prototypes.

Activity-specific HPE. An activity, such as yoga, tennis or dancing, can be represented as a collection of pose classes characteristic for it. In turn, each pose class is represented by its prototypes. Fig. 8(a)-top shows a hierarchical view of the *yoga* activity $\mathcal{Y}$, composed of the pose classes *tree*, *lotus* and *warrior two*, for a total of 6 prototypes automatically learned by our MmH model from the WPC dataset.

We use the activity model $\mathcal{Y}$ as a prior for HPE on novel test images of yoga, collected by querying Google Images with *yoga pose*. The resulting dataset contains *any* yoga pose, not only those in $\mathcal{Y}$. We discard images contained in the WPC sets for *tree, lotus, warrior two*, and annotate the remaining 61 persons with stickmen. This experiment is close to practical application scenarios, where we are likely to know what high-level activity is shown in an image, but not what pose class.

For each prototype $M \in \mathcal{Y}$, we estimate the pose L in a test image by optimizing (3) while keeping M fixed. This results in $|\mathcal{Y}| = 6$ poses. We return the one with the lowest energy according to (3) after ignoring the Γ term (when assessing which pose fits the image best, the compatibility Γ between the pose and the prototype does not matter).

The PCP curves in fig. 8 show that using the activity prior $\mathcal{Y}$ improves the performance of comboPS by 6%. The improvement is due to the compatibility term Γ, which encourages poses similar to the prototype in a smooth manner, adapting it to the image contents.

Generic HPE. We experiment on the generic datasets *Buffy Stickmen* [5] and *ETHZ Pascal Stickmen* [7], which are standard benchmarks to evaluate HPE methods. For this we employ a diverse pose prior composed of all 17 prototypes from the 9 pose classes of WPC. This prior is applied in the same manner as the yoga-specific $\mathcal{Y}$. Following the protocol of [6], we evaluate on episodes $2, 5, 6$ of Buffy Stickmen (276 images) and on the whole ETHZ Pascal Stickmen (549 images).

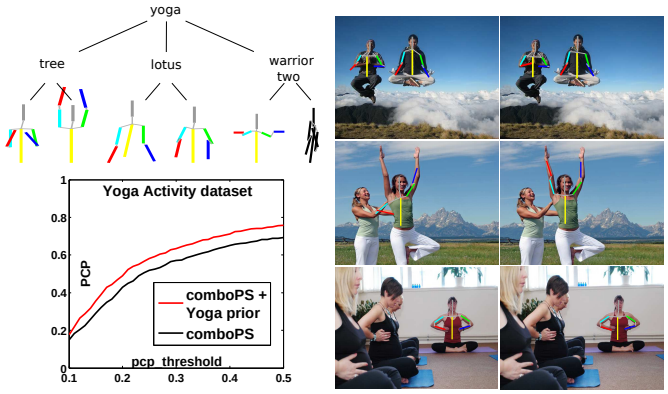
We compare to comboPS without prior and to [1, 6, 23] using code by the respective authors[1] [2] [4]. In order to compare purely pose estimation performance, all methods are given the *same detection windows* [3] as preprocessing. As in the protocol of [6], PCP is evaluated only on correct detections, i.e. covering a ground-truth stickman. These cover 95.3% and 75.1% of the ground-truth for Buffy and ETHZ Pascal Stickmen respectively [8].

Fig. 9 shows the PCP performance curves. As in sec. 5, comboPS outperforms both its components [1, 6], justifying its choice as a competitive baseline. Importantly, comboPS performs better with the pose prior than without, on both datasets and over the entire PCP curve. This shows that a diverse pose prior, learned automatically given just the *names* of 9 pose classes, helps pose estimation even on novel images containing arbitrary poses. This agrees with [14], where multiple pose priors improve over a monolithic one. However, [14] learns from manually annotated stickmen pose priors corresponding to viewpoints. Our priors instead correspond to distinct poses and are learned in a weakly supervised manner. The authors of [22] take the idea to the extreme by computing a pose prior adapted to each test image based on ground-truth stickmen from similar training images.

For completeness, we point out that [22, 23] report better results than ours on both datasets when evaluated on detection windows output by an older detector [6, 11], which covers a smaller subset of the data (also in fig. 9). Finally, [30] reports the best results on both datasets so far.

7. http://www.vision.ee.ethz.ch/~calvin/ethz_pascal_stickmen/

8. The results we reproduce for [1, 6, 23] based on CALVIN detector[3] are close to what they originally published. At pcp-threshold 0.5, the differences are: (i) Buffy Stickmen: +5% for [1], +1.5% for [6], -0.6% for [23]; (ii) ETHZ Pascal stickmen: [1] not reported before, -1.5% for [6], -6% for [23]. The differences are due to the CALVIN detection windows covering about 10% more ground-truth stickmen than those of [6, 11] used in [1, 6, 22, 23].



(a) (b)

Fig. 8: **Yoga activity experiment.** *(a): yoga activity hierarchy and PCP curve for the test dataset (comboPS with/without activity prior); (b) comparison of comboPS with activity prior (right) and without (left). The first two cases improve thanks to the prior. The third case is estimated correctly despite it does not resemble any prototype in the prior. This is possible because the compatibility term Γ gives about the same non-zero probability to all poses far from any prototype.*

6.7 Conclusions

We have presented the novel human Pose Co-Estimation technique (PCE) for joint pose estimation over multiple persons in a common, but unknown, pose. We have demonstrated its benefits for estimating poses in images of synchronized activities and for learning prototypes of pose classes fully automatically, directly by querying Google Images with the class name. Moreover, we have shown the prototypes learned by PCE to form a valuable prior improving pose estimation on novel images of a specific activity and even on generic datasets containing arbitrary poses (sec. 6.6). Importantly, this prior is learned directly from web images without any additional stickmen annotations.

In addition to the applications we demonstrate, we believe PCE could be valuable in other scenarios where several people perform the same activity (e.g. workers in a factory, athletes in sports events, actors on stage). Moreover, the prototypes we learn could be used to query a pose retrieval engine [12] to search for similar poses in a large database. As the prototypes are learned given only a class name, this would effectively enable a new *query-by-text* functionality.

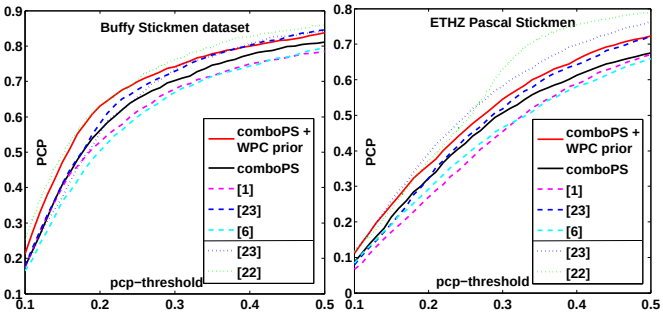


Fig. 9: **Buffy Stickmen and ETHZ Pascal Stickmen.** *Solid: our methods, evaluated on the CALVIN detection windows [3]. Dashed: other methods, evaluated on the same windows. Dotted: other methods, evaluated on the older detection windows [6]. These cover a smaller subset of the data, so they are not directly comparable. We could not evaluate [22] on the CALVIN windows as their code is not available.*

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[3] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who's in the Picture. In *NIPS*, 2004.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] P. Buehler, M. Everinghan, D. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language tv broadcasts. In *BMVC*, 2008.

[6] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.

[7] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010.

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[10] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *ECCV*. Springer-Verlag, May 2004.

[11] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[12] V. Ferrari, M. Marin, and A. Zisserman. Pose search: Retrieving people using their pose. In *CVPR*, 2009.

[13] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *ICCV*, 2001.

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[15] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007.

[16] X. Lan and D. Huttenlocher. Beyond trees: Common-factor models for 2D human pose recovery. In *ICCV*, volume 1, 2005.

[17] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *ECCV*. Springer, 2004.

[18] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004.

[19] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[20] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *CVPR*, 2006.

[21] C. Romesburg. *Cluster Analysis for Researchers*. Krieger Pub. Co., 2004.

[22] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010.

[23] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[24] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[25] L. Sigal and M. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, volume 2, pages 2041–2048, 2006.

[26] L. Sigal and M. Black. Predicting 3D people from 2D pictures. In *In AMDO*, 2006.

[27] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010.

[28] J. Valmadre and S. Lucey. Deterministic 3d human pose estimation using rigid structure. In *ECCV*, 2010.

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001.

[30] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.