Mask R-CNN for Indonesian Shadow Puppet Recognition and Classification

Ida Bagus Kresna Sudiatmika¹*, Made Artana², Nengah Widya Utami³, Made Adi Paramartha Putra⁴, Eka Grana Aristyana Dewi⁵

¹,⁴Informatic Engineering, STMIK Primakara, Denpasar, Indonesia
² Information System, STMIK Primakara, Denpasar, Indonesia
³,⁵Accounting Information System, STMIK Primakara, Denpasar, Indonesia

*Kresna@primakara.ac.id

Abstract. Region-based Convolutional Neural Networks (R-CNN) have achieved remarkable achievements in object recognition and detection. In this paper, Mask R-CNN is introduced to detect and recognize Indonesian Shadow Puppet patterns. Indonesian Shadow Puppet (Wayang) is one of the traditional Indonesian arts that depict stories, one of the stories is the Mahabharata. Our aim for conducting this research is to protect Indonesian Puppets (Wayang) from extinction through systemic pattern recognition. The algorithm used in this study is Mask R-CNN. Mask R-CNN training to recognize Indonesian Shadow Puppet requires many labels that are difficult to obtain. Therefore, the authors carried out data augmentation to add datasets to improve training. Furthermore, testing is done using several learning rates. This test is done using a Cloud GPU. The results of the tests show that the accuracy of the training obtained is 92.04% and the accuracy of validation with the learning rate value set is 0.0001. Testing the model in this study successfully tested the puppet pattern.

1. Introduction

Shadow Puppet is one of Indonesia’s traditional arts that is growing rapidly in Java and Bali. Puppets are usually performed in religious ceremonies and arts. Puppet shows in Indonesia have their own speech and unique style which is an original masterpiece from Indonesia. Wayang (leather puppets) has been recognized by UNESCO as a Masterpiece of Oral and Intangible Heritage of Humanity [1]. Wayang is recognized as great works because they have a high value for human civilization. However, Wayang as cultural heritage can be destroyed if there is not enough appreciation from both the community and the government for puppet artists. Thus, by doing this research aims to preserve Wayang from extinction. By using object recognition can help people, especially young people, to recognize the puppet objects from the patterns found in the puppets themselves.

The dataset used in this study is the Indonesian Shadow Puppet in the Mahabharata story. The author took pictures of various art venues in Bali and puppet museums in Bali. One of the museums used for dataset collection is the Setia Darma Mask and Puppet House which is located in Ubud, Bali. In collecting datasets, the author experiences difficulties that are the limitation of the collected datasets. Therefore, augmentation data is performed to add to the dataset.

Data augmentation is a technique used to generate new training samples from original data by applying random jitters and perturbations (but not changing class labels from data). There are several applications...
for data augmentation carried out in the study. To reduce errors in leaf pattern recognition, Zhang et al. [2] adopted data augmentation to reduce over-fitting. Jiang-Jing Lv et al. [3] propose five data augmentation dedicated to face image, including perturbation and four synthesis landmarks. In this study, the methods used for augmentation are flipping, color casting, cropping, and blur.

Deep learning is an algorithm that can improve detection and recognition capabilities [4]. Compared to traditional methods, deep learning has good abilities such as motion blur, perspective distortion and changes in illumination. For instance, KuoWang et al. [5] conducted research with deep learning using the CNN algorithm for Cross-Spectral iris recognition. Wenjin Tao et al. [6] applied CNN to recognize the pattern of the American Sign Language Alphabet with multiview augmentation and inference fusion. Among existing deep learning algorithms, R-CNN has outstanding performance in many aspects. For example, Dong Wang et al. [7] conducted a study to detect goat dairy activity by using a faster R-CNN algorithm and obtains 92.49% average precision.

Furthermore, in this study, the problems will be solved in pattern recognition that exists in puppets in order to preserve Balinese Shadow Puppet through object recognition. Moreover, Balinese Shadow Puppet have different patterns and carvings on each puppet so that they can be recognized or distinguished from them. In particular, this study shows that Mask R-CNN is a sophisticated algorithm that is for object detection and segmentation quickly and with effective results. The Mask R-CNN was used because this algorithm is the latest algorithm and it has succeeded in detecting and recognizing objects.

This study will focus on analyzing Indonesian Puppets research. Many have conducted research on Balinese Shadow Puppets, as have Sudiatmika et al. [8] conducted research on Indonesian Wayang in terms of puppet image classification with the CNN algorithm. In this research, they compared the architecture of VGG and Alexnet. The best results they got showed that VGG-16 has a very good ability in recognizing puppet objects. Arik Kurnianto et al. [9] conducted research by representing visual puppet characters in games. The type of leather puppets used in this study is Purwa Puppet.

### 2. Methodology

![Mask R-CNN Architecture](image)

The Mask R-CNN is an extension of Faster R-CNN which has two outputs from each object candidate. Where the output of the Mask R-CNN is a label of the object class, bounding-box object and there is an object mask from the recognition. In this section, it would distinguish between parts of the Mask R-CNN and their uses for object recognition [10].

Backbone Network used in this study is the CNN architecture which is used to extract features from puppet images. The Feature Pyramid Network (FPN) is a very important architecture as the extraction
of features in the Mask R-CNN [11]. This study is used the ResNet-50-FPN architecture to extract image features. The FPN uses a top-down architecture to build high-level semantic features on all scales that are suitable for detecting different types of Puppets. ResNet 50 [12] is a Residual Network with 50 layers. ResNet also has several types namely ResNet101 and ResNet152. We use ResNet 50 because ResNet50 is faster since it has fewer layers.

Region Proposal Network introduced in the Faster R-CNN network which is a fully convolutional network that takes feature extraction from the backbone network and proposes candidate objects that will form a bounding box equipped with score values for each object in the image [13]. The performance of the next RPN is that the window moves spatially on the feature map, the size of the sliding window is $n \times n$. In each sliding window, there are 9 anchors that have been generated which have the same center on the point $(x_a, y_a)$. However, each anchor has 3 aspect ratios and different scales. Next, on each anchor the $p^*$ value will be calculated where this value will indicate how much the anchor will overlap with the ground-truth bounding box. The $p^*$ formula is determined as follows.

\[
p^* = \begin{cases} 
1 & \text{if } \text{IoU} > 0.7 \\
-1 & \text{if } \text{IoU} < 0.3 \\
0 & \text{otherwise}
\end{cases}
\]

which IoU is intersection over union and is defined below:

\[
\text{IoU} = \frac{\text{Anchor} \cap \text{GT Box}}{\text{Anchor} \cup \text{GT Box}}
\]

To form bounding box regression, the following formula is adopted to give parameters to 4 coordinates [14]. The formula is shown as follows.

\[
t_x = \frac{x - x_a}{w_a}, \quad t_y = \frac{y - y_a}{h_a}, \\
t_w = \log \left( \frac{w}{w_a} \right), \quad t_h = \log (h - h_a), \\
t_{x'} = \frac{x' - x_a}{w_a}, \quad t_{y'} = \frac{y' - y_a}{h_a}, \\
t_{w'} = \log \left( \frac{w'}{w_a} \right), \quad t_{h'} = \log (h' - h_a),
\]

Which $(x, y, w, h)$ shows the center coordinates of the box, width, and height. Variable $x, y, x', x'$ are used to indicate the predictive box, ground-truth box, and anchor box.

Furthermore, $n \times n$ spatial characteristics that have been extracted from the convolutional network feature map are forwarded to smaller networks which have two tasks: classification and regression. The output of the regression is to determine the outcome of the bounding-box prediction $(x, y, w, h)$. The output object of classification is the probability $p$ whether the prediction box contains an object or only background.

Roi Align One contribution of the Mask R-CNN is Roi Align which does not digitize cell boundaries and makes all targets the same size. It also applies interpolation to calculate the feature map in a cell better. Roi Align made a significant improvement in accuracy. At this stage Roi Align Features has passed through the base of the network where there are 3 parallel tasks, namely classification, bounding box regression and mask segmentation. The Classifier and Regressor understate the features of Roi Align and convert them to vectors by the Fully Connected (fc) layer. In determining the classification at the end of the network there is a Softmax activation that was used to calculate the probabilities of each target class. The formula for the Softmax function is as follows.
\[ x_j = \frac{\exp(x_j)}{\sum_{j=0}^{m} \exp(x_j)} \]  

(7)

The mask branch predicts the \( m \times m \) mask of each Roi using Fully Convolutional Network. The mask out of this network is in the form of instance segmentation. The Mask R-CNN is the most advanced method of detecting targets by extending the framework from Faster R-CNN to make it faster by adding branches at the end of the model. Through the stages, in the Mask R-CNN the expected output is the target classification and instance segmentation which will also generate the classification score, bounding box, and mask segmentation.

Before doing the training, the image was resized to 512 x 512 px. In order to complete the R-CNN requirement (larger training data), the augmentation data described previously was added. The network was evaluated by testing learning rates with different values. The learning rate (lr) was used \([0.0001, 0.001, 0.01]\). The test was carried out at each different learning rate value with epoch value 100. First, the data was printed through the backbone network which was composed of the ResNet 50 network.

In this study, the images of the Mahabharata Puppets were used and had been collected from various places (arts, museums and several temples) in the Bali Region. The types of puppets were tested: Yudistira Puppet (150 images), Arjuna Puppet (200 images), Bima Puppet (210 images), Nakula Puppet (250 images), Sahadewa Puppet (183 images), and Gatot Kaca Puppet (210 images). The dataset was divided into 2 parts, namely for training data and testing data with an 80% ratio for training data and 20% for validation data.

Deep Learning Networks typically require large amounts of data. So one way to add a dataset is to do data augmentation to add more dataset. The addition of datasets with augmentation also aims to reduce the over-fitting as well as our goal to do augmentation data is to improve the generalization of the model [15]. The augmentation data techniques were applied to this study are Translation, Rotation, Shearing, Horizontal and changes in scale. Below is the result of augmentation data.

![Figure 2. Example Image Augmentation Result](image_url)
specification was used in this study that can be accessed by Cloud-based (https://www.floydhub.com/). Furthermore, Tensorflow and Keras Deep Learning Frameworks were implied which both of are python libraries.

3. Result and Discussion
The performance evaluation of the Mask R-CNN in the puppet dataset has been done through the experimental stages. It can be concluded that the results by comparing the accuracy of the training with the accuracy of the validation during training. After conducting training on the dataset that has been provided, then the results were compared. The results of the training are shown in the following figure 3, figure 4 and figure 5.

Figure 3. Training and validation accuracy (lr = 0.01)

Figure 4. Training and validation accuracy (lr = 0.001)
From Figure 3, Figure 4 and Figure 5 show that using learning rate 0.0001 has the best learning rate with accuracy reached up to 92.04%, while accuracy achieved using learning rate 0.001 is 90.38% while learning rate 0.01 reaches learning value the worst is 10.17%. So in making the training model we use a learning rate of 0.0001. Next, some data were tested for testing the Mask R-CNN. In order to validate the model, the data test was applied. The test results are shown below:
From the test results using Wayang Arjuna, the recognition accuracy of 90.70% was obtained and the results of testing using Yudistira puppets obtained an accuracy of 90.76%. This shows that Mask R-CNN successfully carried out pattern recognition in the Balinese Shadow Puppet. Mask R-CNN successfully recognize shadow puppets pattern with region on each image which is contain image information (Figure 7 and 9). Then the use of this algorithm is very likely to be used. So that later this algorithm can be used as an alternative in pattern recognition.

4. Conclusion
In this paper, a convolution network specifically the R-CNN Mask to recognize the Indonesian Shadow Puppet pattern which is formulated for object detection and classification is proposed. In this study, the training on the Indonesian Shadow Puppet dataset was conducted that had been prepared previously. The results of a comprehensive experiment on the recognition of the Indonesian Shadow Puppet and classification with the Mask R-CNN for this problem work properly in class objects. By using several levels of learning for training namely 0.01, 0.001 and 0.0001 obtained the best results using the learning level of 0.0001 with an achievement level of 92.04%. Tests carried out on several test data and the model succeeded in recognizing puppet patterns. From the test results using Wayang Arjuna, the recognition accuracy of 90.70% was obtained and the results of testing using Yudistira puppets obtained an accuracy of 90.76%. This shows that Mask R-CNN successfully carried out pattern recognition in the Balinese Shadow Puppet.

References
[1] S. Purwanto, “Pendidikan Nilai dalam Pagelan Wayang Kulit,” Ta’allum: Jurnal Pendidikan Islam, vol. 6, no. 1, Mar. 2018.
[2] Zhang, Amy & Ballas, Nicolas & Pineau, Joelle. (2018). A Dissection of Overfitting and Generalization in Continuous Reinforcement Learning.
[3] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, “Data augmentation for face recognition,” Neurocomputing, vol. 230, pp. 184–196, Mar. 2017.
[4] Meiyin Wu and Li Chen, “Image recognition based on deep learning,” in 2015 Chinese Automation Congress (CAC), 2015.
[5] K. Wang and A. Kumar, “Cross-spectral iris recognition using CNN and supervised discrete hashing,” Pattern Recognition, vol. 86, pp. 85–98, Feb. 2019.
[6] W. Tao, M. C. Leu, and Z. Yin, “American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion,” Engineering Applications of Artificial Intelligence, vol. 76, pp. 202–213, Nov. 2018.
[7] D. Wang, J. Tang, W. Zhu, H. Li, J. Xin, and D. He, “Dairy goat detection based on Faster R-CNN from surveillance video,” Computers and Electronics in Agriculture, vol. 154, pp. 443–449, Nov. 2018.
[8] I. B. K. Sudiatmika, Pranowo, and Suyoto, “Indonesian Traditional Shadow Puppet Image Classification: A Deep Learning Approach,” in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), 2018.
[9] A. Kurnianto and F. Limano, “Visual representation of character of wayang kulit purwa in the wayang-based games: Case studies of Kurusetra and Mahabarat warrior games,” in 2016 1st International Conference on Game, Game Art, and Gamification (ICGGAG), 2016.
[10] V. Couteaux et al., “Automatic knee meniscus tear detection and orientation classification with Mask-RCNN,” Diagnostic and Interventional Imaging, vol. 100, no. 4, pp. 235–242, Apr. 2019.
[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018.
[12] N. M. Firdaus, D. Chahyati, and M. I. Fanany, “Tourist Attractions Classification using ResNet,” in 2018 International Conference on Advanced Computer Science and Information Systems (ICACISIS), 2018.
[13] Y. P. Chen, Y. Li, and G. Wang, “An Enhanced Region Proposal Network for object detection using deep learning method,” PLOS ONE, vol. 13, no. 9, p. e0203897, Sep. 2018.

[14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[15] M. D Bloice, C. Stocker, and A. Holzinger, “Augmentor: An Image Augmentation Library for Machine Learning,” The Journal of Open Source Software, vol. 2, no. 19, p. 432, Nov. 2017.