Detection of Ethnic Minority’s Symbols Based on Deep Learning

Ye Feng1,3, Zhuo Shi1,2,3, Qian Kong2a, Mingrui Li1, Ming Yang2, Mengxue Zhang2 and Shuzhen Zeng1

1 School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin China
2 School of Arts & Design, Guilin University of Electronic Technology, Guilin China
3 Institute for Artificial Intelligence Interdisciplinary, Guilin University of Electronic Technology, Guilin China
a 847893236@qq.com

Abstract. Symbols are the embodiments of the ideology of ethnic minorities, the detection of ethnic minority’s symbols is an important part of the protection and inheritance of ethnic cultural heritage. Nevertheless, the ethnic symbols are miscellaneous and referred to some special combination style, it is time-consuming and inaccurate to detect them only by traditional detection methods based on machine learning. Therefore, in this paper, we propose a deep learning method based on Faster R-CNN to detect various kinds of Zhuang minority’s symbols from the Z-S dataset. Firstly, the original images of Zhuang minority’s symbols are preprocessed. Secondly, the processed images are sent to the ResNet-50 with FPN to extract feature maps. Moreover, RPN processes those feature maps to generate bounding boxes. Finally, ROI pooling layer in R-CNN converts those bounding boxes into fixed-length feature vectors, which fed into two sibling fully-connected layers for further detection tasks. Through the ablation experiments based on a certain amount of ethnic symbol images, the results indicate that the proposed method has higher detection quality (mAP), and can effectively reduce the training cost.

1. Introduction
Symbols are important carriers and external expressive form of ethnic culture. Compared with other carriers such as language, architecture and craft, ethnic symbols that are attractive and rich in cultural connotations are easier to be promoted to the public. According to the sixth national census, Zhuang minority is the most widely distributed ethnic minority in Guangxi. The symbols on the clothing and brocade of each branch of Zhuang minority are different in shape, color and intensity, which are deeply influenced by the ethnic ideology and have profound implications. Therefore, the detection of Zhuang’s symbols is of great significance to the continuation and spread of Zhuang’s culture.

However, due to the complexity of symbols and lack of standard image dataset, it’s often difficult to realize high-quality detection of ethnic minority’s symbols. In addition, the difference between some classes of symbols are not obvious. Therefore, this relatively new field has attracted the attention of many researchers and proposed corresponding solutions. Chen J G et al. [1] used Scale Invariant Feature Transform (SIFT) algorithm, Speeded up Robust Features (SURF) algorithm to implement feature extraction of embroidery patterns of Manchu court costumes, then used Best-bin-First (BBF) algorithm to process the extracted feature vector. After the gradient direction accumulation, matching point pairs of feature points were obtained. Wu S et al. [2] used k-poselets approach to perform global and local key point prediction, then HOG, LBP, SIFT features of the training images were extracted,
finally a multitask learning classifier model was trained to obtain different styles of minority clothing. Chen J Z et al. [3] proposed a method binding texture block global matching and Grabcut segmentation, the relative orientation between similar patterns is estimated by a local patch match. Then the pre-segmentation of ethnic patterns is co-optimized to preserve the main structures. All of the above are traditional detection methods based on machine learning, which have the advantages of mature technology and complete related technical documents. However, such methods mainly rely on artificially designed feature extractors to detect objects, lacking high-level semantic information. Obviously, in the scene where the symbol pattern is more complicated, omissions and incorrect detections may occur.

In recent years, with the improvement of GPU computing capabilities and the development of convolutional neural networks, deep learning methods have been widely used computer vision tasks such as object detection and semantic segmentation. Deep learning based object detection frameworks can be primarily divided into two categories: (i) two-stage detectors based on region proposals, such as Fast R-CNN [4], Faster R-CNN [5], etc. This type of framework generates anchor boxes according to specific rules, then the classification and correction of coordinates of detection boxes are performed on the basis of anchor boxes. This type of framework has higher accuracy, but it takes up more memory and has a slower detection speed. (ii) one-stage detectors based on regression, such as SSD [6], YOLO [7], etc. This type of framework does not need to generate region proposals, which has fast detection speed and is suitable for the scene requiring real-time detection. However, its detection accuracy is lower than former framework, and it is more sensitive to the size of the object and the environmental interference from the background.

In this paper, we propose a detection method of Ethnic Minority’s Symbols Based on Faster R-CNN. This method first extract symbol image features using ResNet-50 network with FPN structure, then those feature maps are transformed by RPN to generate region proposals, the proposals are subsequently refined by a softmax classifier and a bounding-box regressor. Finally, ROI pooling layer uniformly converts those bounding boxes into fixed-length feature vectors, these vectors are fed into sibling classification layer and regression layer in RCNN—to predict the categories of objects in the detection boxes and modify coordinates of the detection boxes. To analyse the detection ability of the proposed method, we run several ablation experiments comparing with other detection frameworks. The results demonstrate that our method achieves 89.11 mAP, which has certain efficiency improvement.

2. Proposed Method

2.1. ResNet-50 with FPN

Generally, a deeper network can improve both recognition and localization. The original Faster R-CNN uses VGG-16 with shallow layers to extract features, directly increasing the number of layers of this network may cause a problem of vanishing gradients and performance degradation. Resnet [8] is a better choose to avoid such problem.

![Figure 1. Residual module](image-url)
Residual module is the basic component of the ResNet, as shown in Figure 1. ResNet uses BN[9] layers to normalize the input features of the convolutional layer, then the ReLU function is used to activate the features, which initially solves the problem of vanishing gradient. Assuming $x_i$ is the input feature of the i-th residual module, $W_i$ is the weight matrix, $x_{i+1}$ is desired output feature, $G(x_i, W_i)$ is the residual mapping while shortcuts denote skipping one or more layers to get optimal mapping. Therefore, the forward propagation of ResNet can be defined as:

$$x_i = x_i + \sum_{h=i}^{H-1} G(x_h, W_h)$$

(1)

Assuming the calculated loss as $L$, then the expression of backward propagation is derived:

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial x_i} \frac{\partial x_i}{\partial x_i} = \frac{\partial L}{\partial x_i} (1 + \frac{\partial}{\partial x_i} \sum_{h=1}^{H-1} G(x_h, W_h))$$

(2)

The calculation of the gradient consists of two parts, the linearity of them ensures features can be fully propagated back to the shallow layer, which solves the problem of network performance degradation.

![Figure 2. Comparison of residual module and bottleneck structure](image)

In ResNet-50, the residual module should be modified to a bottleneck structure. As shown in Figure 2(b), the 1x1 convolution layers are used to reducing and then increasing dimensions, in order to reduce parameters required for 3x3 convolution. The amount of parameters used as a whole is 1x1x256x64 + 3x3x64x64 + 1x1x64x256 = 69632; if the bottleneck structure is not used, as shown in Figure 2(a), the parameter amount is 1179648, with a difference of 16.94 times. This proves that the bottleneck structure can reduce calculations of ResNet.

As shown in the figure 3, the structure of FPN[10] can be divided into three parts: bottom-up process, horizontal connection of features and top-down process.

![Figure 3. FPN structure and merging detail](image)
The bottom-up process is actually the forward propagation process of ResNet network. In this process, the phase steps of Conv1, Conv2, Conv3, Conv4 and Conv5 stages are 2, 4, 8, 16 and 32 respectively. In the considerations of efficiency, the Conv1 stage is skipped by default, and the feature activation output of the last residual module in the next 4 stages is selected to obtain C2, C3, C4, and C5 feature maps with mentioned scales, which are then combined into basic feature pyramid.

The basic feature pyramid requires horizontal connections of features and a top-down process to synthesize the final feature pyramid. Firstly, perform a 3x3 convolution on C5 to obtain the P5 feature map. Secondly, the P5 feature map is upsampled by a factor of 2. Moreover, perform a 1x1 convolution on C5 to convert feature map to 256 dimension, then merge it with the C4 feature map by horizontal connection to perform element-wise addition. Finally, perform 3x3 convolution on the merged map to generate the P4 feature map, which is to reduce the aliasing effect of upsampling. Iterate this process from top to bottom, until the P2 feature map is synthesized. The final set of feature maps is called \{P2, P3, P4, P5\}, corresponding to \{C2, C3, C4, C5\}.

2.2. Region Proposal Network

The structure of Region Proposal Network (RPN) is shown in figure 4. In general, RPN use 3x3 sliding window to process the feature maps delivered by ResNet-50 with FPN, to generate a certain number of region proposals. The mapping of the sliding window on the feature map is called anchor, combined with 5 scales of \{32^2, 64^2, 128^2, 256^2, 512^2\} and 3 ratio of \{1:1, 1:2, 2:1\}, RPN will produce 15 types of anchors, then convert them into region proposals on the original image. Each anchor will be transformed into a 256-d feature vector after the convolution operation of the intermediate layers, which is refined by a classification layer and a regression layer. The classification layer is used to determine whether the proposal belongs to the foreground (or background), the regression layer is used to fix bounding boxes of the proposals, make them more closer to the ground truth. Subsequently, these revised proposals will be introduced into the R-CNN together with the feature maps.

![Figure 4. RPN structure](image)

2.3. Region-Based Convolutional Network

Actually, the R-CNN section of Faster R-CNN framework follows the configuration of the same section in previous Fast R-CNN framework. ROI pooling layer in R-CNN converts bounding boxes of revised proposals into fixed-length feature vectors, which fed into sibling fully-connected classification layer and regression layer layers for detection tasks. The classification layer is used to detect the specific category of the object in the revised proposals, and the regression layer is used to modify the coordinates of the detection boxes. Additionally, the Soft-NMS [11] algorithm can be used to remove redundant detection boxes. According to the previous descriptions, the structure of Faster R-CNN is shown in Figure 5.
3. Experiments

3.1. Z-S Dataset

The images of Zhuang minority’s symbols used in this experiment are collected through taking photos of Zhuang minority’s costumes and brocade patterns in museum, visiting ethnic settlements for field research and downloading images from the Internet. According to the statistics, there are 1260 images covering 22 kinds of symbols, the sample images are shown in Figure 6. Before model training, the original images need to be preprocessed. Firstly, the format of all images is converted to png, and some images are cropped for their sizes exceeding 2 MB. Secondly, we annotate the objects on these images with the format of Pascal VOC dataset [12], referring to the pre-divided categories. Considering this dataset is related to Zhuang minority’s symbols, we name it Z-S dataset.

3.2. Ablation Experiments and Result Analysis

To analyse the detection ability of the proposed method, we run several ablation experiments. The common environment of these experiments is as follows: The operating system is Linux (Ubuntu 16.04), the deep learning framework is Tensorflow 1.8 and PyTorch 1.0, the CPU is Intel i7 7700k, the GPU is Tesla P100 (16G).

Firstly, we perform ablation experiments on the Z-S dataset to verify the effectiveness of FPN.
which is involved in region generation of RPN. All experiments are based on ResNet-50, and the learning rate is 0.005. Average Recall (AR) is the metric of these experiments, AR\textsuperscript{100} and AR\textsuperscript{1k} denotes 100 and 1000 proposals per images. Experiments results are shown in Table 1.

| RPN          | feature | horizontal | top-down | AR\textsuperscript{100} | AR\textsuperscript{1k} | AR\textsuperscript{1k} |
|--------------|---------|------------|----------|------------------------|-----------------------|------------------------|
| (a) baseline on conv4 | C4       |            |          | 38.3                   | 50.6                  | 33.7                   |
| (b) baseline on conv5 | C5       |            |          | 38.7                   | 46.9                  | 27.2                   |
| (c) bottom-up pyramid | {P2~P5} | √          |          | 39.5                   | 51.2                  | 32.7                   |
| (d) bottom-up pyramid | {P2~P5} | √          | √        | 36.6                   | 47.8                  | 28.9                   |
| (e) FPN      | {P2~P5} | √          | √        | 46.1                   | 58.4                  | 46.9                   |

In Table 1, the columns “horizontal” and “top-down” denote the presence of horizontal connections and the top-down process respectively. The column “feature” denotes the feature maps on which the RPN heads are attached. Table 1(b) shows no advantage over Table 1 (a), suggesting that a single feature map of higher level may not be efficient because of the trade-off between coarser resolutions and stronger semantic information. Table 1(d) shows binding FPN with RPN improves AR\textsuperscript{1k} to 58.4, which is 7.8 points increase over the original RPN with single scale(Table 1(a)). Moreover, the AR\textsuperscript{1k} is boosted by 13.2 points. These results indicate that FPN makes up for the shortcomings of the original Faster R-CNN directly using a single-scale feature map to generate region proposals, enhances the quality of proposed regions, especially for those regions covering small objects.

Then, we perform ablation experiments on the Z-S Dataset to compare the detection efficiency of Faster R-CNN to other frameworks, such as YOLO, SSD, Fast R-CNN, R-FCN. In these experiments, the learning rate is 0.01, the number of training epochs is 5000, and the interval of the checkpoints is 200. The training set consists of 896 images, the verification set consists of 182 images, and the test set consists of 182 images. We select mean Average Precision (mAP) as the metric of these experiments, experiment results are shown in Table 2.

| framework     | backbone | FPN | Soft-NMS | mAP (%) |
|---------------|----------|-----|----------|---------|
| Fast R-CNN    | VGG-16   | √   |          | 78.62   |
| Fast R-CNN    | ResNet-50| √   |          | 84.35   |
| Fast R-CNN    | ResNet-50| √   | √        | 87.11   |
| YOLO          | VGG-16   |     |          | 69.84   |
| SSD           | VGG-16   |     |          | 78.73   |
| R-FCN         | ResNet-50|     |          | 86.95   |
| Faster R-CNN  | VGG-16   | √   |          | 82.55   |
| Faster R-CNN  | ResNet-50| √   |          | 87.17   |
| Faster R-CNN  | ResNet-50| √   | √        | 89.63   |

Table 2 shows that the mAP of our proposed method is 89.63, exceeding other four frameworks, which is 19.79 points higher than YOLO. Compare Faster R-CNN (ResNet-50) with Faster R-CNN (VGG-16), the mAP is boosted by 4.62 points, for Fast R-CNN such modification leads to a promotion of 5.73 points, it’s obvious that a better feature extractor can directly improve detector’s performance. As combing ResNet-50 with FPN, deep features with low resolution and rich semantic information, which play a key role in classification, well integrated with the shallow features with high resolution and rich location information, which are more suitable for localization of small object. Thus, FPN improves RPN’s robustness to object scale variation and enhances detector’s performance of low-pixel objects. Compare our method with Faster R-CNN (ResNet-50,Soft-NMS), the mAP is boosted by 2.46 points, for Fast R-CNN such ascension is 2.76 points. We randomly select two images from the test set, our method’s detection result for the selected samples is shown as Figure 7.
As there are 22 categories of symbols for 9 set of ablations, the explicit AP of them won’t be listed. Checking the logs of all ablations, we find that the APs of Zhuang four-cornered hibiscus symbols and Zhuang eight diagram symbols are relatively low, regardless of which model from checkpoints we choose to use. Delving into the Z-Y datasets, we find that the amount of images of these two types is quite few. Moreover, these symbols are combined with other types of symbols or intersected with some irrelevant figures in the same images, which makes them being neglected or detected as secondary significant objects. To improve the detection results of them, the relevant images should be cropped properly or labeled more precisely. Moreover, it indicates the inherent obstacles of VOC-format dataset and two-stage detectors. The shape of labeled box used by VOC-format dataset must be rectangle, making it not adjust to some complex-shaped object. For two-stage detectors, there’s a similar problem that anchor may be limiting factor for performance improvements. The other format of dataset such as COCO [13], and anchor-free detectors with special design may help to improve the circumstance.

4. Conclusion
In this paper, we propose a method based on deep learning for detection of ethnic minority’s symbols with Faster R-CNN framework. Firstly, the original images of Zhuang minority’s symbols are prepossessed to set up the voc-format Z-S dataset. Secondly, the images loaded from Z-S dataset are sent to the ResNet-50 with FPN to extract feature maps. Moreover, these feature maps are transformed by RPN to generate region proposals. Finally, ROI pooling layer in R-CNN converts bounding boxes of revised proposals into fixed-length feature vectors, which fed into two sibling fully-connected layers to detect the specific category of the object and modify the coordinates of the detection boxes. The mAP of the proposed method is 89.63, exceeding other four frameworks in ablation experiments. In future research, we will collect more images to extend Z-S dataset and transform it to coco-format, try out anchor-free detection methods, or combine anchor-based methods with semantic segmentation methods.

5. Acknowledgments
This work is jointly supported by the National Natural Science Foundation of China (61862018); Guangxi Natural Science Foundation Project(2018GXNSFAA138084); Project of Guangxi Key Laboratory of Intelligent Processing of Computer Image and Graphics (GIIP201404); Guangxi Experiment Center of Information Science (YB1506).

6. References
[1] Chen J G, Cao C, Ren B, Ma L. Research on Application of Feature Extraction and Matching Algorithms in Chinese Costume Patterns. *Journal of Silk*, 2015, 52(05):36-41.
[2] Chen J Z, Zheng J, Lu S F, et al. Co-optimization of ethnic-pattern segmentation based on hierarchical patch matching. *Scientia Sinica Informationis*, 2019, 49(2): 188-203.
[3] Wu S M, Liu L, Fu X D, Liu L J, Huang Q S. Human detection and multi-task learning for minority clothing recognition. *Journal of Image and Graphics*, 2019, 24(04):0562-0572.

[4] Girshick R. Fast R-CNN. *Proceedings of the IEEE international conference on computer vision*, 2015: 1440-1448.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards realtime object detection with region proposal networks. *Advances in neural information processing systems*, 2015: 91-99.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. *European conference on computer vision*, 2016: 21-37.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 779-788.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 770-778.

[9] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017: 2117-2125.

[11] Bodla N, Singh B, Chellappa R, et al. Soft-NMS. *Proceedings of the IEEE international conference on computer vision*, 2017: 5561-5569.

[12] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

[13] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. *European conference on computer vision*, 2014: 740-755.