DEEP COUNTING MODEL EXTENSIONS WITH SEGMENTATION FOR PERSON DETECTION

Sanjukta Ghosh[⋆†] *Peter Amon*[†] *Andreas Hutter*[†] *André Kaup*[⋆]

[⋆]Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[†]Image Processing and Visualization, Siemens Corporate Technology, Munich, Germany

ABSTRACT

Applications like autonomous driving, surveillance, or any application that demands scene analysis requires object detection, semantic segmentation and instance segmentation. In this paper, we focus on the problem of detecting each instance of a specific category of objects, specifically persons. A novel method for object detection is proposed based on a deep counting model. The feature extractor of the deep counting model is extended with additional layers for segmenting specific instances. While the feature extractor of the deep counting model already focuses on the persons in the scene, the segmentation layers help to get a more accurate estimation of the foreground with persons and the instance segmentation is able to estimate separate instances of persons. Our proposed method outperforms other methods on the CUHK08 dataset with an Average Miss Rate (AMR) of 14% and on the PETS09 dataset with an AMR of 41%.

Index Terms— Deep convolutional neural networks, deep counting models, detection, segmentation, multi-task network

1. INTRODUCTION

Deep learning [1] based methods for image and video analytics are extremely successful and popular. Typical applications include classification, detection, and semantic segmentation. While the performance of a lot of algorithms has been demonstrated on datasets, in practical applications the scenarios can be challenging. For example, in a visual surveillance scenario of people, there can be multiple persons walking close together and/or partially occluding one another when the scene is captured. This poses an additional requirement for detection or segmentation algorithms to be able to separate out the instances of the persons. Our paper focuses on detecting each instance of a specific category of objects, in this case persons, by making use of a deep counting model (Fig. 1). The main contributions of this paper are (1) proposing a multi-task deep convolutional neural network (CNN) based architecture for person counting, segmentation and instance separator followed by instance segmentation which in turn is used for person detection and (2) exploiting the synergies of the feature extractor of the deep counting model to optimize the architecture that would otherwise comprise of a full-fledged decoder comprising of deconvolution and unpooling layers.

The research leading to these results has received funding from the German Federal Ministry for Economic Affairs and Energy under the VIRTUOSE-DE project.

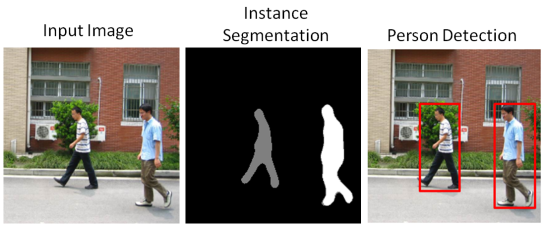


Fig. 1: Object detection obtained from instance segmentation for an image from the PennFudan Pedestrian dataset.

2. RELATED WORK

While numerous deep learning based methods have been developed for classification [2] , detection ([3], [4], [5]) and semantic segmentation ([6], [7]), there are relatively few deep learning based approaches for instance segmentation. Bai et al. [8] achieve instance segmentation by learning the energy function of the classical Watershed Transform. The Mask R-CNN [9] based approach achieves instance segmentation by predicting a segmentation mask for each detected object in an image. Brabandere at al. [10] achieve instance segmentation by using a metric learning based approach at the pixel level. A cost function comprising of distance and variance terms are used to force the model to learn to embed pixels of the same instance close to each other and the pixels of different instances away from each other. However, it has a drawback that it is not able to perform well in datasets with large diversity in images like PASCAL VOC. A set of techniques for instance segmentation use recurrent approaches ([11], [12]) to sequentially predict each object instance.

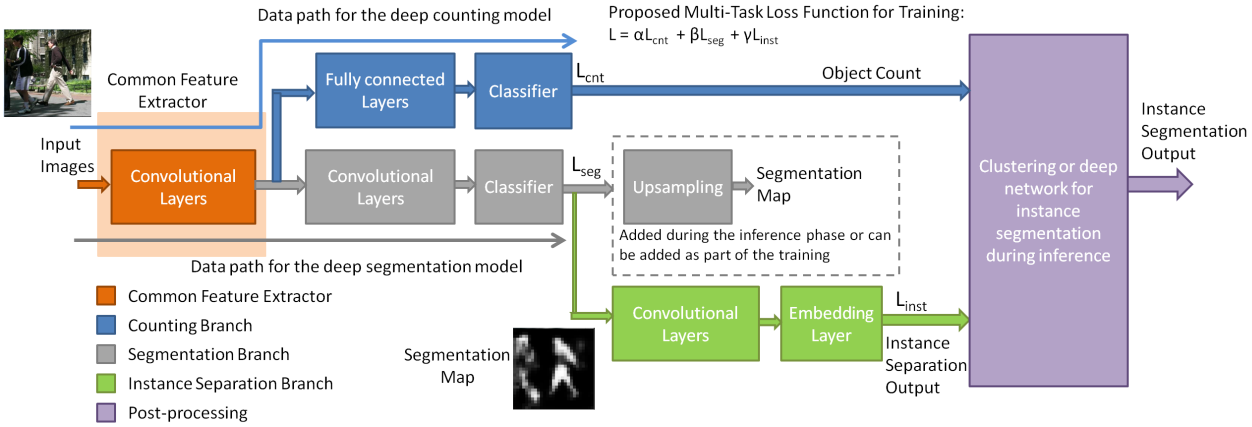


Fig. 2: Architecture of multi-task network for object counting, segmentation and instance separation.

3. PROPOSED METHOD

A framework for performing multiple image analysis tasks like person counting, person segmentation, and instance segmentation/detection is proposed. Our method is based on a deep counting model with extensions. The deep counting model is found to have features which can be exploited to obtain an architecture comprising of fewer layers as compared to any other multi-task network that achieves segmentation/detection in combination with other analytics tasks. Another advantage of the framework is that it can be used to obtain the output either as segmentation maps or bounding boxes or semantic boundaries with simple post-processing steps.

3.1. Deep Counting Model Based Architecture

While the task of counting can be addressed by using regression to predict a count or estimating a density map and obtaining the count by integrating over the area, it has been found that casting the counting problem as a classification problem by using CNNs results in the feature extractor learning features useful for analytics applications as demonstrated in [13] and [14]. We trained a deep counting model as described in [15]. As can be seen in Fig. 3, the feature extractor of a CNN trained for counting persons is able to focus on persons in the scene for different scenarios.

We extend the deep counting model with additional branches for segmentation and instance separation. The architecture shown in Fig. 2 comprises of a common feature extractor followed by three heads, a head or specific layers for object counting, a second one for segmentation, and a third one for instance separation. Since the feature extractor of the deep counting model is able to focus on foreground regions, it is possible to add extensions of only a few layers to achieve the segmentation and instance separation instead of a full-fledged decoder comprising of a series of deconvolution

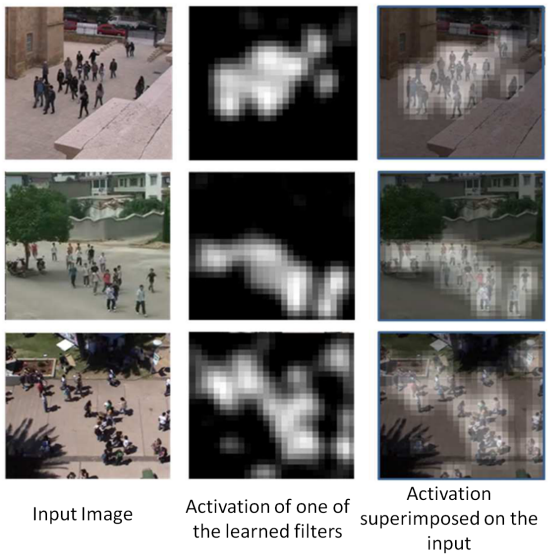


Fig. 3: Focus on foreground with persons using a deep counting model.

(or transposed convolution) and unpooling layers. This helps in reducing the number of computations especially during inference. The probabilistic segmentation map drives the head for instance separation. The function of the instance separator is to create an embedding in a feature space such that the pixel representations corresponding to the same instance are close together and the pixel representations corresponding to different instances are separated. It is not possible to use a simple classifier or cross entropy loss to achieve instance segmentation directly since the number of instances in each frame can vary and there are no clear discriminating features between different instances of the same category of objects.

Moreover, the ordering of the labels of the instances is not important.

The task of instance segmentation is broken down into background removal, foreground localization, and instance separation followed by post-processing steps for instance segmentation. Instead of using separate networks for each of these sub-tasks, using a combined network that comprises of a common feature extractor is beneficial. The different components of the network are synergistic and assist each other during the training resulting in quicker convergence.

3.2. Loss Function

The loss function comprises of a combination for counting, segmentation and instance separation. The loss function for training the combined model is

$$L = \alpha L_{\text{cnt}} + \beta L_{\text{seg}} + \gamma L_{\text{inst}} \quad (1)$$

where α, β, and γ are the weights for the counting branch, segmentation branch, and instance separation branch respectively. L_{cnt} is the cross-entropy loss for the counting branch, L_{seg} is the cross-entropy loss for the segmentation branch and L_{inst} is the loss function for creating pixel-wise embeddings of the instances. The part of the cost function for counting can be a cross entropy loss with classes corresponding to different counts of objects. The segmentation loss function comprises of a cross-entropy loss for two labels: background and foreground. The part of the cost function used for separation of the embedding of the instances is based on the cost function described in [10]. This part of the cost function tries to ensure that pixel embeddings of the same instance are within δ_{v} distance of the cluster center for that instance while the cluster centers for different instances are at least $2\delta_{\text{d}}$ apart. Moreover, it is observed that the feature extractor of the model in the case when the counting model is combined with the instance separation is able to focus better on the foreground objects than in the case where only a pixel level embedding for instance separation is used. Examples on two input images are shown in Fig. 4. While the proposed method has been implemented for the category persons, it can be applied to other categories of objects as well.

3.3. Post-processing

During inference, the trained model described above is used to obtain the prediction of the count of instances and the instance separated output embedded in an n-dimensional space. To obtain the instance segmentation, clustering techniques like mean shift algorithm or K-means can be used in which the predicted count can be used to initialize the number of clusters. Moreover, a deep network with few layers that has been trained for obtaining instance segmentation from the embedded space can be used. The predicted object count will be used to select the appropriate network head. In our case, the output is desired as separate bounding boxes around every object instance. So each object instance as obtained from the instance segmentation is enclosed in a bounding box.

Fig. 4: Effect of combining the counting model with instance separation on the common feature extractor.

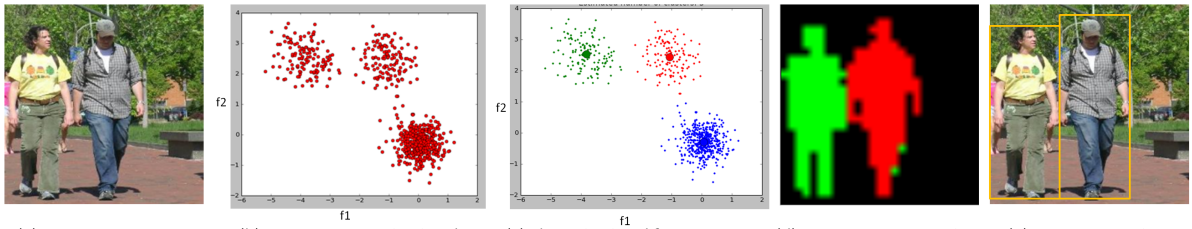
4. EXPERIMENTS

4.1. Dataset and Metrics

Qualitative tests are carried out on the PennFudan Pedestrian dataset [16]. Quantitative tests are performed on the CUHK08 [17] and PETS09 [18] datasets. The datasets have a higher number of pedestrians and occlusion than in typical pedestrian detection datasets. The metric used for measuring the performance of the algorithm is the Average Miss Rate (AMR). It is the area under the ROC curve that plots the Miss Rate vs. the False Positives Per Image (FPPI). The conditions for calculating the AMR include the standard conditions of integrating within 10^{-2} to 10^{0} FPPI. For the evaluation, the toolbox from [19] is used. The common feature extractor is based on the AlexNet [2] architecture followed by extensions for each branch.

4.2. PennFudan Pedestrian dataset

The PennFudan Pedestrian dataset has multiple pedestrians with the heights of labeled pedestrians in a range of 180-390 pixels. Moreover, the backgrounds of the images are not fixed. A model is trained with our proposed method on synthetic data as proposed in [15] and a subset of the PennFudan dataset. Fig. 5(a) shows a test image from this dataset with 2 pedestrians. For the purpose of visualization, the instance separator embeds in a 2d feature space. Three distinct clusters can be observed in Fig. 5(b). Fig. 5(c) is obtained after clustering using the mean shift algorithm[22]. Three clusters are identified corresponding to the background and the 2 pedestrians. Since a pixel level embedding is done in the feature space, the spatial correspondence can be established and the



(a) Input Image (b) Instance separation in 2d feature space (c) Clustering in 2d feature space (d) Instance Segmentation Output (e) Person Detection Output

Fig. 5: Instance separation and detection in 2d feature space.

Table I: AMR on CUHK08 and PETS09 datasets.

Method	Average Miss Rate(%)					
	CUHK08	PETS09				
		S1.L1	S1.L2	S2.L2	S2.L3	Avg.
DPM [20]	47	60	76	60	56	63
Deep Counting Model [13]	26	46	62	48	50	52
DPM + Scale Pior [21]	39	54	68	56	50	57
DPM + Scale Prior + Occlusion Analysis [21]	33	50	63	54	49	54
DPM + Deep Model Specialization [14]	29	45	59	50	47	50
Deep Counting Model + Deep Model Specialization [14]	18	37	56	40	44	44
Our proposed method	14	32	54	38	41	41

instance segmentation obtained as in Fig. 5(d). Fig. 5(e) shows the result of detection after bounding boxes are placed around each identified instance.

4.3. CUHK08 dataset

Table I shows the AMR values for our method in comparison with other methods for the CUHK08 dataset. The benefit of combining the deep counting model with a segmentation and instance segmentation branch is evident as our proposed method achieves the lowest AMR of 14%. While the counting model aids in eliminating the background, any remaining parts of the background are also eliminated by the explicit background classification of the segmentation branch and embedding of the background to a cluster in the feature space away from the instances of the persons. The instance separator then works to effectively separate the instances of persons.

4.4. PETS09 dataset

The PETS09 dataset [18] has sequences with multiple pedestrians and scenarios with occlusion. Table I shows the AMR values for our method in comparison with other methods. The S1.L1 and S2.L2 walking sequence have medium density of people while the S1.L2 and S2.L3 walking sequence have a high density of people. Our proposed method outperforms the other methods. As compared to the other deep counting model based approaches in Table I, our method uses additional supervision during training and at the same time exploits the strengths of the deep counting model resulting in an improved performance. The situation is especially challenging for the sequences with high density of people since there are complex occlusion patterns or people very close to each other. In such cases, our proposed approach helps to separate out the instances.

5. CONCLUSION

We propose a method for detecting instances of a specific category of objects like persons by extending a CNN-based deep counting model by a segmentation branch and instance separator branch. The method facilitates the simultaneous analysis of an image for predicting the count of the objects present in it, the segmentation mask for these objects and a delineation for each object separately by way of an instance segmentation mask or bounding box or semantic boundary. The tasks of our multi-task network assist each other during the training, thus helping to obtain a network with reduced number of layers than the state-of-the-art methods where a series of deconvolution and unpooling layers are involved. This results in reduced computations during inference. Experiments on datasets with multiple persons of varying numbers and occlusion patterns have shown that the proposed method is effective in challenging scenarios.

6. REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2015, NIPS'15, pp. 91–99, MIT Press.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[5] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[8] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *CVPR*. 2017, pp. 2858–2866, IEEE Computer Society.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. International Conference on Computer Vision (ICCV), 2017*, pp. 2980–2988.

[10] B. Brabandere, D. Neven, and L. Gool, "Semantic instance segmentation with a discriminative loss function," *CoRR*, vol. abs/1708.02551, 2017.

[11] B. Romera-Paredes and P. H. Sean Torr, "Recurrent instance segmentation," in *Proc. European Conference on Computer Vision, Part VI*, October 2016, pp. 312–329.

[12] Eunbyung Park and Alexander C. Berg, "Learning to decompose for object detection and instance segmentation," *CoRR*, vol. abs/1511.06449, 2015.

[13] S. Ghosh, P. Amon, A. Hutter, and A. Kaup, "Reliable pedestrian detection using a deep neural network trained on pedestrian counts," in *Proc. IEEE International Conference on Image Processing (ICIP)*, September 2017, pp. 685–689.

[14] S. Ghosh, P. Amon, A. Hutter, and A. Kaup, "Detecting closely spaced and occluded pedestrians using specialized deep models for counting," in *Proc. IEEE Visual Communications and Image Processing (VCIP)*, December 2017, pp. 1–4.

[15] S. Ghosh, P. Amon, A. Hutter, and A. Kaup, "Pedestrian counting using deep models trained on synthetically generated images," in *Proc. 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP),Volume 5*, March 2017, pp. 86–97.

[16] L. Wang, J. Shi, G. Song, and I. Shen, "Object detection combining recognition and segmentation," in *ACCV (1)*, 2007, vol. 4843 of *Lecture Notes in Computer Science*, pp. 189–199.

[17] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 3258–3265.

[18] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *Performance Evaluation of Tracking and Surveillance workshop at CVPR 2009*, Miami, Florida, 2009, pp. 101–108.

[19] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," https://github.com/pdollar/toolbox.

[20] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[21] L. Wang, L. Xu, and M. H. Yang, "Pedestrian detection in crowded scenes via scale and occlusion analysis," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 1210–1214.

[22] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.