2024 한국컴퓨터종합학술대회 (KCC 2024)

Watershed 알고리즘을 이용한 딥러닝 기반 악보 내 음악기호 탐지 방법

Method of Deep Learning-Based Music Symbol Detection Using Watershed Algorithm

Suk-Kyun Hong[O1], Jeoung-Ryeol Lee[1], Kyeong-Min Kim[1], Jun-Young Oh[2], Ji-Hoon Jeong[1*]

[1] School of Computer Science, Chungbuk National University
[2] Department of Computer Science, Chungbuk National University
{goob5748, dlwjdfuf99, kkmlouis, jy.oh, jh.jeong}@chungbuk.ac.kr

1. Introduction

■ **Motivation**

- Even with the same music, scores can be arranged in various ways by the arranger based on factors such as difficulty and the number of performers.
- If a performer wants to convert a score into electronic score, a person must directly read the score in the image or PDF format and enter music symbols (tone, note, rest, etc.)
- Performer should enter hundreds of types of music symbols one by one through the electronic score editing program. **The average number of music symbols per page of the used score dataset is 243 [1]**, and it takes a lot of time and effort to enter each music symbol one by one.

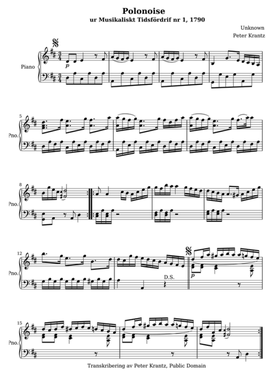
■ **Objectives**

- In this study, we propose a method to reduce the labor required for a performer by directly converting into an electronic score through a deep learning-based music symbol recognition method using the watershed algorithm [2].

2. Materials & Method

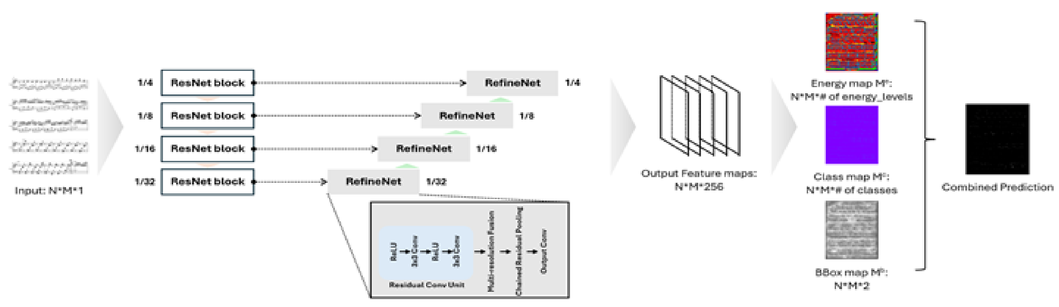
■ **Dataset - DeepScoresV2**

- Classes : 135 music symbols
- Size
  - 255,385 musical score images
  - Total 151M music symbols
- Format
  - Musical score images
  - JSON-type annotation file.



■ **Model Architecture**

- Watershed algorithm
  - Region-based image segmentation methods
  - Distinguishes objects by comparing them with neighboring pixel values based on pixel values.
  - Performs well when dividing images by objects that are touched or overlapped by another objects.
- Deep Watershed Detector
  - Consists of RefineNet, which uses ResNet as a backbone inside and has three 1x1 convolution layers as the ouput layer.
  - The energy map, class map, and bounding box map are extracted as the output of the feature vector extracted from the image through the RefineNet through the 1x1 convolution layer.



■ **Experiment Method**

- Data : 956 musical score data randomly selected from entire dataset
- Preprocessing : Train:Validation:Test = 8:2:2
- Method : Energy, Class, BBox, Combined prediction
- Iteration : Total 2200 iterations

3. Experimental Result & Discussion

We analyze the results through the three comparison groups that conducted the learning by adjusting the hyperparameter that can be changed.

■ **Independent Variable:** Backbone, Optimizer
■ **Controlled Variable:** Learning Rate, Iteration

■ **Three comparison groups**

- Experiment A
- Experiment B: The backbone was changed to ResNet-152 from Experiment A.
- Experiment C: The optimizer was changed to Adam from Experiment B.

Table Ⅰ. Three comparison groups that were trained by adjusting Hyperparameter

	Backbone	Optimizer	Learning Rate	Iteration
A	ResNet-101	RMSProp	0.0001	2200
B	**ResNet-152**	RMSProp	0.0001	2200
C	ResNet-152	**Adam**	0.0001	2200

■ Model performance metrics : F1 score, Average Precision (AP)

- In this study, AP at 0.5, the ratio of predictions with IoU values of 0.5 or more per class

■ **Result**

- **Highest Precision : Experiment C**
- **Highest Recall : Experiment A**
- **Highest F1 Score : Experiment B**

Table Ⅱ. Detecting performance of three comparison groups

	Precision	Recall	F1 Score
A	0.000028	**0.000004823**	0.0000082288
B	0.0001	0.00001642	**0.000028526**
C	**0.0002**	0.000003766	0.0000073972

Table Ⅲ. Occurrences number of classes and AP at 0.5

Class	Class	No. Occurrences	AP at 0.5
	timeSig1	35	0.143
	timeSig8	127	0.067
	keySharp	1120	0.048
	rest16th	234	0.048

In the field of detecting musical symbols within a score, there is no significant difference between false and non-detections.

Non-Detection : a situation where a quarter note is not detected
False : a situation where a quarter note is incorrectly detected as an eighth note

The experiment B with the highest F1 score is the optimal model

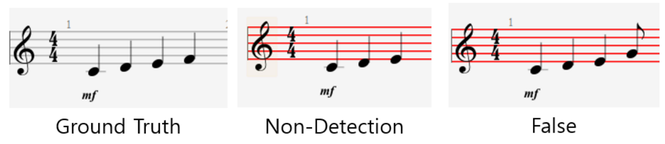


Fig. 1. Images visualizing the ground truth and predicted by model

4. Conclusion

■ Deep learning-based model for automatically recognizing and classifying music symbols within a score image.

■ It is required to improve the performance through expanding the network or tuning other parameters, based on the model that achieved the highest F1 score during experiments.

■ Following the enhancement of performance, our future work can contribute to the digitalization of the music field by reducing the effort required to convert a score image into an electronic score and enabling automated work in order to play a song.

Reference

[1] L. Tuggener, Y.P. Satyawan, A. Pacha, J.Schmidhuber and T.Stadelmann, “The DeepScoresV2 Dataset and Benchmark for Music Object Detection” International Conference on Pattern Recognition, pp. 9198-9195, 2021.

[2] Kornilov, Anton S., I.V. Safonov, “An Overview of Watershed Algorithm Implementations in Open Source Libraries”Journal of Imaging, 2018.

* This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW (2019-0-01183) supervised by the IITP (Institute of Information & communications Technology Planing & Evaluation) (2019-0-01183), the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (RS-2023-00237203), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (RS-2021-II212068, Artificial Intelligenc Innovation Hub), Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Agriculture and Food Convergence Technologies Program for Research Manpower Development Program funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2024-00398561)