Appearance Sharing for Collective Human Pose Estimation

Appearance Sharing for Collective Human Pose Estimation

Marcin Eichner[1], Vittorio Ferrari[2]

[1] ETH Zurich, Switzerland, eichner@vision.ee.ethz.ch
[2] University of Edinburgh, United Kingdom, vferrari@staffmail.ed.ac.uk

Abstract. While human pose estimation (HPE) techniques usually process each test image independently, in real applications images come in collections containing interdependent images. Often several images have similar backgrounds or show persons wearing similar clothing (foreground). We present a novel human pose estimation technique to exploit these dependencies by sharing appearance models between images. Our technique automatically determines which images in the collection should share appearance. We extend the state-of-the art HPE model of Yang and Ramanan to include our novel appearance sharing cues and demonstrate on the highly challenging Leeds Sports Poses dataset that they lead to better results than traditional single-image pose estimation.

1 Introduction

2D articulated human pose estimation (HPE) in still images is a very challenging problem that has received considerable attention in recent years [1–10]. Thanks to the progress in those works, HPE methods can now be applied with some success on uncontrolled still images, without any prior knowledge about poses, the appearance of persons or backgrounds. However, the problem is far from solved. Highly cluttered backgrounds, large scale changes and strong scale variations can cause the failure of even the most recent state-of-the-art methods [10, 9].

A trait common to essentially all approaches [1–10] is to estimate pose *independently* on each image. We believe that this makes the problem harder than its needs to be. In real applications the test images come in *collections*, not one at a time. The user typically runs a pose estimator on a collection and only later inspects the results or inputs them to subsequent stages of a larger system. Importantly, usually there are *dependencies* between the images in a collection. Often some of them show people against a common background, while others show persons wearing very similar clothing (foreground). This happens a lot in sports photography, where both the background (football pitch, gym hall, water pool, tennis court) as well as the foreground recur (different players in the same football team, or even the same athlete in different poses or viewpoints, fig. 1). Images with either foreground or background in common are frequent also in: a) video surveillance (images taken in front of the same background); b) in movies (an actor wearing the same clothes throughout an episode); c) holiday photo collections, where the same person appears in many pictures, often wearing the same clothes and/or repeatedly visiting the same location (e.g. beach, pool, hotel room). Even in pure research papers with no concrete application, the proposed methods are evaluated on entire test sets [1, 2, 4, 7–9] which feature the dependencies mentioned above. One of the

most challenging modern datasets [11], the Leeds Sports Poses (LSP) dataset [7][1] contains numerous images of soccer, baseball, tennis or indoor sports sharing backgrounds or persons wearing very similar outfits (fig. 1a).

In this paper, we propose a novel technique to exploit these phenomena by performing human pose estimation while *sharing appearance* between images. Our method automatically discovers clusters of images with common foreground or background in the collection and thus determines between which images to share appearance (fig. 1a). We robustly estimate color appearance models suitable for all images in a cluster and incorporate them into a state-of-the-art HPE model [10]. Inference on the extended model effectively performs pose estimation jointly over all images in the cluster, guided by the shared appearance models.

Our paper is organized as follows. The next section gives an overview of our approach (sec. 2), followed by a summary of the base HPE model [10] we build on (sec. 3). In sec 4 we present our technique for automatically estimating shared appearance models, and explain how to incorporate them into [10] in sec. 5. We present an extensive experimental evaluation in sec. 6, which demonstrates that our approach for sharing appearance models over images improves performance over [10] run independently on each image.

Related works. Articulated human pose estimation in still images is very challenging due to high variability in person appearance and pose, as well as the presence of background clutter, illumination and self-occlusions. Recent works have addressed these issues with advanced appearance models [1, 4, 3, 12, 7, 5, 10], complex pose priors [7, 5] or non-tree dependencies between body parts [8, 13] or hierarchical models [14]. While the above works build on variants of the Pictorial Structure (PS) model [15], there are also other techniques, e.g. bottom-up body assembling from segmentations [16] or pose estimation by foreground max-covering [17]. A trait common to essentially all approaches is to estimate pose *independently* on each image. In this paper instead we tackle HPE by exploiting multiple images sharing a common appearance.

The importance of good appearance models is reflected by their evolution. Early works employed simple box filters on background subtracted silhouettes [15]. Later, generic appearance models based on image gradients were developed, including generative edge masks [1], discriminatively trained shape-context templates [3], or linear [12] and non-linear [7] HOG templates [18]. In addition to generic templates based on gradients, a few works also employ color appearance models specific to a particular image, like in the iterative image parsing work of [1]. Later [4] extended this idea to transfer color models between body parts of a person, which was also adopted by [6]. In this paper we also propose appearance models based on color, but they are estimated *over multiple images*, following our main spirit of sharing appearance between images.

HPE approaches dedicated to video [19, 20, 2, 21] often employ multi-image models optimizing pose jointly over consecutive video frames. Usually, they exploit pose [2, 21] or appearance [19, 20] consistency over time. In this paper we tackle a different problem. A still image collection lacks temporal continuity and contains many different persons.

[1] http://www.comp.leeds.ac.uk/mat4saj/lsp.html

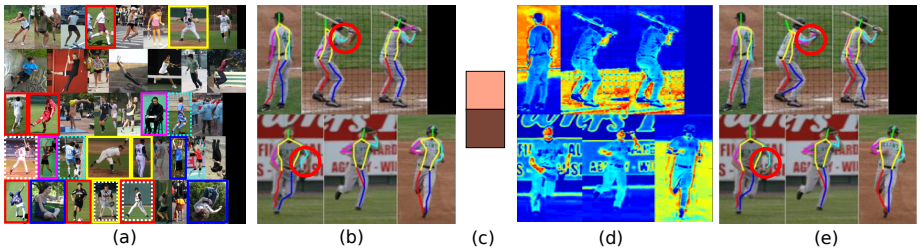


Fig. 1. Approach overview. *(a) random samples from the LSP dataset with example cluster assignments (solid boxes for background, dashed for foreground; colors depict cluster ids); (b) an automatically discovered foreground cluster with initial pose estimate [10] overlaid; the lower arms in the red circles are incorrectly estimated; (c) two high-weight color bins in the automatically estimated shared foreground model of a lower arm; (d) lower arm foreground likelihood computed according to this appearance model (heat-map); (e) improved pose estimation result produced by our extended HPE model which incorporates the shared appearance model; the lower arms in the red circles are now correctly estimated.*

2 Overview of our method

The core idea of our work is to improve HPE by exploiting background and foreground appearance patterns recurring over several images in a large dataset such as *LSP* [7] (fig. 1a). We give here an overview of the processing stages in our pipeline, focusing on the case of foreground appearance. The pipeline for background appearance is analogue. **1)** we run the pose estimator [10] independently for each image (sec. 3). This pose estimator employs person-generic body part appearance models based on gradients. **2)** we group images into clusters likely to have similar foreground appearance in terms of color distribution, based on the initial pose estimates (fig. 1b, sec. 4.2). **3)** for each cluster we robustly estimate a color appearance model *shared across the cluster* by integrating evidence over all images in it (fig. 1c, sec. 4.2). **4)** we use the shared appearance model to derive per-pixel foreground likelihoods for each image in the cluster (fig. 1d, sec. 4.3). **5)** we extend the HPE model of [10] to incorporate these foreground likelihoods as additional unary potentials (sec. 5). **6)** we run inference on the extended model to update the pose estimates in all images (fig. 1e).

Our method exploits the fact that clusters are typically mixed, containing some images with correct and some with wrong pose estimates. Therefore, instead of estimating pose on each image *independently* our method attempts HPE *jointly* over a cluster of images with similar appearance. Stage 3 robustly recovers the underlying shared appearance model, minimizing the impact of incorrect pose estimates. The resulting shared appearance model then helps in stage 6 by guiding the extended HPE towards better pose estimates.

As shown in extensive experiments, our method automatically exploits recurring appearance patters in a large dataset to successfully improve the accuracy of human pose estimation (fig. 1e). Our extended HPE model outperforms the baseline framework of [10] on one of the largest and most challenging HPE datasets available (LSP, sec. 6).

In the remainder of the paper we present each stage of our pipeline in detail.

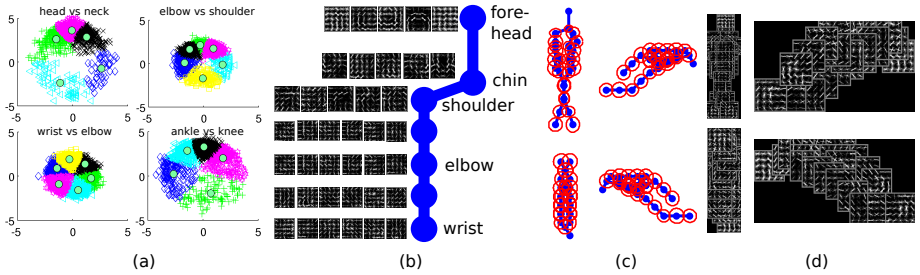


Fig. 2. Mixture of pictorial structures model (MoPS) [10]. *(a) Clusters of relative joint positions for finding body part types in the LSP training set; (b) a path in the model from the head to the left wrist, along with the HOG appearance templates specialized for each body part and orientation type; (c) visualization of 4 trees drawn out of an exponential number of trees that MoPS can generate; (d) HOG appearance visualization for the trees in (c).*

3 Base model [10] (MoPS)[2]

Many human pose estimators build on the pictorial structure model (PS) [1–4, 6–9]. A PS [15] is a conditional random field where nodes explicitly correspond to body parts (e.g. head, torso, left lower arm). The state space of a node is the set of possible (x, y, θ) positions a part can take in the image. The recent work of [10] introduces a novel representation: a mixture of pictorial structures (MoPS) (fig. 2), where each node represents a body part in a particular orientation. Nodes are now roughly corresponding to joints (e.g. elbow, knee) and midpoints of limbs. Hence, a part is represented as a mixture of axis-aligned templates, one per orientation.

The state space of a node is now only its (x, y) location. However, orientation is implicitly captured by different mixture components. These are estimated from the training data by clustering the relative position between neighboring parts, typically into 5-6 components per part (fig. 2a). This representation is highly flexible and enables to model foreshortening, an effect often responsible for failures of classic PS

Following the notation of [10], we write I for an image, p_i for the (x, y) location of part i and t_i for its mixture component, where $i \in \{1, ...K\}$, $p_i \in \{1, ...L\}$, and $t_i \in \{1, ...T\}$[3]. Here, t_i is the *type* of part i, which implicitly models its orientation. The energy of a body part configuration (p, t) is

$$\begin{aligned} S(I, p, t) &= \sum_i w_i^{t_i} \Phi(I, p_i) + \sum_{ij \in \mathcal{E}} w_{ij}^{t_i, t_j} \Psi(p_i - p_j) + S(t) \\ S(t) &= \sum_i b_i^{t_i} + \sum_{ij \in \mathcal{E}} b_{ij}^{t_i, t_j} \end{aligned} \tag{1}$$

where $\mathcal{E}$ is the set of edges connecting body parts in the kinematic tree. $\Phi(I, p_i)$ is a a feature extracted from image I at p_i, so $w_i^{t_i} \Phi(I, p_i)$ is the image likelihood for part i to

[2] http://phoenix.ics.uci.edu/software/pose/

[3] We omit the subscript to indicate the set spanned by it (e.g. $t = \{t_1, ...t_K\}$)

be at location p_i with type t_i. $\Psi(p_i - p_j) = [dx^2, dx, dy^2, dy]$ is the relative location of parts i and j, so $w_{ij}^{t_i,t_j}\Psi(p_i - p_j)$ evaluates a spatial prior defined over locations according to a spring-like deformation model for types t_i, t_j; $S(t)$ is a co-occurrence model that favors assigning certain types to certain parts ($b_i^{t_i}$) and certain combinations of types of pairs of parts ($b_{ij}^{t_i,t_j}$). It is a spatial prior defined over types (orientations).

The model components and their parameters reflect the idea of decomposing parts into types. The appearance model $w_i^{t_i}\Phi(I, p_i)$ is governed by parameters $w_i^{t_i}$ representing a HOG template specialized for part i and type t_i (fig. 2b). Hence each part has different templates for different orientations, which helps to capture the multi-modal appearance of body parts [9, 10]. The work of [10] uses a variant of HOG features [22] for Φ. The deformation model $w_{ij}^{t_i,t_j}\Psi(p_i - p_j)$ is a switching spring model controlling the relative placement of two parts. Each spring $w_{ij}^{t_i,t_j}$ is tailored to a particular pair of types t_i, t_j. This allows fine-grained control over the amount of deformation tolerable for each pair of part orientations.

Inference. Finding the configuration (p, t) that maximizes (1) can be done efficiently because $\mathcal{E}$ forms a tree and the pairwise potentials are quadratic functions. Using dynamic programming and efficient distance transforms [15] exact inference can be performed in complexity $O(KLT^2)$ [10].

Learning. The training set contains positive training images $\{I^\rho, p^\rho, t^\rho\}$ labeled by the ground-truth body part configuration on a person, and negative images I^η containing no person. Let $z^\rho = (p^\rho, t^\rho)$, with p^ρ the ground-truth joint locations in I^ρ and t^ρ is assigned by clustering $p_i^\rho - p_j^\rho$ over the training set into part type clusters (fig 2a). Note how (1) is linear in the model parameters $\beta = (w, b)$, with $w = (w_1^{t_1}, ...w_K^{t_T}, ...w_{ij}^{t_i,t_j}...)$ and $b = (b_1^{t_1}, b_K^{t_T}, ...b_{ij}^{t_i,t_j}...)$. Hence, it can be rewritten as $S(I, z) = \beta \cdot \Theta(I, z)$, with $\Theta = (\Phi(I, p_1), ..., \Phi(I, p_K), ...\Psi(p_i - p_j), ..., 1, 1, ...1....)$. With this reformulation, the model can be learned with a structured prediction objective function similar to [22]

$$\begin{aligned} \arg\min_{\beta,\xi_i \geq 0} \quad & \frac{1}{2}\beta \cdot \beta + C\sum_\rho \xi^\rho + C\sum_\eta \xi^\eta \\ \text{s.t. } \forall \rho \in \text{pos} \quad & \beta \cdot \Theta(I^\rho, z^\rho) \geq 1 - \xi^\rho \\ \forall \eta \in \text{neg}, \forall z \quad & \beta \cdot \Theta(I^\eta, z) \leq -1 + \xi^\eta \end{aligned} \tag{2}$$

The constraints state that positive examples (pos) should score at least 1, whereas all possible configurations z of parts in negative images (neg) should score at most -1.

The quadratic program (2) has an exponential number of constraints but it defines a convex problem which can be optimized using a dual coordinate-descent solver [10]. In practice, [10] first trains each body part template independently. These initial templates then serve as a good initialization for training of the entire model using dual coordinate-descent of [10].

4 Sharing appearance

We present here our technique for sharing appearance models between images (stages 2, 3 and 4 in sec. 2). We start by finding clusters of images likely to have similar fore-

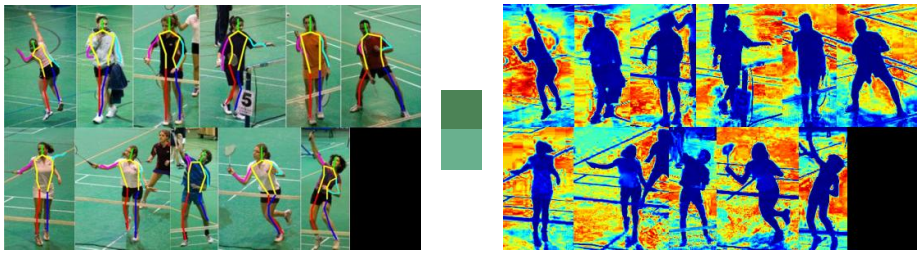


Fig. 3. Sharing background. *(a) example background cluster automatically found in the LSP dataset with initial pose estimates overlaid; (b) two high-weight color bins from the shared background model h^c_{bg}; (c) background likelihood maps I_{bg} computed of all images in the cluster based on h^c_{bg} (visualized as heat maps).*

ground or background appearance (sec. 4.1, 4.2). For each cluster we then estimate appearance models shared by all images in the cluster and derive from them pixel-wise likelihood maps (sec. 4.3). In section 5 we show how to incorporate these likelihood maps as additional unary potentials into the HPE model in order to improve pose estimation performance.

While single-image state-of-the-art HPE techniques typically employ gradients as features [3, 10, 9], our representation is based on color which is better suited to model the appearance of specific backgrounds or clothes.

4.1 Sharing background

The key idea is to exploit images with common backgrounds to generate an informative additional cue for constraining human pose estimation. We start by clustering images according to their background appearance. Next, for every cluster we robustly estimated a shared background appearance model.

Image clustering. For every image m in the input set $\mathcal{I}$, we extract color histograms h^m_{bg} from the entire image except the surface occupied by the initial pose estimate (sec. 2). Even when the initial pose estimate is partially incorrect, the error typically occupies a small portion of the images, so it has a minor influence on the histogram. We then define a pairwise similarity matrix between all images in $\mathcal{I}$ as $W^{mn} = 1 - \frac{\chi^2(h^m_{\text{bg}}, h^n_{\text{bg}})}{2}$. Next, we cluster $\mathcal{I}$ according to W using agglomerative clustering (sec. 4.3). This results in disjoint clusters $\mathcal{B}^c \subset \mathcal{I}$, with $\cup_c \mathcal{B}^c = \mathcal{I}$ (fig. 3a).

An advantage of agglomerative clustering over other techniques (e.g. k-means, GMM [23]) is that it does not require the number of clusters as input. Instead, it requires a parameter controlling cluster compactness (i.e. roughly how dissimilar two points can be and yet be in the same cluster). This is desirable in our setting as we want clusters of images with very similar backgrounds, but we do not know how many different types of background there are in the dataset.

Estimating a shared model. Given an image cluster $\mathcal{B}^c$, we estimate the shared background model h^c_{bg} by averaging the color histograms h^m_{bg} of all images $m \in \mathcal{B}^c$ in it. This estimate is robust to incorrect pose estimates in some of the images, as the moderate amount of foreground within their individual histograms is further diminished by averaging over all images (fig. 3b).

4.2 Sharing foreground

We start by forming image clusters containing persons with similar appearance, i.e. wearing similar clothes. In a typical cluster, the initial pose estimation worked correctly on some persons and failed on others (fig. 6). For each body part, the key idea of our method is to robustly recover its correct color model by finding the one which occurs most frequently over the cluster. These *per-part* color models are then used as additional cues to refine the pose estimate (next section). In this manner, our scheme automatically exploits images with correct pose estimates to improve other images where pose estimation failed.

Image clustering. In sec. 4.1 partially incorrect initial pose estimates had little impact, as a wrong limb covers a small portion of the image. Accordingly, the similarity between two images in sec. 4.1 simply ignored this issue. Importantly, in this section we want to find clusters of persons with similar appearance, despite some of them having partially incorrect pose estimates. To succeed we must design a more robust similarity measure that explicitly takes into account partially incorrect poses, as limbs are large relative to the area of a person.

Hence, we define the similarity between images $m, n \in \mathcal{I}$ as follows: $W^{mn} = \sum_i \delta\left(1 - \frac{\chi^2(h^m_i, h^n_i)}{2} > \varrho\right)$, where h^k_i is the color histogram of the patch covered by part i in image k, ϱ is a threshold on the similarity of corresponding body parts in the two images (e.g. the left lower arm in each image), and $\delta(\cdot)$ is 1 if the argument is true and 0 otherwise. This similarity measure counts how many body parts have similar appearance in the two images. Again we use agglomerative clustering to partition $\mathcal{I}$ into disjoint image clusters $\mathcal{F}^c$ (sec. 4.3). As fig. 4a shows, the robust similarity measure groups together images of persons wearing similar clothing, despite having errors in the initial pose estimates. These are the right conditions for our approach to make an improvement. As we will see below, the correct pose estimates in the cluster will help to fix the incorrect ones.

Estimating shared models. Given an image cluster $\mathcal{F}^c$, for each body part i we want to estimate its correct color model h^c_i shared by the persons in the cluster. Simply averaging the color histograms h^m_i of the patches of part i in all images $m \in \mathcal{F}^c$ would not produce a good shared color model. The images where the part is incorrectly estimated would spoil the average (fig. 4a). This is fundamentally different from the background shared model estimation (sec. 4.1). Instead, we propose here a technique for estimating the correct color model which is robust to incorrect initial pose estimates in some images.

We cast the problem as outliers detection. We cluster the patches $\{h^m_i\}_{m \in \mathcal{F}^c}$ using agglomerative clustering (fig. 4b). Then we find the dominant patch cluster $\mathcal{D}^c_i$

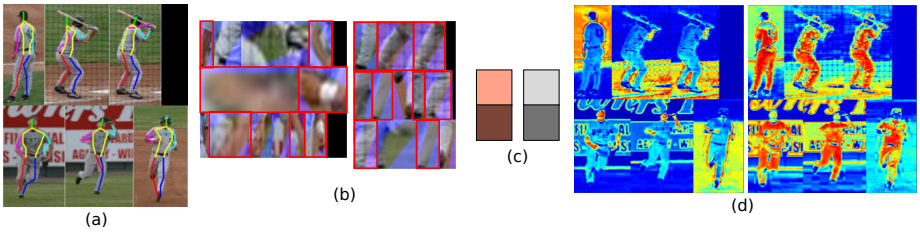


Fig. 4. Sharing foreground. *(a) example foreground cluster automatically found in the LSP dataset with initial pose estimates overlaid; (b) lower arm (left) and lower leg (right) patches from the initial pose estimates; areas under blue overlays do not belong to the patches; the members of the dominant clusters $\mathcal{D}^c_{lowerarm}$ and $\mathcal{D}^c_{lowerleg}$ are marked red; (c) two high-weight color bins in the shared foreground model of the lower arm $h^c_{lowerarm}$ (left) and lower leg $h^c_{lowerleg}$ (right), derived from the patches in the dominant clusters $\mathcal{D}^c_{lowerarm}$ and $\mathcal{D}^c_{lowerarm}$ respectively; (d) foreground likelihood maps $I_{fg,i}$ for all images in the cluster computed based on h^c_i, for i = lower arm (left) and i = lower leg (right).*

(fig. 4b). As the distribution of patch similarities may vary substantially for different kinds of parts i and image clusters $\mathcal{F}^c$, we use the median similarity of patches over all pairs of images in $\mathcal{F}^c$ as the compactness parameter. The key idea behind this dynamic parameter setting is that the similarity of patches within the dominant cluster is higher than the median. Hence, agglomerative clustering will tend to properly form a dominant cluster of highly similar patches, separated from many smaller clusters with other patches (fig. 4b).

Finally, we compute the shared color model h^c_i of part i as the average of the color histograms of the patches in the dominant patch cluster $\mathcal{D}^c_i$ (fig. 4c). This procedure correctly recovers the shared color model although the body part might be incorrectly localized in some images of $\mathcal{F}^c$. In fact it can work when the part is correctly localized even in fewer than 50% of the images, as long as the failures do not have consistent appearance. All it needs is for the correctly localized parts to form a dominant cluster in color space.

4.3 Technical details

Agglomerative clustering. We find background/foreground clusters using agglomerative clustering based on clique partitioning (CP) [24]. We construct a fully-connected graph G_{CP}, where each vertex represents an image $I_m \in \mathcal{I}$ and where edges are weighted according to pairwise similarity matrix $W \in [0,1] - \tau$, where τ controls the desired similarity. We partition then G_{CP} into disjoint cliques using CP. As CP is NP-hard we use the fast approximate clique partitioning technique of [25][4].

Unfortunately, no explicit background/foreground cluster membership annotations are available for the LSP dataset. Therefore, we set the clustering parameters (τ_{bg} and τ_{fg}, ϱ respectively) empirically using the LSP training set (see 6), such that visually appealing clusters are produced.

[4] http://www.robots.ox.ac.uk/ vgg/software/UpperBody/

Likelihood maps. After estimating shared color appearance models, we derive from them a pixel-wise likelihood map for each image in a cluster. The likelihood map is derived by assigning to each pixel its probability according to the color model. In a background cluster $\mathcal{B}^c$, the shared color model h^c_{bg} yields the same likelihood map I_{bg} for all body parts (fig. 3c). In a foreground cluster $\mathcal{F}^c$, there is a separate shared color model h^c_i for each part i, which yields a different likelihood map $I_{\text{fg},i}$ per part (fig. 4d).

5 Extended model (ExMoPS)

In this section we show how to extend the MoPS model [10] to include the likelihood maps I_{bg},$I_{\text{fg},i}$ derived in sec. 4.3 as additional cues to restrict the location of body parts. For this we redefine the base model (1) to have multiple appearance models $a \in \{\text{gen}, \text{bg}, \text{fg}\}$

$$S(I,p,t) = \sum_a \sum_i w^{t_i}_{a,i} \Phi_a(I_{a,i}, p_i) + \sum_{ij \in \mathcal{E}} w^{t_i,t_j}_{ij} \Psi(p_i - p_j) + S(t) \tag{3}$$

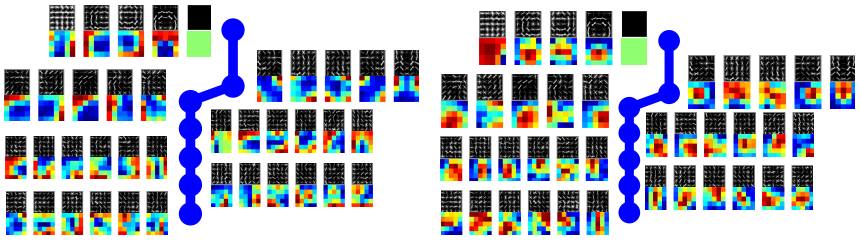
gen is the generic HOG appearance model of [10], whereas bg/fg are our foreground / background color appearance models shared over an image cluster (sec. 4).

In (1) there was one appearance template $w^{t_i}_i$ for each body part and type, where in our extended model (3) there are multiple $w^{t_i}_{a,i}$. Each term $w^{t_i}_{a,i}\Phi_a(I_{a,i}, p_i)$ is defined by a kind of template and a feature image $I_{a,i}$. In this notation the original appearance term of [10] is $w^{t_i}_{\text{gen},i}\Phi_{\text{gen}}(I_{\text{gen},i}, p_i)$, based on HOG templates $w^{t_i}_{\text{gen},i}$ applied to a gradient image I_{gen}. We define the new appearance terms $w^{t_i}_{\text{bg},i}\Phi_{\text{bg}}(I_{\text{bg},i}, p_i)$, $w^{t_i}_{\text{fg},i}\Phi_{\text{fg}}(I_{\text{fg},i}, p_i)$, where the templates $w^{t_i}_{\text{bg},i}, w^{t_i}_{\text{fg},i}$ are weight masks applied to background/foreground likelihood maps $I_{\text{bg},i}$,$I_{\text{fg},i}$ (fig 5). As mentioned in sec. 4.3, the same background likelihood map is used for all body parts $I_{\text{bg},i} = I_{\text{bg}}$, whereas the foreground likelihood map $I_{\text{fg},i}$ is part specific.

Inference. Inference in the extended model is analogous to the one for the base model (sec. 3). As we only introduced additional unary terms, the computation complexity of inference remains the same.

Learning. Introducing additional appearance terms keeps the model (3) in a form $S(I,z) = \beta \cdot \Theta(I,z)$, but $\beta = (w,b)$ now spans over multiple appearance templates per part and type. This leads to an equivalent learning problem as (2). However, we note that our shared appearance cues are defined only on images that contain persons (sec. 4). Therefore, instead of having a negative training set I^η containing no person [10], we sample negative examples from the positive images I^ρ (elsewhere than on the ground-truth stickmen).

Analogue to [10], we first train all appearance templates independently, and then use them as initialization to optimize the full model by dual coordinate-descent. This joint optimization of all parameters of the extended model enables to find an optimal balance between all terms, including multiple appearance cues and pose priors. Moreover, different appearance cues may be weighted differently for different body parts. If a



(a) (b)

Fig. 5. Extended MoPS appearance templates. *Learned appearance templates for the body parts along the path from forehead to left wrist. (a) appearance templates from the shared background model ExMoPS$\{gen, bg\}$; (b) appearance templates from the shared foreground model ExMoPS$\{gen, bg\}$. Top rows: generic appearance templates (HOG) defined on image gradients, one per type (orientation). Bottom rows: shared background (a) or foreground (b) appearance models defined on our color likelihood maps. Note how the expected outline of a body part is recognizable in the templates. Also note how the templates for background likelihoods have high weight in regions* around *parts rather than on the parts themselves.*

cue is not informative for a particular body part, it will get a low weight in the extended model. Note how this is different than a simpler solution that would keep the HOG templates $w^{t_i}_{\text{gen},i}$ as pre-trained in the base model, and then trains the new color templates $w^{t_i}_{\text{bg},i}, w^{t_i}_{\text{fg},i}$ on top of them. Fig. 5 shows ExMoPS models with shared foreground and background templates learned by by our method.

Missing data. Not all appearance terms may be available for every training image, e.g. the foreground sharing cue does not make sense on singleton foreground clusters (sec 4.2). One way out would be to train the model only to a subset of images where all appearance cues are available. However, this might lead to over-fitting and at test time it would require applying different models depending on the availability of cues for a particular test image.

Instead, we propose to replace the likelihood maps for the missing cues with null maps filled with a uniform value (e.g. 0 in foreground null maps). The learned model is then able to discard a cue when it is unavailable.

This effectively enables us to train the extended model on the *entire dataset* despite the missing appearance cues. At test time we have a single model which benefits from whatever cues are available for a particular image.

6 Experiments

Dataset. There are several data-sets for evaluating 2D HPE algorithms [1, 2, 4, 7, 9], some having specific shortcomings: [1] has only a small number of images, while [2, 4] have only moderate pose variability [8]. Hence, we focus the evaluation of our method on the *LSP dataset* [7]. With 2000 images, it is the largest dataset with fully accurate ground-truth annotations (as opposed to the even larger [9]). It is also considered one of the hardest datasets in terms of pose variability and background clutter [11]. The official

model	**avg**	t	lul	rul	lll	rll	lua	rua	lla	rla	h
LSP testset - full, person-centric annotations (PC)											
[7]	**55.1**	78.1	64.8	66.7	60.3	57.3	48.3	46.5	34.5	31.2	62.9
MoPS [10]	**50.9**	82.0	53.5	55.3	50.3	52.9	43.6	38.4	30.7	26.1	75.8
ExMoPS{gen}	**54.2**	83.5	59.5	61.4	54.1	58.5	45.3	42.7	31.3	28.7	77.1
LSP testset - full, observer-centric annotations (OC)											
MoPS [10]	**60.8**	84.1	69.5	69.4	64.8	66.4	53.1	51.6	37.3	34.5	77.1
ExMoPS{gen}	**63.7**	84.9	74.0	72.3	67.9	68.6	55.7	55.9	39.9	37.3	80.1
ExMoPS{gen, bg} 1img	**63.6**	86.2	74.9	73.7	68.5	68.1	55.4	54.9	38.0	36.3	80.1
ExMoPS{gen, bg}	**64.3**	86.5	75.6	74.1	68.9	69.8	57.5	55.4	38.7	36.0	80.1
ExMoPS{gen, fg}	**64.2**	85.6	75.2	72.5	68.3	68.0	56.6	56.6	38.4	39.7	80.4
LSP testset - foreground clusters (141 img), OC annotations											
ExMoPS{gen}	**70.0**	91.5	81.6	83.0	75.9	77.3	57.5	63.1	41.8	41.8	86.5
ExMoPS{gen, fg}	**72.5**	91.5	84.4	81.6	80.9	79.4	58.9	70.2	40.4	47.5	90.1

Table 1. PCP results on the LSP dataset. The **avg** column reports an average PCP over all the body parts. The remaining 10 columns show PCP for torso (t), left-right lower-upper leg-arm (lul, rul, lll, rll, lua, rua, lla, rla) and head (h).

protocol [7] has equal test and train splits of 1000 images each, covering various sport activities. This dataset is big enough for our approach to discover clusters of images with common background/foreground appearance.

The annotations in LSP are person-centric (PC), i.e. right/left body parts are marked according to the viewpoint of the person. The right ankle of a person facing the camera is *left* in the image, but it is *right* in the image if the person faces away from the camera. As we do not expect MoPS [10] nor any other state-of-the-art HPE model [1–6, 8, 10, 11] to distinguish between these two situations, we convert all annotations to observer-centric (OC)[5]. Using OC annotations helps reducing confusion during training, e.g. resulting in a more accurate pose prior. Note how OC annotations are by far the most widely used in the 2D HPE community [1–6, 8, 10, 11].

Evaluation Measure. We quantify performance using the PCP measure (Percentage of Correctly estimated Parts) introduced by [2] and used in many other works [6, 6, 4, 3, 7, 9, 10, 8, 26, 27]. An estimated part is considered correct if both its segment endpoints lie within 50% of the length of the corresponding ground-truth segment from their annotated location[6].

We follow the typical evaluation protocol used for datasets containing a single person per image [3, 7, 15, 27, 26] and evaluate *only* the MAP solution returned by the HPE model.

[5] we normalize the orientation of the body such that the torso is upright, and then we flip arms/legs according to shoulder/hips annotation points if necessary, so that the left limb is always on the left side of the normalized torso.

[6] [28] noted a discrepancy between PCP measures used across the HPE community, here we exactly follow the PCP measure used in [3, 7]

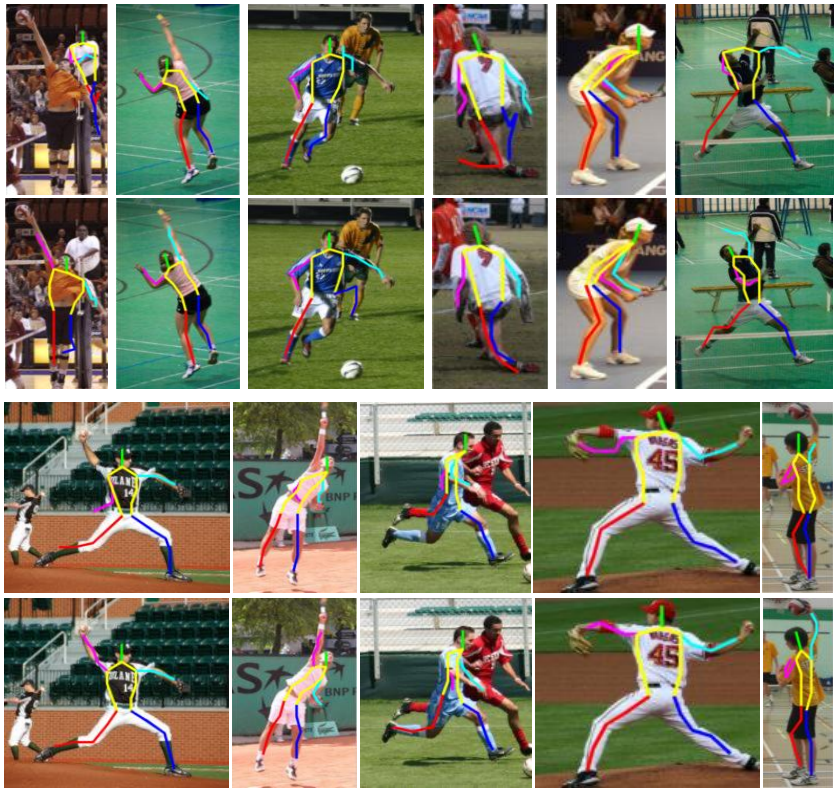


Fig. 6. Qualitative results on the LSP dataset. *Rows 1 and 3 show the initial pose estimate by MoPS, and rows 2 and 4 show the result of our extended models ExMoPS* $\{gen, bg\}$ *or ExMoPS* $\{gen, fg\}$*. Sharing background/foreground appearance models across images help to refine pose in a variety of situations.*

Setup. Following [10], we use a simplified deformation model for MoPS/ExMoPS with $w_{ij}^{t_i,t_j} = w_{ij}^{t_i}$, i.e. the deformation model depends only on the type of the child part only, not on the parent.

The full body MAP output from MoPS/ExMoPS trained/tested on the LSP dataset is a 26 part body configuration (joint locations and midpoints along the kinematic structure except of the head mid-point). As in [10], we convert it to 10 physical parts for PCP evaluation (torso, left-right lower-upper leg-arm and head; abbreviated from now on as t, lul, rul, lll, rll, lua, rua, lla, rla, h).

Results. Table 1 reports our results. We start by comparing the base MoPS model [10] to [7, 9]. These are the two works that have published results on LSP before, and they employed the original PC annotations. On these annotations, MoPS performs slightly below [7]. Following [10], MoPS mines negative training examples from negative part of the INRIA pedestrian dataset [18]. Instead, the extended ExMoPS$\{$gen$\}$, mines negative training examples from positive training images. It performs better then MoPS [10]

and it is on par with [7]. The method of [9] achieves higher PCP (62.7) but it is not directly comparable as it uses a much larger training set in addition to LSP (10000 images, so $11\times$ more training data). In summary, the baseline model we adopt ExMoPS$\{$gen$\}$ already performs on par with the state-of-the-art [7], even without appearance sharing between images.

In all following experiments we employ OC annotations, which we believe are more natural for HPE algorithms relying purely on low-level features [1, 4, 3, 6], as they may not be able to distinguish between front and back views of a person. On these OC annotations, the reference baseline MoPS reaches 60.8% PCP. For a fully transparent comparison, before we investigate the impact of our shared appearance models, we first evaluate MoPS when training from the alternative negative set as above (ExMoPS$\{$gen$\}$). With 63.7%, ExMoPS$\{$gen$\}$ achieves higher PCP than MoPS. As both use only the generic HOG appearance term, the difference is completely due the negative examples being better suited for evaluation on the LSP testset.

We are now ready to investigate the improvements brought by incorporating our new shared appearance models. Adding our new background sharing cue yields an improvement to 64.3% PCP (ExMoPS$\{$gen, bg$\}$). An interesting experiment is to remove the sharing element from this cue by enforcing each image to be in its own cluster (ExMoPS$\{$gen, bg$\}$ 1img). This degenerates to a technique analogue to [1], where a background color model is estimated from a *single image*, but integrated into MoPS. The performance of this model drops to the level of ExMoPS$\{$gen$\}$, demonstrating that the improvement brought by our background sharing is truly due to *sharing between images*. It supports our claim that collective HPE by sharing background models improve over independent single-image pose estimation, even when augmented with an analogue background term. Importantly, our method discovers between which images to share *fully automatically*.

We now investigate our new foreground sharing, which we apply to foreground clusters with 3 or more images, as foreground sharing is undefined on a single image. These clusters contain 141 of the 1000 test images. On these images, the foreground sharing model ExMoPS$\{$gen, fg$\}$ improves PCP performance by +2.5% over the best baseline ExMoPS$\{$gen$\}$. We can also evaluate foreground sharing on the *entire* LSP dataset, by adding a null cue to images in smaller clusters, as described in 5. This also improves performance (64.2%) compared to the best baseline (63.7%), which demonstrates our foreground sharing is a useful new component for HPE. Interestingly, both our newly proposed foreground and background sharing methods achieve similar PCP performance (and they are both equally fully automatic).

Finally, we also investigated a method combining both foreground and background appearance sharing ExMoPS$\{$gen, bg, fg$\}$, its performance however turns out on par with methods using either of the shared components.

Conclusions. We have presented a novel technique to perform human pose estimation over multiple images by sharing foreground/background appearance models. As demonstrated on the highly challenging *Leeds Sports Pose* datasets, our collective pose estimation via appearance sharing improves performance over the baseline method [10] applied independently on each image. In future work we plan to share more elements between images, e.g. texture appearance models.

References

1. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. (2006)
2. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009)
4. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC. (2009)
5. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: CVPR. (2010)
6. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: ECCV. (2010)
7. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. (2010)
8. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: ECCV. (2010)
9. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. (2011)
10. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
11. Andriluka, M., Sigal, L., Black, M.J. In: Benchmark Datasets for Pose Estimation and Tracking. Springer Verlag (2011)
12. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Efficient discriminative learning of parts-based models. In: ICCV. (2009)
13. Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. In: CVPR. (2008)
14. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV. (2011)
15. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV (2005)
16. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: CVPR. (2004)
17. Jiang, H.: Human pose estimation using consistent max-covering. In: ICCV. (2009)
18. Dalal, N., Triggs, B.: Histogram of Oriented Gradients for Human Detection. In: CVPR. (2005)
19. Bissacco, A., Yang, M.H., Soatto, S.: Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: CVPR. (2007)
20. Sapp, B., Weiss, D., B.Taskar: Parsing human motion with stretchable models. In: CVPR. (2011)
21. Park, D., Ramanan, D.: N-best maximal decoders for part models. In: ICCV. (2011)
22. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. on PAMI (2010)
23. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. John Wiley and Sons (2001)
24. Graham, R.L., Groetschel, M., Lovasz, L.: Handbook of Combinatorics. Elsevier (1995)
25. Ferrari, V., Tuytelaars, T., Van Gool, L.: Real-time affine region tracking and coplanar grouping. In: CVPR. (2001)
26. Tian, T.P., Sclaroff, S.: Fast multi-aspect 2d human detection. In: ECCV. (2010)
27. Singh, V.K., Nevatia, R., Huang, C.: Efficient inference with multiple heterogeneous part detectors for human pose estimation. In: ECCV. (2010)
28. Pishchulin, L., Jain, A., Andriluka, M., Thormaehlen, T., Schiele, B.: Articulated people detection and pose estimation: Reshaping the future. In: CVPR. (2012)