SCALE MAPPING AND DYNAMICALLY RE-DETECT 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

ABSTRACT

Though Convolutional neural networks (CNNs) have shown strong ability to extract semantics from images in object detection, the extracted semantics are usually object-related, not scene-related. In the natural scene, heads usually have strong scale priors to a specific circumstance. This paper investigates the influence of head scale and contextual information, then propose a scale-invariant method for head detection. Our proposed 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 use this feature map for auxiliary detection. Particularly, we exploit several new techniques including context information, scale-invariant, and hard example mining. We evaluate our method in three head datasets and achieve state-of-the-art results in *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 the modern people-counting-relevant applications and intelligent monitoring. Though tremendous strides have been made in general object detection, head detection in crowd scene is still a challenging task due to high diversity, heavy occlusion, dynamic blur, low resolution and rare feature.

Many methods have been proposed to handle the head detection task. Gao *et al.* generate proposals by HOG and uses CNN-SVM classifier to score the area[1]. Stewart *et al.* applies LSTM to decode representations into a set of detections[2]. Li *et al.* combines region score and local score to judge a human head[3]. However, all these approaches have a limited performance.

Differ from face detection, head as an object has fewer features on itself. For instance, sunglasses or gauze mask on a face can be features in a way, whereas a backward head in distance can only be regard as a dot. What's worse, heads always encounter the situation of low resolution, blur, and occlusion. Identify heads only form heads itself is difficult. HR[4], GBD-net[5] indicate that contextual information may help. Inspired by them, we use contextual information to assist detection, instead of regarding the second step as classification task in Gao *et al.*[1]. We also discuss how much contextual information should be kept for best performance in Sec. 2.4

Another crucial challenge is scale-invariance. Most prior work set anchors with various size and aspect ratio to match different objects [6] [7] [8] [9], but such method can't eliminate the impact of scale thoroughly. Hu *et al.*[4] utilize the image pyramid but is compute and memory intensive. Feature pyramids use different hierarchy's feature as different scales. Yang *et al.*[10] Zhang*et al.*[11] shows that modeling different filters for objects with different sizes surpass to giving results on different feature maps, though is considered expensive in computational resources.

So we wonder if there is a "one template fits all" method to solve the multi-scale problem. We first conclude two hypothesis from empirical evidence: (i) roughly predict the size of the head is easier than predicting the location and boundary precisely; (ii) some regions become easy to be detected when resizing the area into an appropriate size. Based on these observations, we propose the techniques of ScaleMap and area normalization in region proposal session, which makes the second subnetwork sensitive to a specific scale.

In this paper, we propose a new method called Scale Map Detector(SMD) for head detection. First, a multi-task network roughly predict the bounding box of the heads, besides, giving the ScaleMap that depict the scale information of the scene. Then, the network decides where the weak-detected regions are with the ScaleMap. After that, the weak-detected region will be normalized and re-detected by the second subnetwork, giving a precise result. In our approach, choosing the patch felicitously is critical. So we concern scale-invariance, contextual information and hard example mining and elaborate the patch decision section based on that.

Our contributions are summarized below:

- We present a novel method to detect different complex image in different computational sources, It can detects easy images in 0.1s, while less than 0.4s for complex images.
- Scale-invariance, context information and hard example mining are proposed and demonstrated to be useful for small object detection.
- We achieve the state-of-the-art results in three head detection datasets, including Brainwash dataset, Holly-

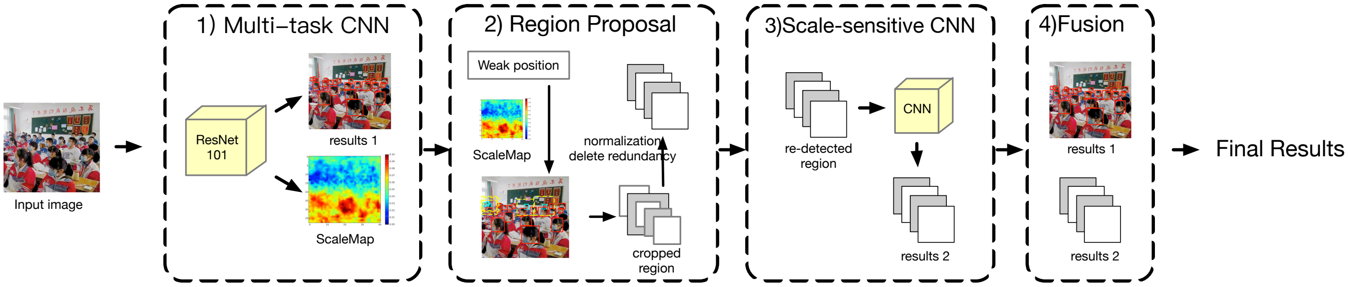

Fig. 1: The overall architecture of ScaleMap Detector(SMD): (1) a multi-task CNN is applied to give ScaleMap (2) Region Proposal gives the weak detected position and propose the re-detect area and its size using ScaleMap (3) a light-weight CNN then re-detects the region (4) final results are given by fusing the previous results

woodhead dataset, and our SCUT-HEAD dataset.

2. SCALEMAP DETECTOR

2.1. Overall Architecture

A light-weight network is efficient but may fail to recognize complex images, while an expensive model is a waste of computational resources to the numerous easy images. Therefore we propose a method that can automatically take time based on the complexity of the image. The 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 which cues where the weak detection areas are. A ScaleMap can tell how large should the region be cropped and then normalize the region to 300px, aiming to simplify the second detection. The second subnet is sensitive to specific size heads and is used to re-detect the hard but centralized object in an easier way. The result will map to original image and fusion by the non-maximun suppression (NMS). In this way, our approach can detect any size of the head with similar accuracy, ignoring the distribution of the training datasets.

2.2. Scale Map

In our approach, predicting the scale of the head correctly is of vital importance. While in many weak detected occasions such as taking human hands or clothes as heads, directly detect the expanded region may not help. Traditional methods to estimate the scale is object-relevant. While in the natural scene, the size of the head is highly relevant to the scene (e.g., the size of a bike or shoes near the head is highly relevant to the head). We present a method to generate ScaleMap. We define every value on the map as the scale of the head it should be on that scene. So detection can have a more global view and be assisted by every other object nearby.

To obtain ScaleMap, we first present how to transfer the ground truth label to the ScaleMap. To each point on the ScaleMap p_{ij}, we traverse all the ground-truth boxes' center p_k and calculate Euclidean distance with 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 label's scale. Formulas are shown as follow:

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

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

where $S(p_k)$ is the ratio between the side length of the bounding box and the image, ranges in (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 is, the heavier the weight is. Apparently we can get $\lim_{p_{ij} \to p_k} SM(p_{ij}) = S(p_k)$, which means the value on ScaleMap is continuous.

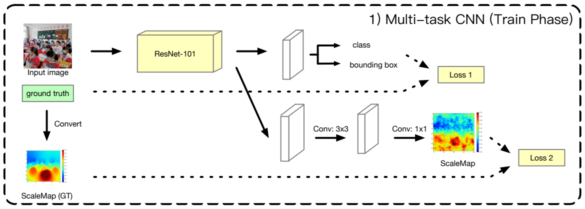

Fig. 2: Multi-task CNN: generate a scale map additionally

The architecture of the multi-task CNN is depicted in Fig. 2. In training phase, the network first processes the input image into ScaleMap label. We then add the output ScaleMap and calculate its loss. We add a 3x3 kernel after the backbone to enlarge the receptive field, then followed by a 1x1 kernel so as to map to the ScaleMap.

We employ our multi-task loss as follow:

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

where the first two items are the classification and regression loss defined in RFCN [8]. L_{scale} represents the difference between the estimated scale map and the ground truth scale map converted from the label. The coefficient α is set to 3e-4 in experiments.

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

where Θ is the set of parameters of the CNN model and 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 Euclidean distance to calculate the L_{scale}. The loss is minimized using mini-batch gradient descent and backpropagation. Fig.3 gives an example of visualized ScaleMap.

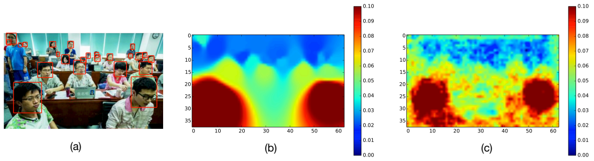

Fig. 3: Convert from label to ScaleMap: (a) shows the input image with head labels. (b) is the converted ScaleMap. (c) is the output of the ScaleMap from the first CNN.

2.3. Region Proposal

This section present our method to decide if the image is difficult and how to re-detect it in a better way in case a difficult image is detected. Fig.4 shows the flow to propose regions.

1).Determine the location: A bounding box with an extra high or low score rarely make mistakes, while middle confidence always leads to uncertain circumstances. So we use the confidence to speculates the hard region. We first denote the output bounding box's center as p_d. Then we set all the center of bounding box with a confidence range [0.3, 0.7] as the hard position, denoted as $P_w = \{p_d|Conf(p_d) \in [0.3, 0.7]\}$, where $Conf(p_d)$ is the confidence of the bounding box p_d.

2).Determine the scale: In our method, we decide the size of the hard region by looking up the ScaleMap. In particular, the side length of the cropped region denoted as $l_w = \alpha \cdot SM(p_w)$, where α is the contextual coefficient of the region. We have a discussion of its value further in Sec. 2.4.

3).Delete redundancy and normalization: In order to improve the speed of the model, we need to delete redundant hard areas. Traverse obtained location p_w, if the location is contained in the remaining hard areas, delete p_w from a set of elements made up of p_w, denoted as P_w. The set of hard areas after traversing and deleting redundant elements is formed as D. then we use the bilinear interpolation to normalize the set of region to a fixed side length of 300px. The region obtained at this time is served as the input for the second network.

4).Fusion: Non-maximum suppression (NMS) is applied twice throughout the network. The first is to filter R from Multi-task CNN's output, denoted as R', while the second is to acquire the fusion of R and N(all the output of the second subnet), which is the final output of SMD.

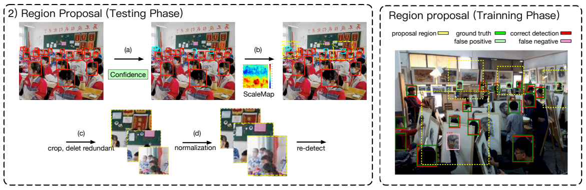

Fig. 4: Region proposal: (a)in the test phase, our method first decide the hard-detected location combine the confidence as shown in blue boxes. (b) then, give a prior estimate of the head scale from the ScaleMap and extend the hard-detected region by α times, shown in yellow boxes. (c) we also delete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

In training phase, since we can supervise the model by the ground-truth, we only have to find the missing or wrong detection bounding box, then extend the bounding box to α times as same as the testing phase, then take it as the training date of the second subnet. As shown in Fig.4 (right), the yellow dotted rectangle is the proposed region.

2.4. Analysis

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

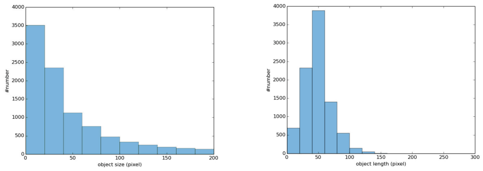

Fig. 5: scale invariance: (a) the distribution of object scales in original images. (b) the object scales distribution in cropped patch processed (patch region are normalized to 300px). The size is resized from a large range [10, 150] to around 50px.

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

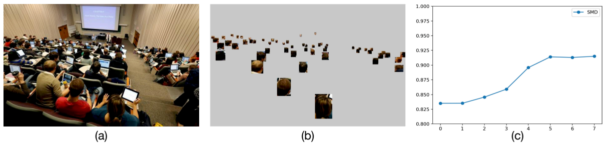

Fig. 6: contextual information: increasing contextual information improves accuracy.

3)Hard example mining: Eliminate the influence of the majority of the simple example in training phase is essential, OHEM [12] first used in Fast RCNN[13] or Focal Loss [14] in RetinaNet all proven this. So in training phase, we only take false positive and false negative example as training data in the second stage, making the subnet more sensitive on hard examples. Analysis is shown in Sec. 3.1.

3. EXPERIMENTS

3.1. Model analysis

To understand our model better, we carry out a series of ablation experiments and analysis how each component affect the final performance. All models are trained and tested on SCUT-HEAD[1], which contains 4,405 images with 111,251 labeled heads. This dataset have 25.2 objects per image on average and is separated into two part. PartA includes 2000 images taken from classroom monitoring videos from 7 classrooms during six days. PartB contains 2405 images crawled from the internet, which is variety in scale and environment.

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 consider RFCN that tiles anchors as same as our approach as the baseline architecture. From Table 1, we can see that directly enlarge the region without contextual information will have no improvement. We then crop the image using sliding widow methods step every 100px, scale it to 300px and re-detect, having a result of 0.873, we later adopt hard example mining strategy, achieving a result of 0.881. We further use our ScaleMap method to generate the size of the crop region, given a result of 0.915.

3.2. Scale performance analysis

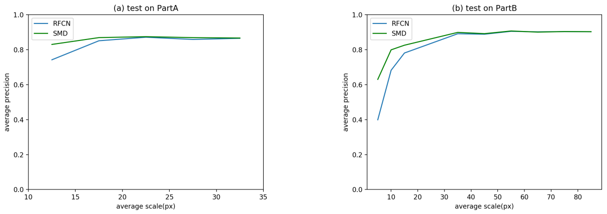

Fig. 7: Different scale performance

We also compare the performances of the RFCN and our method(SMD) on heads in different scales. SMD can normalize various scales of heads to concentrated size, so deteriorates slightly when heads become smaller. Fig. 7 shows that our method can handle the situation better when heads are small, occluded and blur heavily.

3.3. Evaluation on benchmarks

1) SCUT-HEAD dataset

The comparisons with different methods using the SCUT-HEAD datasets are given in Table 2. It can be seen that our method significantly outperforms all other methods with a large margin.

Table 2: Comparation 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. Our method also performs well. Results are shown in Table 3.

Table 3: Comparation on 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

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. Results are shown in Table.4. It can be seen that, again, our method produces the best result.

Table 4: Comparation on Hollywood dataset

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

4. CONCLUSION

This paper propose a new method named ScaleMap to represent the scale information of the scene, rather than the object. We demonstrate its efficacy by our proposed SMD method, which has a better performance, especially on the small blur heads comparing with previous methods. We ascribe this to the better utilization of contextual information in a scale-invariance way and give a heuristic thought that the scene may have more potential ability to assist object prediction.

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

5. 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.