SCALE MAPPING AND DYNAMIC RE-DETECTING IN DENSE HEAD DETECTION

Zikai Sun, Dezhi Peng, Zirui Cai, Zirong Chen, Lianwen Jin

School of Electronic and Information Engineering, South China University of Technology
{eeszk,eedzpeng, eezrchen,eezrcai}@mail.scut.edu.cn, eelwjin@scut.edu.cn

ABSTRACT

Convolutional neural networks (CNNs) have demonstrated a strong ability to extract semantics from images during object detection; however, the extracted semantics are typically have strong scale priors for a specific circumstance. In this paper, we investigate the influence of head scale and contextual information, and then propose a scale-invariant method for head detection. Our method can dynamically detect heads depending on the complexity of the image. It uses an extra feature map to represent the scale information of the spatial relationship, and then uses this feature map for auxiliary detection. Particularly, we exploit several new techniques, including contextual information, scale invariance, and hard example mining. We evaluated our method on three head datasets and achieved state-of-the-art results for the Brainwash dataset, HollywoodHeads dataset, and SCUT-HEAD dataset.

Index Terms— Object detection, convolutional neural network

1. INTRODUCTION

Human head detection plays an essential role in modern people-counting-relevant applications and intelligent monitoring. Although tremendous strides have been made in general object detection, head detection in a crowd scene is still a challenging task because of high diversity, heavy occlusion, dynamic blur, low resolution, and rare features.

Many methods have been proposed to address this task. Gao *et al.*[1] generated proposals using HOG and used a CNN-SVM classifier to score the area. Stewart *et al.*[2] applied LSTM to decode representations into a set of detections. Li *et al.*[3] combined the region score and local score to assess a human head. However, all these approaches have limited performance.

Unlike faces, heads have few features. For instance, sunglasses or a gauze face mask can be features in some sense, whereas the back of a head in the distance can only be regarded as a dot. Worse still, heads always encounter the scenario of low resolution, blur, and occlusion. Identifying heads merely from the heads themselves is difficult. Inspired by HR[4] and GBD-net[5], which demonstrate the utility of contextual information, we introduce this concept to assist re-detection in our method rather than regarding the second step in Gao *et al.*s approach [1] as a classification task. We discuss the amount of contextual information that should be preserved for best performance in Section 2.4.

Another challenge is scale invariance. Most prior work set anchors to various sizes and aspect ratios to match different objects [6] [7] [8] [9]; however, these methods cannot eliminate the impact of scale thoroughly. Hu *et al.*[4] used the image pyramid; however, it is memory intensive. Yang *et al.*[10] and Zhang *et al.*[11] showed that modeling different filters for objects with different sizes is superior to providing results on different feature maps, although it is considered as expensive in terms of computational resources.

Thus, we question whether there is a "one-template-fits-all method to solve the multi-scale problem. We first confirm two hypotheses from empirical evidence: (i) roughly predicting the size of a head is easier than predicting the location and boundary precisely; and (ii) some regions become easy to detect when the area is resized into an appropriate size. Based on these observations, we propose the techniques of ScaleMap and area normalization, which make the second sub-network sensitive to a specific scale.

In this paper, we propose a new method called the ScaleMap detector (SMD) for head detection. First, a multi-task network roughly predicts heads and provides a ScaleMap that contains scale information about the scene. The weakly detected regions are then determined and normalized using the ScaleMap. The second sub-network then re-detects the normalized area and provides a more precise result. In our approach, choosing regions that are suited to the circumstances is critical. We consider scale invariance, contextual information, and hard example mining and then elaborate the region proposal section based on this.

Our contributions are summarized as belows:

- We present a novel method that can detect images from 0.1s to 0.4s on NVIDIA TitanXp, depending on the complexity of the image.
- Scale invariance, context information, and hard example mining are proposed to be useful for small objects.
- We achieve state-of-the-art results on three head detection datasets: the Brainwash dataset, HollywoodHead dataset, and our SCUT-HEAD dataset.

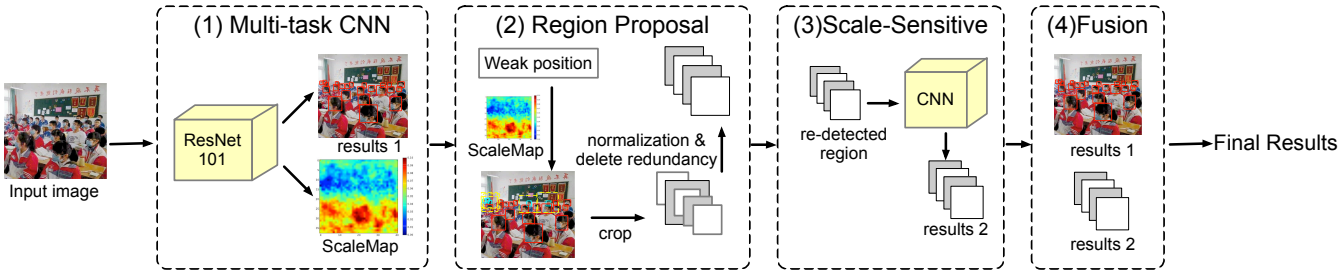

Fig. 1: Overall architecture of SMD: (1) a multi-task CNN is applied to provide a ScaleMap; (2) region proposal provides the weak detected position and proposes the re-detected area and its size using ScaleMap; (3) a lightweight CNN then re-detects the region; and (4) final results are provided by fusing the previous results.

2. SCALEMAP DETECTOR

2.1. Overall architecture

A lightweight network is efficient, but may fail to recognize complex images, whereas an expensive model is a waste of computational resources for numerous easy images. Therefore, we propose a method that can automatically take time based on the complexity of the image. An overview of the approach is shown in Fig. 1. Given a test image, the first multi-task CNN is used to provide a coarse result that cues the location of the weak detection areas. A ScaleMap can determine how much the region should be cropped, and then normalizes the region to 300px, thereby aiming to simplify the second detection. The second subnet is sensitive to specific-sized heads and is used to re-detect the hard but centralized object in an easy manner. The results map to the original image and are fused using non-maximum suppression (NMS). Thus, our approach can detect any size of head with similar accuracy by ignoring the distribution of the training datasets.

2.2. ScaleMap

In our approach, predicting the scale of the head correctly is of vital importance. In many weak detected occasions, such as viewing human hands or clothes as heads, directly expanding regions may not help. We present a method to generate a ScaleMap and define every value on the map as the scale that the head should be in that scene. Thus, detection can have a more global view and be assisted by every other object nearby. Regarding obtaining the ScaleMap, we first present an approach to transfer the ground-truth label to the ScaleMap. For each point on the ScaleMap p_{ij}, we traverse all the ground-truth boxes' centers p_k and calculate the Euclidean distance between them. Then we set its reciprocal and to the power of γ as the weight w_k, where γ is a modulating term and set to two in this paper. Point p_{ij}'s scale $SM(p_{ij})$ is the weighted average of all the label's scales:

$$SM(p_{ij}) = \frac{\sum_{k=1}^{K} w_k S(p_k)}{\sum_{k=1}^{K} w_k} \tag{1}$$

$$w_k = (\frac{1}{||p_{ij} - p_k||_2})^{\gamma} \tag{2}$$

where $S(p_k)$ is the ratio between the side length of the bounding box and the image, and is in the range (0, 1); that is, for each location, the size of the head is determined by all known head sizes, the weights are related to the distance, and the closer the head, the heavier the weight. We can obtain $\lim_{p_{ij} \to p_k} SM(p_{ij}) = S(p_k)$, which means that the value the ScaleMap is continuous.

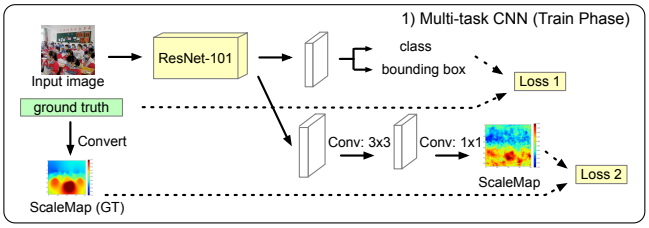

Fig. 2: Multi-task CNN: generate the ScaleMap additionally

The architecture of the multi-task CNN is shown in Fig. 2. In the training phase, the network first processes the input image into a ScaleMap label. We then add the output ScaleMap and calculate its loss. We add a 3×3 kernel after the backbone to enlarge the receptive field, followed by a 1×1 kernel so as to map to the ScaleMap. We use our multi-task loss as follows:

$$L_{total} = L_{cls} + L_{reg} + \alpha \cdot L_{scale} \tag{3}$$

where L_{cls} and L_{reg} are the classification and regression loss defined in RFCN [8], respectively. L_{scale} represents the difference between the estimated ScaleMap and the ground-truth scale map converted from the label; coefficient α is set to 3e-4 in the experiments:

$$L_{scale}(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} ||F_s(X_i, \Theta) - F_i||_2^2 \tag{4}$$

where Θ is the set of parameters of the CNN model, N is the number of training samples, X_i is the input image, and F_i is the ground truth scale map of image X_i. We use the Euclidean distance to calculate L_{scale}. The loss is minimized

using mini-batch gradient descent and backpropagation. Fig.3 shows an example of a visualized ScaleMap.

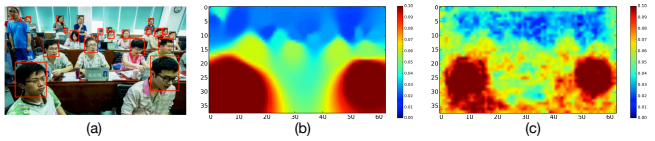

Fig. 3: Convert from a label to a ScaleMap: (a) the input image with head labels; (b) the converted ScaleMap; and (c) the output of the ScaleMap from the multi-task CNN.

2.3. Region proposal

In this section, we present our method to determine whether the image is difficult and how to re-detect it in a better manner in the case in which a difficult image is detected. Fig.4 shows the flow to the proposed regions.

1).Determine the locations: Prediction with moderate confidence always leads to uncertain circumstances compared with an extremely high or low score. Hence, we estimate the hard location using confidence scores. We first denote the output bounding box's center as p_d. Then we set all the bounding box centers within confidence range [0.3,0.7] as hard positions, denoted by $P_w = \{p_d \,|\, conf(p_d) \in [0.3, 0.7]\}$, where $conf(\cdot)$ represents the confidence of the bounding box.

2).Determine scales: We convert locations to regions by looking up the ScaleMap. In particular, we formulate a hard region as a tuple $\{p_w, l_w\}$, where p_w is the location in the image and l_w denotes the side length of the cropped region. We obtain $l_w = \beta \cdot SM(p_w)$, where β is the contextual coefficient of the region and set to 5 in the experiments. Hence, the set of hard regions can be represented as $D' = \{\,(p_w, l_w) \,|\, w = 1, 2, ..., W\}$.

3).Delete redundancy and normalization: To improve the speed of the model, we make redundant regions obsolete. We traverse each tuple $\{p_w, l_w\}$ in D', and if location p_w is covered by other regions in the set of regions D', then we delete $\{p_w, l_w\}$. The set of hard areas after traversing and deleting redundant elements is formed as D. We then use bilinear interpolation to normalize the set of regions to a fixed side length of 300px. The regions obtained at this time serve as the input for the second network.

4).Fusion: NMS is applied twice throughout the network. The first time is to narrow down the multi-task CNN's outputs as R. The second time is to acquire the fusion of R and N (all the output of the second subnet), which is the final output of SMD.

In the training phase, because we can supervise the model using the ground truth, we only have to determine the missing or wrong bounding boxes, and then extend the area β times, the same as the testing phase, the same as the testing phase, and regard it as the training data of the second subnet. As shown in Fig.4 (right), the yellow dotted rectangles represent the proposed regions.

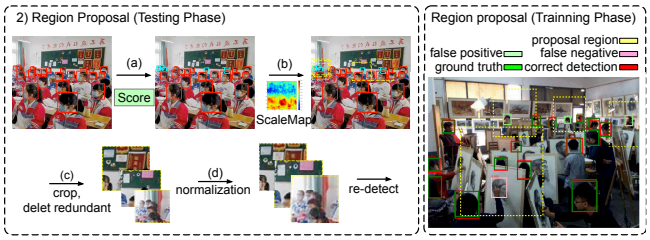

Fig. 4: Region proposal: (a) in the test phase, our method first determines the hard-detected location combined with the confidence, as shown in the blue boxes; (b) we provide a prior estimate of the head scale from the ScaleMap and extend the hard-detected region by β times, as shown in the yellow boxes; (c) we also delete the redundant region to consider efficiency; and (d) then we normalize the rest to a fixed size for further detection.

2.4. Analysis

1)Scale invariance: As shown in Fig.5, head sizes processed during the test phase are gathered together; the second model is also sensitive to a certain size. This matching strategy between the training and testing phases reduces the difficulty of the secondary network.

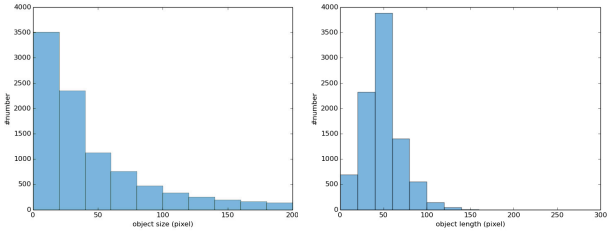

Fig. 5: Scale invariance: (a) the distribution of object scales in the original images; and (b) the object scale distribution in the cropped region processed (normalized to 300px). The size is resized from a large range [10, 150] to approximately 50px.

2)Contextual information: Heads without any context are difficult to recognize (as shown in fig.6 (b)). Therefore, an appropriate extended area is helpful. To integrate context, we extend the fields of view at different times in the original proposal box centered on the object. The relation curves for which accuracy varies with context are shown in Fig.6(c).

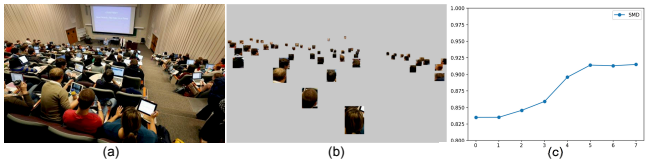

Fig. 6: Contextual information: increasing contextual information improves accuracy.

3)Hard example mining: Eliminating the influence of the majority of simple examples in the training phase is essential; OHEM [12], which was first used in Fast RCNN[13], and Focal Loss [14] in RetinaNet have all proven this. Thus, in the training phase, we only consider false positive and false negtive examples as training data in the second stage, which makes the subnet more sensitive for hard examples. The analysis is presented in Section 3.1.

3. EXPERIMENTS

3.1. Model analysis

To understand our model better, we conducted a series of ablation experiments and analyzed how each component affected the final performance. All models were trained and tested on SCUT-HEAD[1], which contains 4,405 images with 111,251 labeled heads. This dataset has 25.2 objects per image, on average, and is separated into two parts. Part A includes 2,000 images taken from classroom monitoring videos and Part B contains 2,405 images crawled from the internet

Table 1: Method ablation analysis

Methods	Results
R-FCN[8]	0.835
+fixed local enlarge without context	0.835
+fixed local enlarge	0.873
+hard example mining	0.881
SMD	0.915

We considered RFCN as the baseline architecture. From Table 1, Directly enlarging the region without contextual information resulted in no improvement. We then cropped the image using sliding widow methods with a step every 100px, scaled it to 300px and re-detected, with a result of 0.873, We later adopted a hard example mining strategy, and achieved 0.881. We further used our ScaleMap method to generate the size of the crop region, which had a result of 0.915.

3.2. Scale performance analysis

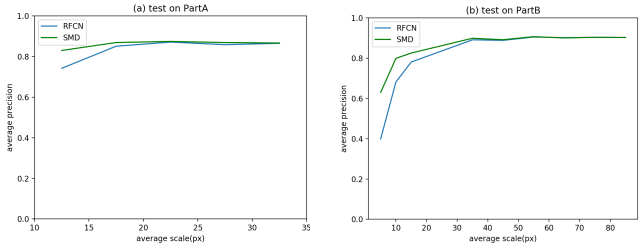

Fig. 7: Different scale performance

We also compared the performance of the RFCN and our method (SMD) on heads with different scales. SMD can normalize various scales of heads to a concentrated size; thus, the results deteriorate slightly when heads become smaller. Fig. 7 shows that our method can manage the scenario better when heads are small, occluded, and heavily blurred.

3.3. Evaluation on benchmarks

1) SCUT-HEAD dataset

A comparison of SMD with previous methods using the SCUT-HEAD dataset are presented in Table 2, where P, R, and H represent precision, recall, and harmonic mean, respectively. It can be seen that our method significantly outperformed all other methods with a large margin.

Table 2: Comparison between previous methods and SMD

Methods	PartA			PartB		
	P	R	H	P	R	H
YOLOv2[15]	0.91	0.61	0.73	0.69	0.69	0.69
SSD[7]	0.84	0.68	0.76	0.80	0.66	0.72
FRCN[16]	0.86	0.78	0.82	0.87	0.81	0.84
R-FCN[8]	0.87	0.78	0.82	0.90	0.82	0.86
SMD	**0.92**	**0.90**	**0.91**	**0.94**	**0.89**	**0.91**

2) Brainwash head dataset

The Brainwash head dataset[17] has 91,146 heads annotated in 11,917 images. All images are clipped from one coffee shop's surveillance camera. Our method performed well again. The results are shown in Table 3. AP denotes the average precision.

Table 3: Comparison on the Brainwash dataset

Methods	Con-local[18]	ETE-hung[17]	R-FCN	f-localized[19]	**SMD**
AP(%)	45.4	78.4	84.8	85.3	**90.04**

3) HollywoodHeads dataset

The HollywoodHeads dataset[18] contains 369,846 human heads annotated in 224,740 video frames from 21 Hollywood movies. It has a large number of images, but few heads per image. The results are shown in Table.4. It can be seen that, again, our method produced the best result.

Table 4: Comparison on the HollywoodHeads dataset

Methods	DPM face[20]	Con-local[18]	R-FCN[8]	**SMD**
AP(%)	37.4	78.4	86.3	**87.6**

4. CONCLUSION

In this paper, we proposed a new method called ScaleMap to represent the scale information of a scene rather than the object. We demonstrated its efficacy using our proposed SMD method, which performed better compared with previous methods, particularly on small blurred heads. We ascribe this to the better use of contextual information in a scale-invariance manner and give a heuristic thought that the scene may have more potential ability to assist object prediction.

5. ACKNOWLEDGMENT

This research was supported in part by the National Key Research and Development Program of China (Grant No. 2016YFB1001405), NSFC (Grant Nos. 61472144, 61673182, and 61771199), GD-NSF (Grant No. 2017A030312006), GDSTP (Grant Nos. 2015B010101004 and 2017A030312006), and GZSTP (Grant No. 201607010227).

[1]The SCUT-HEAD dataset can be downloaded from https://github.com/HCIILAB/SCUT-HEAD-Dataset-Release

6. REFERENCES

[1] Chenqiang Gao, Pei Li, Yajun Zhang, Jiang Liu, and Lan Wang, "People counting based on head detection combining adaboost and cnn in crowded surveillance environment," *Neurocomputing*, vol. 208, pp. 108–116, 2016.

[2] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng, "End-to-end people detection in crowded scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2325–2333.

[3] Yule Li, Yong Dou, Xinwang Liu, and Teng Li, "Localized region context and object feature fusion for people head detection," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 594–598.

[4] Peiyun Hu and Deva Ramanan, "Finding tiny faces," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1522–1530.

[5] Xingyu Zeng, Wanli Ouyang, Junjie Yan, Hongsheng Li, Tong Xiao, Kun Wang, Yu Liu, Yucong Zhou, Bin Yang, Zhe Wang, et al., "Crafting gbd-net for object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[6] Joseph Redmon, 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*, 2016, pp. 779–788.

[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, vol. 1, p. 4.

[10] Shuo Yang, Yuanjun Xiong, Chen Change Loy, and Xiaoou Tang, "Face detection through scale-friendly deep convolutional networks," *arXiv preprint arXiv:1706.02863*, 2017.

[11] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[12] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[13] Ross Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[15] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[16] Shaoqing Ren, 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*, 2015, pp. 91–99.

[17] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng, "End-to-end people detection in crowded scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2325–2333.

[18] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev, "Context-aware cnns for person head detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2893–2901.

[19] Yule Li, Yong Dou, Xinwang Liu, and Teng Li, "Localized region context and object feature fusion for people head detection," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 594–598.

[20] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "A deep pyramid deformable part model for face detection," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–8.