HIERARCHICAL PART DETECTION WITH DEEP NEURAL NETWORKS

Esteve Cervantes[⋆†], *Long Long Yu*[†], *Andrew D. Bagdanov*[⋆], *Marc Masana*[⋆], *Joost van de Weijer*[⋆]

[†]Wide Eyes Technologies, Barcelona, Spain
[⋆] Computer Vision Center Barcelona, Universitat Autonoma de Barcelona, Spain

ABSTRACT

Part detection is an important aspect of object recognition. Most approaches apply object proposals to generate hundreds of possible part bounding box candidates which are then evaluated by part classifiers. Recently several methods have investigated directly regressing to a limited set of bounding boxes from deep neural network representation. However, for object parts such methods may be unfeasible due to their relatively small size with respect to the image. We propose a hierarchical method for object and part detection. In a single network we first detect the object and then regress to part location proposals based only on the feature representation inside the object. Experiments show that our hierarchical approach outperforms a network which directly regresses the part locations. We also show that our approach obtains part detection accuracy comparable or better than state-of-the-art on the CUB-200 bird and Fashionista clothing item datasets with only a fraction of the number of part proposals.

***Index Terms*—** Object Recognition, Part Detection, Convolutional Neural Networks

1. INTRODUCTION

Parts are believed to be an essential part of object category models [1, 2]. Methods vary in the way they model spatial relations between parts, the nature of the parts (semantic or unsupervised), and the number of parts. Apart from their use for object detection [2], parts have been applied in action recognition [3] and fine grained detection [4, 5].

Approaches based on sliding windows have long dominated the field of object recognition. The ability to implement these methods as a convolutional filter allows them to quickly evaluate many windows, however the number of windows to consider is vast. As a solution, object proposal methods were developed [6, 7] which use bottom-up image analysis to propose a limited set of object regions. The success of object proposals has sparked its application for part-based object detection [5, 8]. In [5] the selective search object proposal method was used to generate part proposals for bird recognition. However, part detection is of different than object detection. In object detection, prior knowledge of the expected location and size of objects is limited and the generation of thousands of object proposals based on low-level image evidence is reasonable. However, parts have in general more restricted statistics especially when we consider their position with respect to the object location and size. Exploiting these restrictions on the expected position and size of the part proposals is the main objective of this paper.

Alternatives using regression to directly estimate object proposals from CNN representations have been proposed [9]. This technique proposed for object detection is class agnostic and still requires hundreds of proposals per image. Regressing directly to parts was studied by Liang et al. [10], who directly estimate bounding boxes of clothing items given a person bounding box. Their method has the advantage that only a single bounding box per clothing item class needs to be evaluated. However, their method separates the object detection (in their case the human) from the part detection.

In this paper we propose an end-to-end hierarchical object and part detection framework. Given a CNN representation of an image our method regresses a *single* object bounding box. Next, based on the CNN representation within the object bounding box we regress a *single* proposal for each of the parts. We train the hierarchical object and part detection network in an end-to-end fashion. To the best of our knowledge, we are the first to investigate such a hierarchical network for part detection. Our method has the advantage over object proposal methods [6, 7, 5, 8, 9] that we evaluate significantly fewer bounding boxes. With respect to [10], our work integrates object and part detection in a single network.

2. TOP-DOWN PART REGRESSION

In the recognition problems we consider in this paper, objects we wish to localize consist of an ensemble of sub-objects, or object parts. In the fashion recognition problems we consider in Section 3, for example, we localize clothing items (e.g. hat, glasses, boots, skirt, and handbag) present in images. These problems are often characterized by having a relatively large number of potential parts, some or most of which may not be present in a given image.

This work was supported by TIN2014-52072-P and TIN2013-42795-P of the Spanish Ministry of Science. We also thank the NVIDIA Corporation for support in the form of donated GPUs.

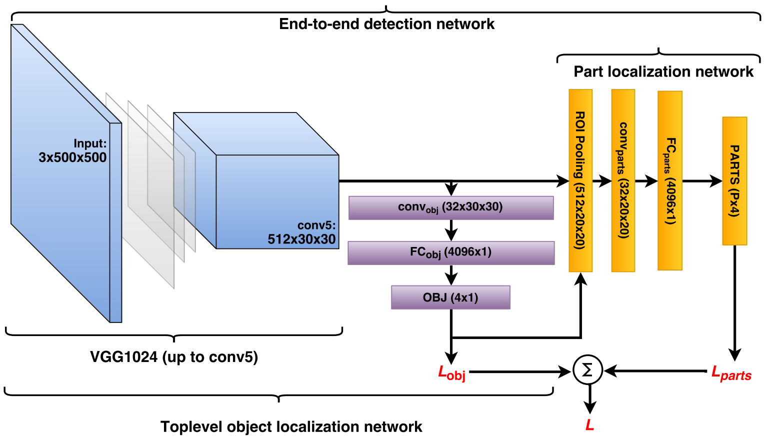

Fig. 1. A CNN for simultaneously detecting object and object-parts. See text for detailed explanation.

2.1. Object and object-part detection by CNN

Our approach is an adaptation of some of the key ideas from proposal-based object recognition methods [11, 12] which are based on the evaluation of many object proposals. Some of them apply regression to improve the proposed bounding box, which typically leads to a small shift of the bounding box. In [9] they directly regress to potential object bounding boxes from the images, but they do so without taking classes into account, and [10] regress to parts, however segmentation is their final goal.

The main novelty of our method is that we propose a hierarchical approach to object and part detection. Our aim is to prevent having to evaluate the thousands of part proposals which are typically proposed by object proposal methods [5, 8]. We first regress to the object bounding box based on the image and next regress to the part bounding boxes based only on the features within the object bounding box.

Given an image I, we assume there is one top-level object (e.g. "person" for clothing item detection). This top-level object is then assumed to consist of a subset of possible parts (e.g. clothing items like "hat", "skirt", or "boots"). Our approach is to learn how to predict a *single* candidate top-level object box, and then a *single* candidate object box for each of the P potential object parts:

$$\begin{aligned}\hat{B}_{\text{obj}}(I) &= [x, y, w, h]\\ \hat{B}_{\text{parts}}(I, B_{\text{obj}}) &= \{\hat{\mathbf{b}}_p \mid p = 1, \ldots P\},\end{aligned}$$

where each candidate part box $\hat{\mathbf{b}}_p = [x_p, y_p, w_p, h_p]$ is dependent on both the image I and the top-level object box $B_{\text{obj}}(I)$. Note also that the total number of boxes which are generated by our network is $P + 1$.

The object and object-part boxes can be estimated using regression based on the internal representation of images in intermediate layers of Convolutional Neural Networks (CNNs) [9, 12]. Instead of directly estimating the parts from the image, we model them to be dependent on the estimation of the top-level object. Therefore, it is necessary to regress the object and parts in two stages. This will allow us to train the network end-to-end in order to simultaneously estimate an object and its parts.

2.2. An end-to-end network for object part recognition

The network we use for object and object-part detection is illustrated in Fig. 1. The network is designed to simultaneously estimate $\hat{B}_{\text{obj}}(I)$ and $\hat{B}_{\text{parts}}(I, B_{\text{obj}})$. We first discuss each component in the architecture of our network, then describe the loss function that is optimized for training.

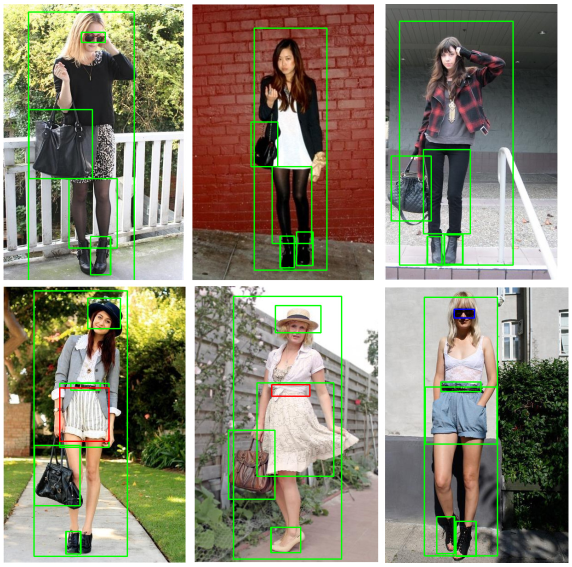

Fig. 3. Results on Fashionista. (top row) Images with correctly detected clothing items (green). (Bottom row) Images with some items wrongly detected (red) or missed (blue).

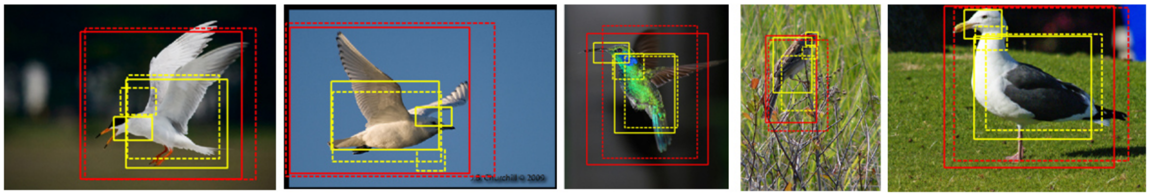

Fig. 2. Results of the non-hierarchical (in dotted lines) and the hierarchical network (in solid lines). Bird detection is given in red and part ('body' and 'head') detection in yellow. Note the improvement in localization due to hierarchical network.

Network architecture. Our network is based on the VGG1024 network described in [13]. Since our architecture cascades in sequence two detection phases, it is important to balance performance with network size and complexity. The main components of our network are:

- The initial part of the network (**Input** through **conv5**) is identical to the VGG1024 network. The pretrained weights from VGG1024 are used to initialize the first part of the network, and fine-tuning is performed from there.
- The *object prediction sub-network* (in violet in Fig. 1) which proposes a *single* object box in the image and consists of: conv_{obj}, a 1×1, fully convolutional layer which reduces the dimensionality of the incoming feature map to $32 \times 30 \times 30$; FC_{obj}, a 4096-unit, fully-connected layer which provides the high-level feature representation from which the object box position will be predicted; and **OBJ**, a fully connected layer which predicts the four bounding box coordinates of the top-level object.
- The *part proposal* subnetwork (in orange in Figure 1 which, given the top-level object prediction, predicts the positions of the P parts and consists of: **ROI Pooling**, a pooling layer that extracts and pools the **conv5** features from the predicted toplevel bounding box location; $\text{conv}_{\text{parts}}$, which reduces the dimensionality of the incoming feature map to $32 \times 20 \times 20$; FC_{parts}, a 4096-unit, fully-connected layer which provides the high-level feature representation from which the part box position will be predicted; and **PARTS**, a fully connected layer which predicts the four bounding box coordinates for each of the P potential object parts.

The network produces two loss outputs: the *object loss* L_{obj}, which determines how well the toplevel object is estimated, and the *part loss* which is an average of how well each object sub-part is predicted.

Learning. We train the network in Fig. 1 end-to-end to minimize an average per-box loss over all training images. For a training image I denote the groundtruth toplevel object box as $B^*_{\text{obj}} = [x^*, y^*, w^*, h^*]$, and similarly the set of groundtruth sub-part boxes as $B^*_{\text{parts}} = \{\mathbf{b}^*_p \mid p = 1, \dots, P\}$. We also denote by $y^*_p \in \{0, 1\}$ an indicator of whether part p is present in the training image.

	# Boxes	Head	Body	Bbox
NH	3	0.39	0.83	0.95
H	3	0.63	0.91	0.96
Fast-RCNN & EB	50	0.36	0.82	0.97
Fast-RCNN & EB	1000	**0.82**	0.90	**0.99**
Fast-RCNN & H	3	0.81	**0.92**	0.98
Zhang et al. [5]	>1000	0.61	0.70	-

Table 1. Results on the CUB-200 dataset in percentage of correctly localized parts. We report results for both Hierarchical (indicated by 'H') and non hierarchical (indicated by 'NH') predictions. Fast-RCNN results are combined with Edge Box (EB) and our hierarchical (H) method.

The loss we optimize is the sum of the toplevel object loss, and the constituent part losses:

$$L(\hat{B}_{\text{obj}}, \hat{B}_{\text{parts}}) = L_{\text{obj}}(\hat{B}_{\text{obj}}) + L_{\text{parts}}(\hat{B}_{\text{parts}}), \tag{1}$$

where L_{obj} measures localization error for the toplevel object:

$$L_{\text{obj}} = \text{smooth}_{L_1}(\hat{B}_{\text{obj}} - B^*_{\text{obj}}),$$

L_{parts} measures the average sub-part localization error:

$$L_{\text{parts}} = \frac{1}{\sum_p y^*_p} \sum_p y^*_p \text{smooth}_{L_1}(\mathbf{b}^*_p - \hat{\mathbf{b}}_p),$$

and $\text{smooth}_{L_1}(\cdot)$ is the smooth ℓ_1 loss function:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{otherwise.} \end{cases}$$

Optimizing Eq. 1 simultaneously learns multiple regressors: one for the toplevel object and one for each part.

3. EXPERIMENTAL RESULTS

In this section we report on a number of experiments to evaluate our part detection method.

Datasets: In the experiments we consider three datasets. The CUB-200 dataset [15] consists of 11,788 images of 200 bird species. Each image is annotated with bounding box,

	#Boxes	Bag	Belt	Glasses	Hat	Pants	L-Shoe	R-Shoe	Shorts	Skirt	Tights	Mean	Bbox
EB	50	0.37	0.09	0.11	0.50	0.65	0.14	0.15	0.34	0.50	0.51	0.34	N.A.
EB	1000	**0.72**	**0.28**	**0.60**	**0.81**	0.82	0.35	0.38	**0.77**	0.68	0.75	0.62	N.A.
NH	11	0.46	0.13	0.32	0.63	0.80	0.31	0.37	0.60	0.67	0.72	0.50	N.A.
H	11	0.51	0.22	0.35	0.70	**0.97**	**0.61**	**0.62**	0.73	**0.92**	**0.92**	**0.65**	0.99
[14]	>1000	0.23	0.14	**0.22**	0.36	0.57	0.29	**0.33**	0.37	0.29	0.41	0.31	N.A.
H	11	**0.38**	**0.16**	0.16	**0.56**	**0.85**	**0.30**	0.23	**0.54**	**0.56**	**0.60**	**0.45**	1.0

Table 2. Results in AP on the Wide Eyes fashion dataset (top four lines) and the Fashionista dataset (bottom 2 rows). We report results for both Hierarchical (indicated by 'H') and non hierarchical (indicated by 'NH') predictions, and edge Box (EB). The final column shows that the hierarchical approach obtains almost perfect person bounding box detection.

part location (head and body), and attribute labels. The Fashionista data set consists of 685 images containing full body views of persons and covering a variety of clothing items [16]. Here we consider clothes as parts of the object class person. Images are annotated with groundtruth clothing labels for 53 categories. This is an example of a part-based recognition problem where images only contain a small subset of the possible parts. The Wide Eyes fashion image dataset consists of 19,000 images (18,000 for training and 1,000 for testing) of full body views of persons. The images are similar to the Fashionista dataset.

Part detection on the CUB-200 bird dataset: In this section we report results on the CUB-200 dataset. Most previous work considers that the bounding box of the bird is provided [4, 17, 18]. Here we consider the more realistic scenario where both the bird as well as the parts should be detected in the image; this scenario was also considered by Zhang et al. [5]. The results are presented in Table 1 in terms of the percentage of correctly classified bounding boxes. A bounding box is considered correctly classified when they have intersection over union (IoU) overlap [19] with the groundtruth bounding box of at least 0.5.

First we consider the main novelty which is the introduction of hierarchical detection of object and parts. A non-hierarchical approach would directly estimate the part bounding boxes from the image. The results in Table 1 clearly show the superiority of the hierarchical approach. It obtains significant gains for both parts; especially noteworthy is the 20% gain for heads. Next, we train a Fast-RCNN with 1000 bounding boxes [11], and test this network with 50, and 1000 object EdgeBox proposals, and the three part proposals from our method. The improvement in the results of our method combined with Fast-RCNN is due to the extra regression step which improves results especially for 'head' parts. Fast-RCNN with 1000 bounding boxes obtains similar results as our method with only three proposals. Finally, we compare to the results of Zhang et al. [5]. We outperform their results considerably, and we do so by only considering a fraction of bounding boxes they do.

Clothing item recognition: Next we evaluate our part proposal method for the task of clothing item detection. We select the same ten classes considered in [14] to compare results. Here we first detect the person and based on the features within the person bounding box, we propose ten boxes for the various clothing parts. We compare our approach against the object proposal method EdgeBox [20] with a varying number of proposals. We train a Fast-RCNN [11] object detector based on 1000 EdgeBox proposals per training image. At testing time we test the same Fast-RCNN network providing it with either our proposed eleven boxes, or a varying number of boxes proposed by EdgeBox. The results are summarized in Table 2 (first four rows). Again results show that our hierarchical approach significantly improves results. Note that our results obtained with *only eleven part proposals* obtain similar results as EdgeBox with 1000 proposals per image.

Finally, we compare results to [14]. That method is based on over 1000 proposals [7] and a deep neural network feature representation, and incorporates pose prediction and geometrical priors. To obtain results on Fashionista we finetune our network trained on the WideEyes dataset on the Fashionista dataset [16].[1] Next we provide our part proposals to a Fast-RCNN network which was only trained on the Fashionista train dataset. We report results in Table 2 (bottom two rows). The results show that we significantly outperform the reported results of [14], although we only require a fraction of the number of part proposals. In Fig. 3 several examples of the detections with our network are provided.

4. CONCLUSIONS

In a single end-to-end network we propose a hierarchical approach to object and part detection. Experiments show that the proposed network outperforms its non-hierarchical counterpart, and obtains similar or better results than state-of-the-art on two benchmark datasets while only considering a fraction of the number of part proposals.

Acknowledgements: This work has been funded by the project TIN2013-41751, the CHISTERA project PCIN-2015-251, and the NEOTEC fund IDI-20140496 of the Spanish Ministry of Science.

[1] The Fashionista dataset was too small for training our network.

5. REFERENCES

[1] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2003.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] Fahad Shahbaz, Joost van de Weijer, Muhammad Anwer Rao, Michael Felsberg, and Carlo Gatta, "Semantic pyramids for gender and action recognition," *IEEE TIP*, pp. 1–1, 2014.

[4] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[5] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[6] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010.

[7] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[8] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r*cnn," in *IEEE International Conference and Computer Vision (ICCV)*. IEEE, 2015.

[9] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov, "Scalable object detection using deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2155–2162.

[10] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Liang Lin, and Shuicheng Yan, "Deep human parsing with active template regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[11] Ross Girshick, "Fast r-cnn," in *International Conference on Computer Vision (ICCV)*, 2015.

[12] R. Girshick, J. Donahue, T. Darrellr, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 580–587.

[13] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[14] Kota Hara, Vignesh Jagadeesh, and Robinson Piramuthu, "Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 2016.

[15] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[16] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3570–3577.

[17] Fahad S Khan, Joost van de Weijer, Andrew D Bagdanov, and Maria Vanrell, "Portmanteau vocabularies for multi-cue image representation," in *Advances in neural information processing systems*, 2011, pp. 1323–1331.

[18] Ning Zhang, Ronan Farrell, Forrest Iandola, and Trevor Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 729–736.

[19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*. Springer, 2014.