Better Appearance Models for Pictorial Structures

Better appearance models for pictorial structures

Marcin Eichner
eichner@vision.ee.ethz.ch
Vittorio Ferrari
ferrari@vision.ee.ethz.ch

Computer Vision Laboratory
ETH
Zürich, Switzerland

Abstract

We present a novel approach for estimating body part appearance models for pictorial structures. We learn latent relationships between the appearance of different body parts from annotated images, which then help in estimating better appearance models on novel images. The learned appearance models are general, in that they can be plugged into any pictorial structure engine. In a comprehensive evaluation we demonstrate the benefits brought by the new appearance models to an existing articulated human pose estimation algorithm, on hundreds of highly challenging images from the TV series *Buffy the vampire slayer* and the PASCAL VOC 2008 challenge.

1 Introduction

Pictorial structures (PS) [10, 21, 23] are a popular paradigm for articulated pose estimation. Although PS are typically used for humans [10, 11, 21, 22, 23], they are well suited for any articulated object class (e.g. cows [16] and horses [21]). PS are probabilistic models where objects are made of parts tied together by pairwise potentials carrying priors over their spatial relations (e.g. kinematic constraints). The local image likelihood for a part to be in a particular position is measured by a unary potential carrying an appearance model of the part (e.g. the torso is red). Inference in a PS involves finding the MAP spatial configuration of the parts, i.e. the *pose* of the object.

The success of PS depends critically on having good appearance models, which constrain the image positions likely to contain a part. Because of their importance, previous works have put great care in estimating appearance models. The most reliable way, but the least automatic, is to derive them from *manually segmented* parts in a few video frames [2]. Another approach is to apply *background subtraction*, and use the number of foreground pixels at a given position as a unary potential [10, 17, 18]. The *strike-a-pose* work [22] searches all frames for a predefined characteristic pose, easier to detect than a general pose. In this pose all parts are visible and don't overlap, enabling to learn good appearance models, which are then used to estimate pose in all other frames (as part appearance is stable over time).

The above strategies cannot be applied to a single image as they require video. Moreover, background subtraction is only reliable on static backgrounds, and strike-a-pose is limited to videos containing a predefined characteristic pose. In an effort to operate on a single image, with unknown part appearances, Ramanan [21] proposes *image parsing*, where inference is first run using only generic edge models as unary potentials. The resulting pose is used to build appearance models specific to this particular person and imaging conditions, and

inference is repeated using both edges and appearance. Ferrari et al. [11] extend this approach with a preprocessing stage called *foreground highlighting*, which removes part of the background clutter to restrict the space parsing needs to search for body parts. Foreground highlighting depends on a generic detector to find the location and scale of the person in the image.

In this paper, we present a new approach for estimating part appearance models from a single image. As in recent pose estimation works [4, 11, 14], we use a generic detector to determine an approximate location and scale reference frame on the object. The whole object [8, 19] needs not be detected, as a part of it is sufficient (e.g. a person's face [28] or head-and-shoulder profile [11]). Two observations motivate our approach: (i) relative to the reference frame, some parts have rather stable location (e.g. the torso is typically below the face); (ii) the appearance models of different parts are statistically related. For example, the lower arms of a person are colored either like the torso (clothing) or like the face (skin). Only rarely they have an entirely different color. The legs of a horse have the same color as its torso, as the whole horse is covered by the same fur. This implies that the appearance of some parts can be predicted from the appearance of other parts.

We learn the relative location distribution of parts wrt the reference frame and the dependencies between the appearance of different parts from training data. These relations are exploited to generate appearance models for body parts in a new image. In this fashion, parts which are well localized wrt to the reference frame (e.g. torso) help determining the appearance model for more mobile parts (e.g. lower arms). If no inter-part dependencies exist, our approach naturally degenerate to estimating each part independently.

2 Pictorial structure framework

We briefly review here the general framework of pictorial structures (PS) for human pose estimation. A person's body parts are tied together in conditional random field. Typically, parts l_i are rectangular image patches and their position is parametrized by location (x,y), orientation θ, scale s, and sometimes foreshortening [7, 10]. The posterior of a configuration of parts $L = \{l_i\}$ given an image I is

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

The pairwise potential $\Psi(l_i,l_j)$ corresponds to a prior on the relative position of parts. It embeds kinematic constraints (e.g. the upper arms must be attached to the torso) and, in a few works, other relations such as occlusion constraints [26] or coordination between parts [18]. In many works the model structure E is a tree [10, 11, 21, 22, 23], which enables exact inference, though some works have explored more complex topologies [6, 7, 18, 26, 27, 30]. Inference returns the single most probable configuration L^* [6, 10], or posterior marginal distributions over the position of each part [11, 21]. The unary potential $\Phi(l_i|I,\theta)$ corresponds to the local image evidence for a part in a particular position (likelihood). It depends on appearance models θ describing how parts should look like. It computes the dissimilarity between the image patch at l_i and the appearance model for part i. The appearance models are parameters of the PS and must be provided by an external mechanism (see section 1).

A few recent works [4, 11, 14] first run a generic detector [8, 19] to find the approximate location and scale (x,y,s) of the person, and then run pose estimation only within the detection window. This reduces the search space for body parts, making it possible to tackle complex, highly cluttered images.

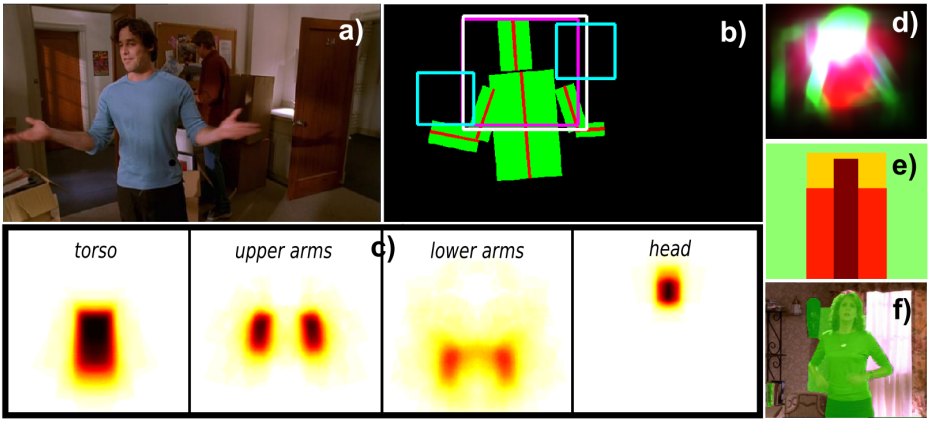

Figure 1: **Learning location priors.** *(a) A training image. (b) Detection windows (cyan), expected window hallucinated from the stickman (magenta), window associated to the stickman (white), body part rectangles (green) obtained by widening the stickman line segments (red). (c) Learnt location priors. By estimating left/right arm parts together we increase the number of training examples (this exploits the appearance similarity of symmetric parts, as done in [7, 15, 27]). (d) Pose estimate returned by [27] for the image in (f) (without using foreground highlighting). (e) Initialization regions for foreground highlighting. (f) Original image with foreground highlighting superimposed.*

3 Better appearance models

In this paper we present a new method for estimating good body part appearance models from a single image. These can be then plugged into any PS engine. For the experiments (section 6) we build a full pose estimation system by plugging our appearance models into [21].

The input to our method are candidate detection windows output by a person (part) detector. In the case of a detector tuned to a part (e.g. the upper-body detector of [11], or a face detector [13, 28]), we enlarge the window by a predefined factor, to make it cover the whole person (as done in [11]). For simplicity, in the remainder of the paper we say *detection window* to indicate this enlarged area.

Our approach is motivated by two main observations: (i) the location of some parts relative to the detection window $W = (x, y, s)$ is rather stable (e.g. the torso is typically in the middle of an upper-body detection window); (ii) the appearances of different body parts are related (e.g. the upper-arms often have the same color as the torso).

As the two observations hold in a statistical sense, we learn (i) a location prior capturing the distribution of the body part locations relative to W (section 3.1); (ii) an appearance transfer mechanism to improve the models derived from the location prior by combining models for different body parts (section 3.2). The training data consists of images with ground-truth pose annotated by a stickman, i.e. a line segment for each body part (figure 1b).

After learning, our method is ready to estimate appearance models on new, unannotated test images (section 3.3). Initial appearance models are estimated given W and the learnt location priors. These models are then refined by the appearance transfer mechanism.

While we present our approach on human upper-bodies, it can be applied to any object class for which a detection window can be provided (e.g. human full bodies [8, 19], horses [25], sheep [9])

3.1 Training: learning location priors

For each body part i, we learn a *location prior* $LP_i(x, y) \in [0, 1]$: the prior probability for a pixel (x, y) to be covered by the part, before considering the actual image data (figure 1a).

	torso	upper arms	lower arms	head
torso	1	0.11	0.16	0
upper arms	0	0.89	0.31	0
lower arms	0	0	0.34	0
head	0	0	0.19	1

Table 1: **Learned appearance transfer weights.** *Each entry w_{it} denotes the contribution of part i (row) to the appearance model of part t (column).*

Importantly, pixel coordinates are relative to the detection window, so that LPs can be employed later on test images. Thanks to LPs, we can estimate initial appearance models before running a pictorial structure inference (as opposed to [21]). As in our implementation appearance models are color histograms $P_i(c|fg)$, they are obtained by weighting pixel contributions by $LP_i(x,y)$ (details in section 4).

We learn LPs from training images with ground-truth pose annotated by a stickman (figure 1a). We first obtain detection windows by running the generic object detector on these images. Next, we associate stickmen to detection windows as in figure 1b. Based on the detection windows, we now project all training stickmen to a common coordinate frame, where they are roughly aligned in location and scale. In this common coordinate frame, the LPs are learnt in maximum likelihood fashion: $LP_i(x,y)$ is the fraction of training images where part i covers pixel (x,y). LPs are computed for every pixel in the detection window.

Example LPs are presented in figure 1c. LPs for the head and torso are quite sharply localized, while LPs for the arms are more diffuse. Interestingly, the location of lower arms appears very uncertain a priori, matching our expectation that they can move around freely.

Notice how LPs are learned in the coordinate frame obtained by actually running the object detector on the training images, as opposed to deriving ideal detection windows from the stickmen. This procedure delivers realistic LPs, tuned to the behavior we expect at test time, as they already account for the uncertainty in the localization of the detection window.

3.2 Training: transferring appearance models between body parts

Given an image of a person with lower arms behind her back, can we predict their color based on the visible body parts ? Intuitively, we can, because we know that usually people wear either a rather uniformly colored pullover with long-sleeves, in which case the lower arms are colored like the torso, or wear short sleeves, in which case the lower arms have skin color (the same as the face). While external factors might help our reasoning, such as scene type (e.g. beach vs office) and season (winter vs summer), our ability to predict is rooted in the intrinsic relations between the appearance of different body parts.

Inspired by the power of the above relations, here we learn a transfer mechanism to combine the appearance models of different body parts. The input appearance models are derived from LPs (section 3.1). The appearance transfer mechanism estimates the new appearance model of a part as a linear combination of the input appearance models of all parts.

Learning mixing weights. The new appearance model AM_i^{TM} for a part t is given by

$$AM_t^{TM} = \sum_i w_{it} AM_i^{LP} \tag{2}$$

where w_{it} is the mixing weight of part i, in the combination for part t, and AM^{LP} is the initial appearance model (derived from the location prior).

The parameters of the transfer mechanism are the mixing weights w_{it}. We learn them by minimizing the squared difference between the appearance models produced by the transfer

mechanism (AM_i^{TM}) and those derived from the ground-truth stickmen (AM^{GT}):

$$\min_{w_t} \quad \sum_s \sum_k \left(\sum_i w_{it} AM_{ski}^{LP} - AM_{skt}^{GT} \right)^2 \tag{3}$$
$$s.t: \qquad 0 \le w_{it} \le 1, \qquad \sum_i w_{it} = 1$$

where i runs over all parts, s runs over training samples, and k runs over the components of the appearance model (entries of a color histogram, in our case). Ground truth color histograms are computed over rectangular part masks obtained by widening the line segments of the annotated stickman by a predefined factor (figure 1b). Since this is a quadratic optimization problem with linear inequality constraints, we find its global optimum efficiently using quadratic programming [20]. The mixing weights w_t are found for each part t separately by solving a new quadratic program (3) for each part.

Table 1 shows the mixing weights learnt based on the location prior of figure 1c. Two interesting observations can be made: (i) for parts that are rather stationary wrt the detection window (torso, head), the refined appearance model is identical to the input model from LP; (ii) mobile parts benefit from the contribution of stationary parts with similar appearance. Upper arms models are improved by appearance transfer from the torso. Lower arms, which have the highest localization uncertainty, get strong contribution from all other parts. This because people tend to either wear uniformly colored clothes with long sleeves (contribution from upper arms and torso), or wear short sleeves (contribution from head, which is also skin-colored). These results confirm our intuition that exploiting relations between the appearance of different body parts leads to better appearance models.

3.3 Test: estimating appearance models for a new image

After learning LPs and mixing weights, our method is ready to estimate good appearance models for new test images. For clarity, we explain the procedure here for the case where appearance models are color histograms. However, our scheme can be applied for other appearance models as well, such as texture histograms.

Color models. The procedure entails three steps. First, the detection window W is transformed to the standard coordinate frame where the LPs were learned from, by cropping W out of the image and rescaling it to a fixed size. Second, initial color models are derived from the LPs, as described in section 4. Third, the color models are refined by applying appearance transfer as in equation (2), leading to the final color models $P_i(c|fg)$.

Soft-segmentations. The color models estimated above characterize the appearance of the body parts themselves. Following [21], we also estimate here a background model $P_i(c|bg)$ for each body part, derived from the complement of the LP (i.e. $1 - LP_i(x,y)$). The foreground $P_i(c|fg)$ and background $P_i(c|bg)$ models are used to derive the posterior probability for a pixel to belong to a part i (using Bayes theorem, assuming $P_i(fg) = P_i(bg)$)

$$P_i(fg|c) = \frac{P_i(c|fg)}{P_i(c|fg) + P_i(c|bg)} \tag{4}$$

The posterior foreground probabilities are then used to derive a soft-segmentation of the image for each body part, which is the cue used in the unary term of the pictorial structure (Φ in equation (1), details in section 4).

4 Computing appearance models and soft-segmentations

Our implementation uses color histograms as appearance models. These can be derived from a (soft-)segmentation of the image, and vice-versa.

Color models from soft-segmentations. The contribution of a pixel (x,y) with color c to the histogram is weighted by the value of the soft-segmentation at that pixel. We also apply trilinear interpolation to let each pixel vote into multiple histogram bins. Each pixel contributes to the eight bins closest to c. When estimating color models AM^{LP} in equation (2), the location prior LP_i is used as the soft-segmentation. When estimating ground-truth color models AM^{GT}, we use the segmentation defined by the widened stickman (figure 1b).

Soft-segmentations from color models. The color models are used to derive foreground posteriors $P_i(fg|c)$, as in section 3.3. A soft-segmentation is obtained from them by applying the inverse of the procedure above. The value of the segmentation for a pixel with color c is interpolated over $P_i(fg|c)$ for neighbouring colors. This soft-voting procedure makes histograms more robust to color and brightness variation.

5 Image parsing and extensions

We briefly introduce here the *image parsing* PS engine [21] and its extensions [11]. In the experimental section we will compare to both methods. Moreover, we evaluate the impact of our appearance models on pose estimation by plugging them into the image parsing engine.

a) Image parsing [21]. In Ramanan's work [21], body parts l_i are oriented patches of fixed size, with position parametrized by location (x,y) and orientation θ (equation (1)). The model structure E is a tree with edges $\Psi(l_i,l_j)$ carrying kinematic constraints. Since the parts' appearance is initially unknown, a first inference uses only edges in the unary potential Φ. A soft-segmentation for each body part is obtained from the resulting marginal distribution over the part position, by convolving it with a rectangle representing the body part (figure 1d). Color histograms for the part and background are then derived from the soft-segmentations. Finally, inference in repeated with Φ using both edges and color.

In this scheme, the first inference stage is the mechanism to obtain appearance models. Unfortunately, the edge-based model is not specific enough and the first inference stage typically fails in the presence of sufficiently cluttered background [11], leading to poor appearance models and, eventually, incorrect pose estimation.

b) Extensions [11]. Ferrari et al. [11] extend [21] with two pre-processing stages aiming at reducing the search space for body parts: (1) *detection*: find the location and scale of the person with a detector generic over appearance and pose [8]; (2) *foreground highlighting*: apply Grabcut [24] within the detection window to exclude part of the background clutter. The initial foreground and background regions for Grabcut are manually designed to be likely to contain the head and torso (for foreground) and away from this area for background (figure 1e). Only the image region returned by foreground highlighting is passed on to parsing (figure 1f).

While these extensions resulted in a more robust system, working in heavily cluttered images, the initialization regions for foreground highlighting are manually tuned to upright human upper-bodies, requiring to be re-designed for every new object class (e.g. full-bodies, sheep). Moreover, parts lost by foreground highlighting cannot be recovered during parsing.

c) Our pose estimator. In order to evaluate our appearance models, we use the following pose estimation procedure: (1) detect windows on people using the detector of [11]; (2) estimate part-specific color models as in section 3.3; (3) run image parsing [21] within the detection window using directly our color models in the unary potential (i.e. skipping the initial edge-based inference). Note that location priors are only involved in the estimation of color models, and are *not* used to constrain the position of parts during parsing.

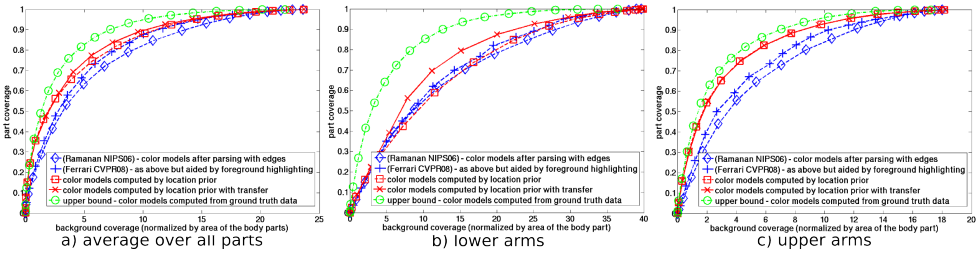

Figure 2: **Evaluation of segmentation induced by color models.** *Each curve is averaged over all images in the Buffy test set (episodes 2,5,6). Points on the curve are obtained by thresholding to the soft-segmentation with an increasing threshold. The Y-axis shows how much of the area A of the ground-truth rectangle for a part is covered by the segmentation, in percentage. The X-axis shows how much of the segmentation lies out of the ground-truth rectangle (i.e. on another part or on the background), in multiples of A.*

6 Experiments and Conclusions

We present a comprehensive evaluation on two levels: (i) the quality of the soft-segmentations derived from the proposed appearance models; (ii) their impact on pose estimation.

Datasets. We experiment on video frames from the 'Buffy: the vampire slayer' TV show [11] and still images from the PASCAL VOC 2008 challenge [9]. We use annotated stickman data from episodes 2-6 of Buffy's season 5, for a total of 748 annotated frames (available for download [1]). This data is challenging due to the uncontrolled conditions, with very cluttered images, often dark illumination, persons appearing at a wide range of scales and wearing clothing of any kind and color (figure 3). The PASCAL data is even more demanding, as it consists mainly of amateur photographs with difficult illumination and low image quality (figure 3). We have annotated 549 images [1] in the same way as the Buffy data set: in each image one roughly upright, approximately frontal person is annotated by a 6-part stickman (head, torso, upper and lower arms). The person must be visible at least from the waist up.

Below we compare to [11] on their test set, consisting of Buffy episodes 2,5,6, and learn location priors and appearance transfer weights from Buffy episodes 3,4 and PASCAL images (this gives the LPs of figure 1c). When testing on the PASCAL images instead, we re-train from all 5 Buffy episodes.

Soft segmentation. We compare the quality of part-specific soft-segmentations derived from appearance models generated by several approaches (figure 2) on the Buffy test set. These segmentations are important for pose estimation, as they form the unary term Φ of the pictorial structure equation (1). We compare our method to three alternative approaches for estimating color models: (a) edge-based parsing [21]; (b) edge-based parsing aided by foreground highlighting [11]; (c) derive color models from the widened ground-truth stickmen (AM^{GT} in the equation (3)) – this provides an upper bound on the quality of the segmentation that can be achieved with this kind of appearance models. For all approaches, we derive a soft-segmentation from the color models as detailed in section 4. All approaches start from detection windows obtained by the upper-body detector of [3, 11].

As figure 2a shows, on average over all body parts, we obtain better segmentations than the competing methods already from the initial color models based on location priors (section 3.1). Results improve further after appearance transfer (section 3.2). Interestingly, the color models generated from LPs produce a rather poor segmentation of the lower arms,

[1] We released this new dataset on the web [1].

	[10]	LP	LP+AT	LP+AT+FGH	[12]	LP+AT+FGH
Buffy	61.4%	65.5%	69.5%	72.2 %	74.5 %	78.1 %
PASCAL	52.2%	53.2%	54.5%	59.1 %	64.4 %	67.5 %

Table 2: **Evaluation of pose estimation.** *Each entry reports correctness (always wrt to the new evaluation protocol, see main text). The first four columns use the PS model of [12]:* **[10]***: the method of [12];* **LP***: our system using color models based on location priors (no appearance transfer);* **LP+AT***: our system using also appearance transfer.* **LP+AT+FGH***: our full system aided by foreground highlighting. The last two columns use the enhanced PS model of [12]:* **[12]***: the method of [12];* **LP+AT+FGH***: our full system aided by foreground highlighting.*

which have the most diffuse LP (figure 2b). However, segmentation performance improves substantially after refining the color models by appearance transfer, surpassing the competing approaches. As figure 2c shows, we obtain a considerable improvement also for upper arms. Arms are especially interesting because they move more than head and torso wrt the detection window, making their pose harder to estimate, and because they carry most of the semantic pose information necessary to recognize gestures and actions. Importantly, even the ground-truth color models don't lead to perfect segmentation, because the same color might occur on another body part or on the background. On average over all parts, the segmentations derived from our color models are not far from the upper bound (figure 2a). The largest margin left for improvement is for the lower arms (figure 2b).

As a side note, figure 2 also shows that foreground highlighting helps [21] finding better appearance models, thus providing a deeper explanation for the improved pose estimation results reported by [10].

Pose estimation. We evaluate the impact of the proposed appearance models on pose estimation performance on the Buffy and PASCAL test sets. Performance is measured by *PCP*: the Percentage of Correctly estimated body Parts. Following the criterion of [10], an estimated body part is considered correct if its segment endpoints lie within 50% of the length of the ground-truth segment from their annotated location. PCP is evaluated only for stickmen that have been correctly localized by the initial upper-body detector (according to the standard intersection-over-union $> 50\%$ criterion from the PASCAL VOC challenge; this is the same criterion used to associate detections to stickmen when learning LPs, figure 1b). This protocol allows to cleanly evaluate the person detection and pose estimation tasks separately [2] . Note how the pose estimation algorithm we use [21] operates on individual video frames, ignoring the temporal dimension.

We compare our new pose estimator (section 5c) to [10] without multi-frame stages [3] (section 5b), which did not bring an improvement in later investigation [29]. On the Buffy dataset, the upper-body detector [10] correctly detects 85% of the 276 annotated stickmen. PCP is evaluated on these 235 frames. As table 2 shows, the color models from our complete method (LP+AT) raise PCP by 8.1% over [10]. Interestingly, color models from LPs alone already lead to a 4.1% gain, confirming that both ideas contribute to the overall improvement. Finally, adding foreground highlighting [10] to our pose estimator (as in section 5b), further raises performance, giving an overall improvement of 10.8% over [10].

[2] We modified the protocol used in [10], to make it simpler, more reproducible, and defined fully on still images (not video). In [10] we computed PCP over all images with *any* detection, not necessarily on the stickman. Moreover, [10] tracked detections over time, and the track with the largest number of correct body parts was considered in the evaluation. This under-estimates PCP of persons split over two tracks. In the new protocol, each image is evaluated independently, and only detections on the stickman are considered. Note how the new protocol corresponds to the one in the independent work of [6]. A Matlab implementation of the evaluation routine is available [1].

[3] Performance is higher (61.4%) on Buffy than what reported in [10, 29] (57.9%) due to the new protocol.

Figure 3: **Example results.** *Color coding: head = purple, torso = red, upper arms = green, lower arms = yellow. (a1-3) a few failures of [] on Buffy. (b1-3) the corresponding improvements brought by our LP+AT+FGH method. (a4-5) two more examples of our method on Buffy. The rest of the figure covers the PASCAL dataset. Notice the variety of poses and how our method makes no assumption about skin color (e.g. image c-5 contains a dark-skinned person), although only Caucasians are in the corresponding training set of Buffy images). (a6) a failure due to occlusion of the upper arms; (b6) a failure due to a wrong scale estimate of the detector; (c6) interestingly, our algorithm confuses the lower arm of the leftmost person with his neighbour's leg.*

For the PASCAL dataset we complement the upper-body detector [] with a multi-scale version of the face detector proposed by []. This allows to detect more people in this very challenging dataset, featuring harder imaging conditions and a wider variety of poses than Buffy (figure 3d). Overall, 73.1% of the 549 annotated persons were correctly detected. Pose estimation performance is evaluated on these images only. We start the pose estimator from each detection window independently. If there is more than one correct detection window for a stickman (e.g. one from the face and one from the upper-body detector), we consider the pose with the largest number of correct body parts. As table 2 shows, our full system (LP+AT+FGH) improves over [] by 6.9% correctly estimated body parts.

Recently we have proposed an enhanced PS model []. Employing the new appearance models of this paper in conjunction with the PS model of [], improves also over []

(table 2). Finally, we obtain a last performance boost by learning the proper scale factor between a detection window and the body part sizes expected by the PS model. This leads to 72.3% on PASCAL and 80.3% on Buffy (compared to about 74% by both works [5, 12]).

Conclusions. We have presented a new approach for estimating appearance models from a single image, and demonstrated experimentally that they considerably improve the performance of an existing PS engine [11, 12, 21] on two uncontrolled, very challenging datasets [9, 10]. We obtain better performance on the Buffy dataset than the two different state-of-the-art approaches [5, 12]. In future work we plan to include the more distinctive body part detectors of [5], and tackle open issues such as occluded body parts and joint pose estimation of multiple nearby people (which can confuse the pose estimator).

References

[1] http://www.robots.ox.ac.uk/∼vgg/data/stickmen/index.html.

[2] http://www.vision.ee.ethz.ch/∼vferrari/datasets.html.

[3] http://www.robots.ox.ac.uk/∼vgg/software/upperbody/index.html.

[4] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

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

[6] M. Bergtholdt, J. Knappes, and C. Schnorr. Learning of graphical models and efficient inference for object class recognition. In *DAGM*, 2008.

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

[8] N. Dalal and B Triggs. Histogram of Oriented Gradients for Human Detection. In *CVPR*, volume 2, pages 886–893, 2005.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html, 2008.

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

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

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

[13] B. Froba and A. Ernst. Face detection with the modified census transform. In *In the Proc. of the sixth IEEE international conference on automatic face and gesture recognition*, 2004.

[14] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, and L. Van Gool. Articulated multi-body tracking under egomotion. In *ECCV*, 2008.

[15] Jiang H. and Martin D. R. Global pose estimation using non-tree models. In *CVPR*, 2008.

[16] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *ICCV*, 2005.

[17] X. Lan and D. P. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *CVPR*, volume 1, pages 722–729, 2004.

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

[19] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[20] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag, 2006.

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

[22] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, volume 1, pages 271–278, 2005.

[23] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV*, 2002.

[24] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[25] J. Shotton, A. Blake, and R. Cipolla. Contour-Based Learning for Object Detection. 2005.

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

[27] L. Sigal, M. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *NIPS*, 2003.

[28] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *CIVR*, 2005.

[29] Ferrari V., Marin M., and Zisserman A. 2d human pose estimation in tv shows. In *Dagstuhl post-proceedings*, 2009.

[30] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008.