International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Pose-Guided and Scale-Aware Human Semantic Part Segmentation in Natural Multi-Person Scenes

Fangting Xia · **Peng Wang** · **Alan Yuille**

Received: date / Accepted: date

Abstract Parsing articulated objects like humans into semantic part regions (e.g. head, body and arms, etc.) from a natural image is a fundamental yet challenging problem in computer vision. The major difficulties of this problem derive from multi-instance confusion and large variability in human pose and scale. Current state-of-the-art methods use deep neural networks to predict part labels directly, and then refine the labels by a graphical model such as the dense CRF. These methods are still limited in complex natural scenes because they have no efficient mechanisms to handle multi-person overlapping or to adapt to the scale of human instances. In this work, we propose a part segmentation framework that handles these hurdles effectively. Our framework is scale-aware: we design an hierarchical "auto-zoom" strategy to allow the model to adapt to the size of human instances and their corresponding parts. Our framework is also pose-guided: it predicts human pose and part segmentation jointly, letting the two tasks benefit each other. We extend the PASCAL VOC part datasets with pose joints and perform extensive experiments on it. We show that our method achieves state-of-the-art part segmentation performance, and is especially better at handling small human instances and small parts.

Keywords semantic part segmentation · pose estimation · auto-zoom

Fangting Xia
Google Inc.
E-mail: sukixia@gmail.com

Peng Wang
Baidu Inc.
E-mail: pengwangpku2012@gmail.com

Alan Yuille
Johns Hopkins University
E-mail: alan.yuille@jhu.edu

Fig. 1 Challenges for current human pose estimation algorithms (top row) and semantic part segmentation algorithms (bottom row) in natural multi-person scenes.

1 Introduction

When people look at natural images, they often first locate regions that contain objects, and then zoom in or out on the object regions to perform the more detailed task of object semantic part segmentation, i.e. decomposing each object instance region into its semantic parts (e.g. head, body, lower-arms, etc.). A closely correlated task to object semantic part segmentation is object pose estimation, which aims to predict the position of joints (e.g. forehead, neck, left shoulder, etc.) for each object instance. Though closely correlated, semantic part segmentation [1–4] and pose estimation [16–18] are studied individually most of the time. They are both crucial to object interaction understanding and many high-level tasks, e.g. fine-grained recognition [19–21], action recognition [22–24], person identification [25,26], and video surveillance [27,28].

Currently, both semantic part segmentation and pose estimation face unsolved challenges in natural multi-person scenes (see Fig. 1). Traditional semantic part segmentation

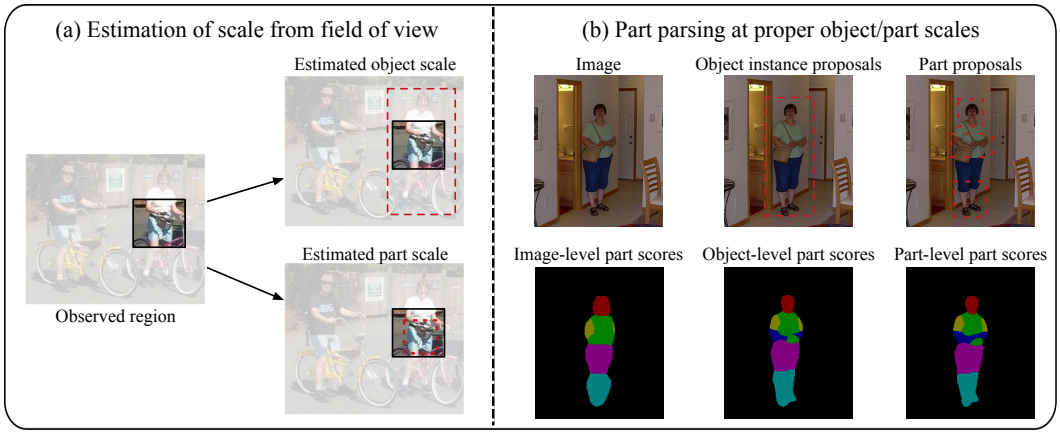

Fig. 2 Intuition of our hierarchical auto-zoom model (HAZN), which handles the large variation of object and part scale in natural multi-person scenes. (a) The scale and location of an object and its parts (the red dashed boxes) can be estimated from the observed field of view (the black solid box) of a neural network. (b) Part segmentation can be more accurate by using proper object and part scales. In the top row, we show our estimated object and part scales. In the bottom row, our part segmentation results gradually become better by increasingly utilizing the estimated object and part scales.

methods [5–9] usually generate part segment proposals first and then use graphical models to select and assemble the part segment proposals into human instances. These methods are often time-consuming and only work well in constrained scenes, which pre-suppose known scale, fairly accurate localization, clear appearances, and/or relatively simple poses. Recently, dramatic progress has been made in semantic part segmentation due to the advent of fully convolutional neural networks (FCNs) [2] and the availability of object part annotations in large-scale datasets like PASCAL [33]. These FCN-based methods [10, 3, 4] usually compute pixel-wise part labels directly in a simple and fast way, and then optionally refine the labels by a graphical model such as the dense CRF [11]. Although these FCN-based methods work well generally, they still suffer from the following two problems when handling natural multi-person scenes. They can make mistakes (e.g. missing small instances, producing poor boundary details or incomplete part regions, etc.) on small scale or extra large scale human instances for short of an effective mechanism to adapt to the size of the object instance. They also tend to produce erroneous predictions when the human instance is in an unusual pose or the appearance cues are weak, due to lack of object-level shape prior to regularize the part segments.

Similar to the trend of semantic part segmentation, traditional pose estimation approaches adopt graphical models to combine spatial constraints with local observations of joints, based on low-level features, like color intensities, HOG [34], shape-context [35], and so on. Recent strategies rely on deep-learned joint detectors, and use a carefully designed graphical model to select and assemble joints into valid pose configurations. Traditional approaches suffer from limited feature representation power and can only work in simple datasets with small pose and scale variation. Recent deep-based approaches have much better invariance to pose/scale variation, but their localization of joints is still inaccurate (e.g. joints are sometimes outside the human body) and they still struggle in multi-person overlapping scenes.

In this paper, we present a scale-aware and pose-guided part segmentation framework that effectively improves human semantic part segmentation in natural images with large pose/scale variation. The framework mainly contains two novel models: (1) a hierarchical auto-zoom model (HAZN) that handles the large scale variation of human instances and human parts; (2) a joint prediction model that combines pose estimation and semantic part segmentation together, letting the two tasks benefit each other. Here we give a brief introduction to the two proposed models.

The hierarchical "auto-zoom" model (HAZN) model is a FCN-based model performing object/part scale estimation and part segmentation at the same time, adapting to the size of objects and parts. It's partially motivated by the proposal-free end-to-end detection strategies [29–32]. To get some intuition of this approach, observe in Fig. 2(a), that the scale and location of a target object, and of its corresponding parts, can be estimated accurately from the field-of-view (FOV) window by applying a deep neural network. Using estimated object and part scales, the semantic part segmentation results can become better and better, see Fig. 2(b). Please refer to Sec. 3 for a detailed explanation of the HAZN model.

The joint prediction model is also a FCN-based framework aiming to solve the pose estimation task and the semantic part segmentation task jointly, in which the estimated pose provides object-level shape prior to regularize part segments (e.g. helping part segments align with human instances over the details of arms and legs where appearance cues are missing) while the part-level segments constrain the variation of pose locations (e.g. helping pose joints locate within

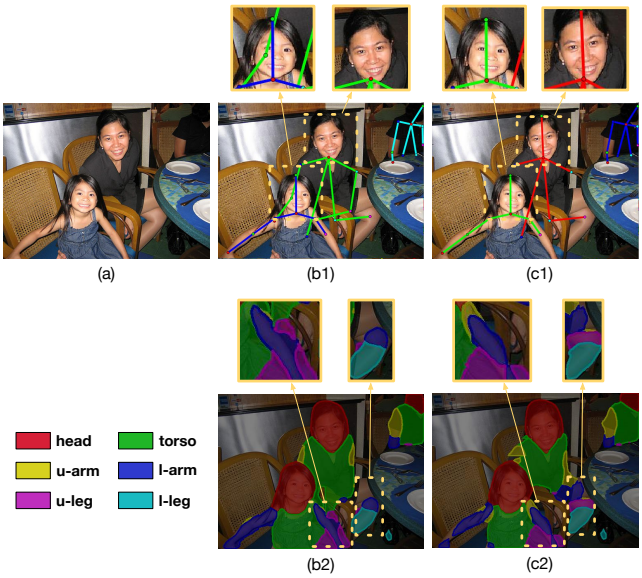

Fig. 3 Joint human pose estimation and semantic part segmentation improve both tasks. (a) input image. (b) pose estimation and semantic part segmentation results before joint inference. (c) pose estimation and semantic part segmentation results after joint inference. Note that comparing (b1) and (c1), our result recovers the missing forehead joint and corrects the location error of right elbow and right wrist for the woman on the right. Comparing (b2) and (c2), our result gives more accurate details of lower arms and upper legs than (b2) for both people.

their corresponding part segments). As shown in Fig. 3, by taking advantage of the complementary property between pose estimation and semantic part segmentation, this joint prediction model produces better results on both tasks. See Sec. 4 for detailed introduction of the model.

In this paper, we address the task of human semantic part segmentation in "the wild" where there are large variations in scale, pose, location, and occlusion. This motivates us to work with PASCAL images [36] because these were chosen for studying multiple visual tasks, do not suffer from dataset design bias [37], and include large variations of objects, particularly of scale and pose. Hence human semantic part segmentation in PASCAL is considerably more difficult than in datasets, such as Fashionista [7], that were constructed solely to evaluate human parsing.

We illustrate our approach by extensive experiments on PASCAL-Person-Part dataset [33], which contains detailed part annotations for every person in a subset of PASCAL VOC 2010. We further augment this dataset with 14 pose joint locations through manual labeling to evaluate our joint prediction model. Our experiments show that our approach outperforms previous state-of-the-art methods for both semantic part segmentation and pose estimation tasks. We are particularly good at recovering small human parts, giving clearer details of arms and legs, and correcting local confusions of people.

2 Related Works

As human semantic part segmentation and human pose estimation are two fundamental topics in computer vision, many graphical models on the two topics have been proposed during the past thirty years. With the advent of powerful deep learning techniques, some models start to use deep-learned features, or even perform the whole task within a deep neural network. Recently, there are also works that combine pose estimation and segmentation in graphical models. In this section, we give an overall review of previous literature in these aspects.

2.1 Human Semantic Part Segmentation

Traditional methods on human semantic part segmentation fall into two major categories. One type of methods first generate region/segment proposals for each semantic part, then select and assemble the region proposals by a graphical model. Bo et al. [5] generate region proposals by UCM segmentation [38], rank the region proposals using shape and appearance features, and assemble the proposals with simple geometric constraints. Dong et al. [8] get region proposals by UCM and CPMC [39], extract a rich set of appearance features for each region proposal encoded by Fisher Kernel and second-order pooling, and assemble the region proposals with a dedicated And-Or graph (AOG). The other type of methods treat semantic part segmentation as scene parsing, adopting pixel-wise conditional random fields (CRFs) to infer the pixel-wise part labels. Yamaguchi et al. [7] build a CRF based on super-pixels and try to classify each super-pixel into one of the semantic part types. They learn unary part classifiers for super-pixels based on traditional features, and use the output scores as unary terms in the CRF. For the pairwise terms, they train logistic regression models to predict whether two neighboring pixels should have the same label, and also learn a consistency prior between each part type pair from the training data. These traditional methods perform well on simple images, but struggle in natural images with large pose/scale variations.

Over the past few years, deep convolutional neural networks (DCNNs) [41] have pushed the performance to new heights in many computer vision tasks such as image classification and object detection. DCNNs have also been applied to semantic object segmentation in the wild [4, 14, 15], showing that DCNNs can also be applied to segment object parts in the wild. Some extract deep-learned features for region proposals, and assemble the region proposals using a graphical model based on those features [12]. Some use fully convolutional networks (FCNs) [2] to output pixel-wise part labels directly, tailed by an optional CRF to smooth the part labels [2, 4, 3]. Graphical models with DCNN features perform well on relatively simple datasets and give

good details of parts, but they struggle in complex natural datasets with large pose/scale variation. FCN-type approaches handle pose variation better, but they still suffer from large scale variation (e.g. missing parts or giving coarse part boundary details), and can produce local confusion errors (e.g. labeling arm regions as legs, labeling background regions as arms, etc.) if the person is in a non-typical pose, or when there are some other object/person nearby with similar appearance.

Recently, several works explore object-level context to produce better part segmentation results. Hariharan et al. [10] perform object detection, object segmentation and part segmentation sequentially, in which the object is first localized by a RCNN [40], then the object (in the form of a bounding box) is segmented by a FCN to produce an object mask, and finally part segmentation is performed by partitioning the mask. The process has two potential drawbacks: (1) it is complex to train all components of the model; (2) the error from object masks, e.g. local confusion and inaccurate edges, propagates to the part segments. Our hierarchical auto-zoom net model (HAZN) follows this general coarse-to-fine strategy, but is more unified and more importantly, we do not make premature decisions. Wang et al. [3] employ a two-stream FCN to jointly infer object and part segmentations for animals, where the part stream was performed to discover part-level details and the object stream was performed to find object-level context. Although this work proves the usefulness of object-level context, it only uses a single-scale network for both object and part score prediction, where small-scale objects might be missed at the beginning and the scale variation of parts still remains unsolved.

Some other recent works try to handle the scale issue within a DCNN structure. They commonly use multi-scale features from intermediate layers, and perform late fusion on them [2, 10, 4] in order to achieve scale invariance. Chen et al. [13] propose a scale attention model, which learns pixel-wise weights for merging the outputs from three fixed scales. These approaches, though developed on powerful DCNNs, are all limited by the number of scales they can select and the possibility that the scales they select may not cover a proper one. Our HAZN model avoids the scale selection error by directly regressing the bounding boxes for objects/parts and zooming the regions into proper scales. In addition, this mechanism allows us to explore a broader range of scales, contributing a lot to the discovery of missing object instances and the accuracy of part boundaries.

2.2 Human Pose Estimation

2.3 Joint Pose Estimation and Part Segmentation

3 Hierarchical Auto-Zoom Net

4 Joint Prediction of Pose Estimation and Semantic Part Segmentation

5 Conclusion

Acknowledgements AAA

References

1. Rauschert, Ingmar, and Robert T. Collins. "A generative model for simultaneous estimation of human body shape and pixel-level segmentation." In European Conference on Computer Vision, pp. 704-717 (2012)
2. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440 (2015)
3. Wang, Peng, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L. Yuille. "Joint object and part segmentation using deep learned potentials." In Proceedings of the IEEE International Conference on Computer Vision, pp. 1573-1581 (2015)
4. Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." arXiv preprint arXiv:1606.00915 (2016)
5. Bo, Yihang, and Charless C. Fowlkes. "Shape-based pedestrian parsing." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2265-2272 (2011)
6. Eslami, S., and Christopher Williams. "A generative model for parts-based object segmentation." In Advances in Neural Information Processing Systems (NIPS), pp. 100-107 (2012)
7. Yamaguchi, Kota, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. "Parsing clothing in fashion photographs." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3570-3577 (2012)
8. Dong, Jian, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. "A deformable mixture parsing model with parselets." In International Conference on Computer Vision (ICCV), pp. 3408-3415 (2013)
9. Zhu, Long Leo, Yuanhao Chen, Chenxi Lin, and Alan Yuille. "Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation." In International Journal of Computer Vision, 93(1), pp 1-21 (2011)
10. Hariharan, Bharath, Pablo Arbelez, Ross Girshick, and Jitendra Malik. "Hypercolumns for object segmentation and fine-grained localization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 447-456 (2015)
11. Krhenbhl, Philipp, and Vladlen Koltun. "Efficient inference in fully connected crfs with gaussian edge potentials." In Advances in Neural Information Processing Systems (NIPS), pp. 109-117 (2011)
12. Xia, Fangting, Jun Zhu, Peng Wang, and Alan L. Yuille. "Pose-Guided Human Parsing by an AND/OR Graph Using Pose-Context Features." In AAAI, pp. 3632-3640 (2016)
13. Chen, Liang-Chieh, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. "Attention to scale: Scale-aware semantic image segmentation." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 3640-3649 (2016)

14. Dai, Jifeng, Kaiming He, and Jian Sun. "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1635-1643 (2015)
15. Papandreou, George, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. "Weakly-and semi-supervised learning of a DCNN for semantic image segmentation." arXiv preprint arXiv:1502.02734 (2015)
16. Yang, Yi, and Deva Ramanan. "Articulated pose estimation with flexible mixtures-of-parts." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1385-1392 (2011).
17. Tompson, Jonathan, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. "Efficient object localization using convolutional networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 648-656 (2015)
18. Chen, Xianjie, and Alan Yuille. "Parsing occluded people by flexible compositions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3945-3954 (2015)
19. Branson, Steve, Grant Van Horn, Serge Belongie, and Pietro Perona. "Bird species categorization using pose normalized deep convolutional nets." arXiv preprint arXiv:1406.2952 (2014)
20. Zhang, Ning, Jeff Donahue, Ross Girshick, and Trevor Darrell. "Part-based R-CNNs for fine-grained category detection." In European Conference on Computer Vision, pp. 834-849 (2014)
21. Krause, Jonathan, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. "The unreasonable effectiveness of noisy data for fine-grained recognition." In European Conference on Computer Vision, pp. 301-320 (2016)
22. Wang, Yang, Duan Tran, Zicheng Liao, and David Forsyth. "Discriminative hierarchical part-based models for human parsing and action recognition." In Journal of Machine Learning Research, 13(Oct), pp. 3075-3102 (2012)
23. Zhou, Yang, Bingbing Ni, Richang Hong, Meng Wang, and Qi Tian. "Interaction part mining: A mid-level approach for fine-grained action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3323-3331 (2015)
24. Wang, Chunyu, Yizhou Wang, and Alan L. Yuille. "An approach to pose-based action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 915-922 (2013)
25. Ma, Lianyang, Xiaokang Yang, Yi Xu, and Jun Zhu. "Human identification using body prior and generalized EMD." In Image Processing (ICIP) 18th IEEE International Conference, pp. 1441-1444 (2011)
26. Zhao, Liming, Xi Li, Jingdong Wang, and Yueting Zhuang. "Deeply-learned part-aligned representations for person re-identification." arXiv preprint arXiv:1707.07256 (2017)
27. Gallego, Jaime, Montse Pardas, and Jose-Luis Landabaso. "Segmentation and tracking of static and moving objects in video surveillance scenarios." In Image Processing (ICIP) 15th IEEE International Conference, pp. 2716-2719 (2008)
28. Liu, Si, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. "Surveillance video parsing with single frame supervision." In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops, pp. 1-9 (2017)
29. Huang, Lichao, Yi Yang, Yafeng Deng, and Yinan Yu. "Densebox: Unifying landmark localization with end to end object detection." arXiv preprint arXiv:1509.04874 (2015)
30. Liang, Xiaodan, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan. "Proposal-free network for instance-level object segmentation." arXiv preprint arXiv:1509.02636 (2015)
31. Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in Neural Information Processing Systems (NIPS), pp. 91-99 (2015)
32. Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788 (2016)
33. Chen, Xianjie, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. "Detect what you can: Detecting and representing objects using holistic models and body parts." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1971-1978 (2014)
34. Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893 (2005)
35. Belongie, Serge, Jitendra Malik, and Jan Puzicha. "Shape context: A new descriptor for shape matching and object recognition." In Advances in Neural Information Processing Systems (NIPS), pp. 831-837 (2001)
36. Everingham, Mark, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes challenge: A retrospective." In International Journal of Computer Vision 111, no. 1, pp 98-136 (2015)
37. Li, Yin, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. "The secrets of salient object segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 280287 (2014)
38. Arbelaez, Pablo, Michael Maire, Charless Fowlkes, and Jitendra Malik. "From contours to regions: An empirical evaluation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2294-2301 (2009)
39. Carreira, Joao, and Cristian Sminchisescu. "Cpmc: Automatic object segmentation using constrained parametric min-cuts." In IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7), pp 1312-1328 (2012)
40. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587 (2014)
41. LeCun, Yann, Lon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." In Proceedings of the IEEE, 86(11), pp. 2278-2324 (1998)