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

Pictorial structures (PS) are a popular paradigm for articulated pose estimation. 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.

The success of PS depends critically on having good appearance models. Because of their importance, previous works have put great care in estimating appearance models. As in this work we are interested in fully automatic pose estimation in single images for persons of unknown appearance, neither manual segmentation [3] nor background subtraction [6] are viable. Ramanan [10] 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. [7] extend this approach with a preprocessing stage that removes part of the background clutter to restrict the space parsing needs to search for body parts.

In this paper, we present a new approach for estimating part appearance models from a single image. As in recent pose estimation works [1, 7, 9], we use a generic detector to determine an approximate location and scale reference frame on the object. 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.

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 the detection window 1a (ii) an appearance transfer mechanism to improve the models derived from the location prior by combining models for different body parts 1b. The training data consists of images with ground-truth pose annotated by a stickman. After learning, our method is ready to estimate appearance models on new, unannotated test images. Initial appearance models are estimated given the detection window and the learnt location priors. These models are then refined by the appearance transfer mechanism. 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. 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 [4], horses [11], sheep [5])

We present a comprehensive evaluation both of the quality of the soft-segmentations derived from the proposed appearance models and of their impact on pose estimation. We report results on the challenging 'Buffy: The Vampire Slayer' dataset [7] and on a set of newly annotated still images from PASCAL VOC 2008 [5]. The experiments show that our technique improves in terms of soft-segmentation quality over both approaches [7, 10] and that it is close to the best possible quality, achievable using appearance models derived from ground-truth stickmen. Moreover, by plugging our appearance models in the pose estimator [8] we obtain 80.3% and 72.3% correctly estimated body parts on the Buffy and PASCAL datasets respectively. This represents an improvement of about $6-7\%$ over the state-of-the-art on Buffy [2, 8].

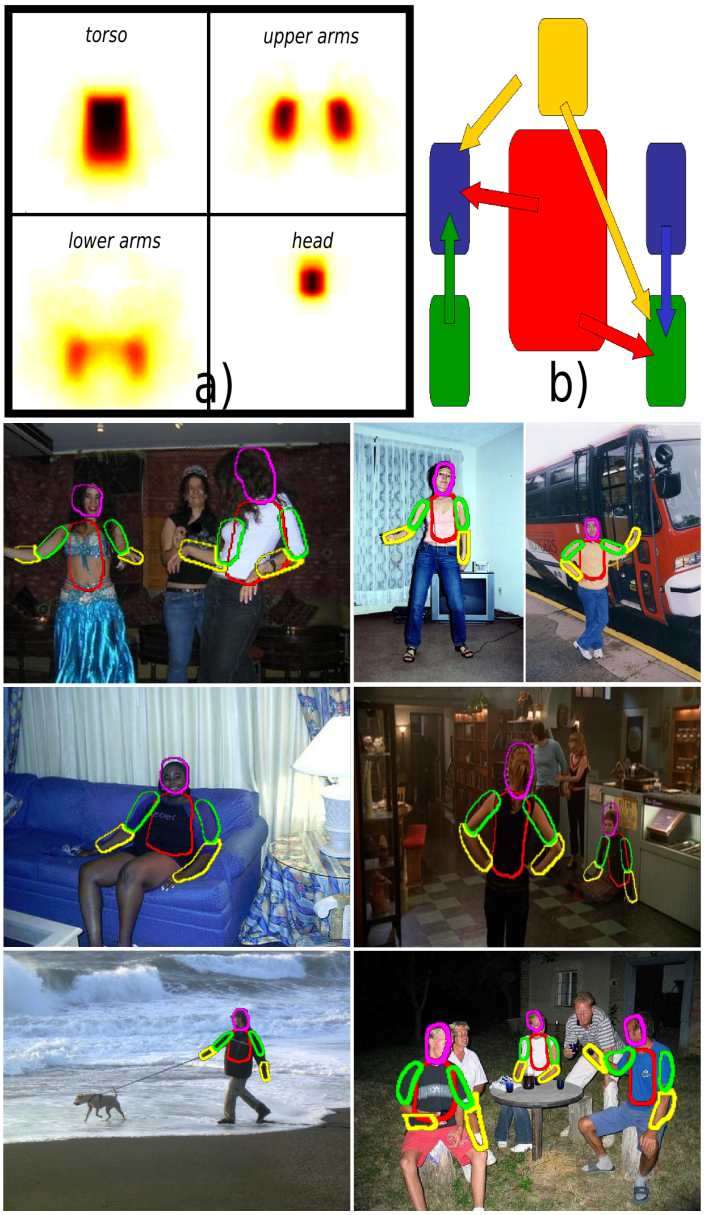

Figure 1: **Training stage and test time examples** *a) An example of Location Prior b) Visualization of appearance transfer mechanism* Images on the bottom present some interesting examples of output of our system at the test time.

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

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

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

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

[5] 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.

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

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

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

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

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

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