Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 283-290, Kasım 2020
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 283-290, November 2020
Copyright © 2020 EJOSAT
Research Article

Recognizing Musical Notation Using Convolutional Neural Networks

Ahmad Othman[1], Cem Direkoğlu[2*]

[1] Middle East Technical University Northern Cyprus Campus, Centre for Sustainability, Department of Electrical and Electronics Engineering, 99738 Kalkanlı, Güzelyurt, Northern Cyprus, Mersin 10, Turkey (ORCID: 0000-0001-8156-8965)

[2] Middle East Technical University Northern Cyprus Campus, Centre for Sustainability, Department of Electrical and Electronics Engineering, 99738 Kalkanlı, Güzelyurt, Northern Cyprus, Mersin 10, Turkey (ORCID: 0000-0001-7709-4082)

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2020 – 22-24 October 2020)

(**DOI:** 10.31590/ejosat.823266)

ATIF/REFERENCE: Othman, A. & Direkoğlu, C. (2020). Recognizing Musical Notation Using Convolutional Neural Networks. *European Journal of Science and Technology*, (Special Issue), 283-290.

Abstract

Musical scores are the essential of music theory and its development. Musical notation was developed by Greeks around 521 BCE, considering that music was developed a long time ago will will find a gap between new musical technology and old scrpits of music theory since they were written in. However, having music scores in written form has rised various kinds of problems for music information retrieval (MIR). Music notation recognition is a type of optical character recognition (OCR) applications, which allow us to recognize musical scores and convert it to a format that can be editied or played on computer such as musicXML (for page layout). In this paper, we introduce a Convolutional Neural Networks (CNN) based framework for musical notation recognition in images. We use a popular pre-trained CNN network, namely ResNet-101 to extract global features of notation and rest images. Then, a Support Vector Machine (SVM) is employed for training and classification purpose. ResNet-101 is one of the state-of-art pre-trained network for image recognition, ResNet-101 trained with more than a million images. Multiclass SVM classifiers using a fast-linear solver is also very powerful classifier. We also evaluated the proposed approach on a dataset that was derived from Attwenger, P RecordLabel and OMR-dataset, and then labeled manually by music theory. As a result, we can separate notes and rests from each other with an average accuracy of 99.02%. We can also classify five different note types. This is the first time that Resnet-101 and a SVM is combined together to perform musical notation recognition, and results are very promising.

Keywords: Optical music recognition, convolutional neural networks, support vector machine, notation recognition.

Evrişimli Sinir Ağlarını Kullanarak Müzik Notasyonunu Tanıma

Öz

Müzik notaları, müziğin gelişiminde kritik bir rol oynar. Yüzyıllar boyunca müzik, ister bestecisinin el yazması isterse herhangi bir yazılı versiyon olsun, resim biçiminde tutulmuştur. Bununla birlikte, müzik notalarının resim biçiminde arşiv edilmesi, müzik bilgilerinin alınması için birçok zorluğu doğurmuştur. Müzik notası tanıma, MIDI (çalma için) ve musicXML (sayfa düzeni için) gibi, müzik notalarının düzenlenebilecek veya çalınabilecek şekilde tanınmasına izin veren optik karakter tanıma (OCR) uygulamalarından biridir. Bu yazıda, görüntülerde nota tanıma için Evrişimli Sinir Ağları (CNN) tabanlı bir çerçeve öneriyoruz. Not ve dinlenme görüntülerinin genel özelliklerini çıkarmak için, önceden eğitilmiş popüler bir CNN ağı, yani ResNet-101'i kullanıyoruz. Ardından, eğitim ve sınıflandırma amacıyla bir Destek Vektör Makinesi (SVM) kullanılır. ResNet-101, görüntü tanıma için son teknoloji ürünü önceden eğitilmiş ağlardan biridir, ResNet-101 bir milyondan fazla görüntüyle eğitilmiştir. Hızlı bir doğrusal çözücü kullanan çok sınıflı SVM sınıflandırıcılar da çok güçlü bir sınıflandırıcıdır. Çalışmamızı test etmek için, deneyimizde veri seti Attwenger, P RecordLabel ve OMR-veri setinden türetildi ve ardından müzik teorisi ile manuel olarak etiketlendi. Sonuç olarak, notaları ve dinlenmeleri birbirinden %99.02 oranıda doğru bir şekilde ayırabiliriz. Ayrıca beş farklı not türünü sınıflandırabiliriz. Bu çalışmada, Resnet-101 ve bir SVM'in ile kez birleştirilerek müzik notası tanıma için bir araya getirilmiştir ve sonuçlar çok umut vericidir.

Anahtar Kelimeler: Optik müzik tanıma, evrişimli sinir ağları, destek vektör makinesi, nota tanıma.

[*] Corresponding Author: Middle East Technical University Northern Cyprus Campus, Centre for Sustainability, Department of Electrical and Electronics Engineering, Kalkanlı, Güzelyurt, Northern Cyprus, Mersin 10, Turkey. ORCID: 0000-0001-7709-4082, cemdir@metu.edu.tr

1. Introduction

Optical Music Recognition (OMR) is mainly musical notation recognition in images. OMR is a difficult task and canbe considered as application of optical character recognition (OCR) (Casey et al., 2008; Bainbridge and Bell, 2001). Generally, a conventional OMR system contains four basic steps as follows (Good and Actor, 2003): (1) Pre-processing of images, (2) recognition of musical symbols, (3) transformation of musical notation, and (4) encoding of musical notation. On the other hand, modern techniques are based on Deep Learning. Two main approaches are used for Optical Music Recognition (OMR): Approaches that are based on object detection and approaches that are based on sequential recognition. Related work is summarized below.

1.1. Approaches based on Object Detection

Due to the rapid developments in computer vision technologies recently, the OCR applications have been positively involved in the CNN applications (LeCun, et al., 2015). Many object detection algorithms appeared and classified into 2 categories: One-stage detection algorithms and two-stage detection algorithms. The extra stage in the second category is basically the regional proposal stage, this stage gives higher accuracy in the two-stage detections but with lower speed compared to the single stage algorithms. YOLO (LeCun, et al., 2015; Rebelo et al., 2012; Redmon et al., 2015), SSD (Liu et al., 2016) and retina-net (Lin et al., 2017) are examples for the one-stage algorithms, while Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), and R-FCN (Dai et al., 2016) are examples on the two-stage algorithms. (Pacha et al., 2018) offered a framework for the detection of music objects. In his work, he used three different object recognition methodes (u-net (Ronneberger et al., 2015), Faster R-CNN and RetinaNet) on three different datasets (deep scores (Tuggener et al., 2018), MUSCIMA++ (Hajič and Pecina, 2017) and capital) to recognize musical notes. (Hajič et al., 2018) also uses two steps to detect the notes. In the initial stage, score images are input that are segmented as binary images, which lead into classification problems composed of a set of binary pixels. The experiments were carried out on MUSCIMA++ dataset. Their results are presented as symbols according to f-score measures. (Tuggener et al., 2018B) projected the Deep Watershed Detector. They used ResNets (He et al., 2015) to find the density map and use it to predict location class of each symbol this method let us use an entire image without cropping each staff. It was working well on small symbols, but it raised a problems such as inaccurate bounding boces and unrecognizing rare classes.

1.2. Approaches based on Sequential Recognition

In this approach instead of training a single separate musical symbol, the whole line of music paper is translated at the same time. Convolutional neural networks (CNN) and recurrent neural networks (RNN) are most used in a such approchies, CNNs can extract features and combine them with the structure of an image, therefore convolutional architecture has become a famous among objects recognition. Another sequence-to-sequence model is RNN which is often used in machine translation (Cho, et al., 2014; Sutskever et al., 2014). (Van der Wel and Ullrich, 2017) presented the Convolutional Sequence-to-Sequence network, which utilizes CNN in order to convert the input into series of vectors. Then RNN encodes the vector sequence into a fixed-size illustration. Finally, it uses the RNN to decode the fixed size into a tag sequence represents the output. The experiment gived good results and was able to demonstrate the pitch, note accuracy, and time. (Calvo-Zaragoza and Rizo, 2018) extracted features from printed music sheet using a CNN and feed them into a Recurrent Neural Network. They faced a problem with misalignment of different scores image labels, but they were able to solve it using a Connectionist Temporal Classification (CTC) loss function.

Proposed approach is explained in Section 2. Section 3 clarifies musical notation basics. Section 4 talk about the dataset and implementations. Evaluations and results are discussed in Section 5. Section 6 is conclusions.

2. Proposed Approach

We use ResNet-101 (ResNet, 2015) ResNet-101 is a pre-trained CNN model that is uses image feature extraction. Note and rest images are input to ResNet-101 for feature extraction. Our network contains up to 152 layers. ResNet-101 was generated with residual representation-based learning. This network has an atvantage of skipping connection that can fit the input from the previous layer into the next layer without any need of modification from the input. Skip connection yields a deeper network. ResNet became the Winner of ILSVRC 2015 in image recognition and classification tasks. In addition to that it also won MS-COCO 2015 image detection, and segmentation challenge.

Figure 1 shows ResNet101 architecture (ResNet, 2015). It consists of there parts: (1) VGG-19, (2) 34-layer plain network and (3) 34-layer residual network. (1) The VGG-19 is improved version of the VGG-16. It is a CNN with 19layers. It was built by stacking convolutions (2) 34-layer plain network was inspired from VGG-19, which is a deeper network that contains more convolutions layers. (3) Finally, 34-layer residual network (ResNet) is a deep network that includes additional skip/shortcut connections. ResNet contains three kinds of shortcuts/skips connections: (A) shortcut achives identity mapping, with extra zero fillings for increasing dimensions. (B) The projection shortcut performes increasing dimensions only, the other shortcuts are identity. Which mean we need extra parameters. (C) All shortcuts are projections. Thus, we need more extra parameters. CNN network of ResNet-101 is trained using a million images from the ImageNet database. As a result, the network is rich with features from a range of images. After feature extraction with Resnet101, we train multiclass SVM classifiers using a fast-linear solver. Then, test notation images are input to SVM for classification.

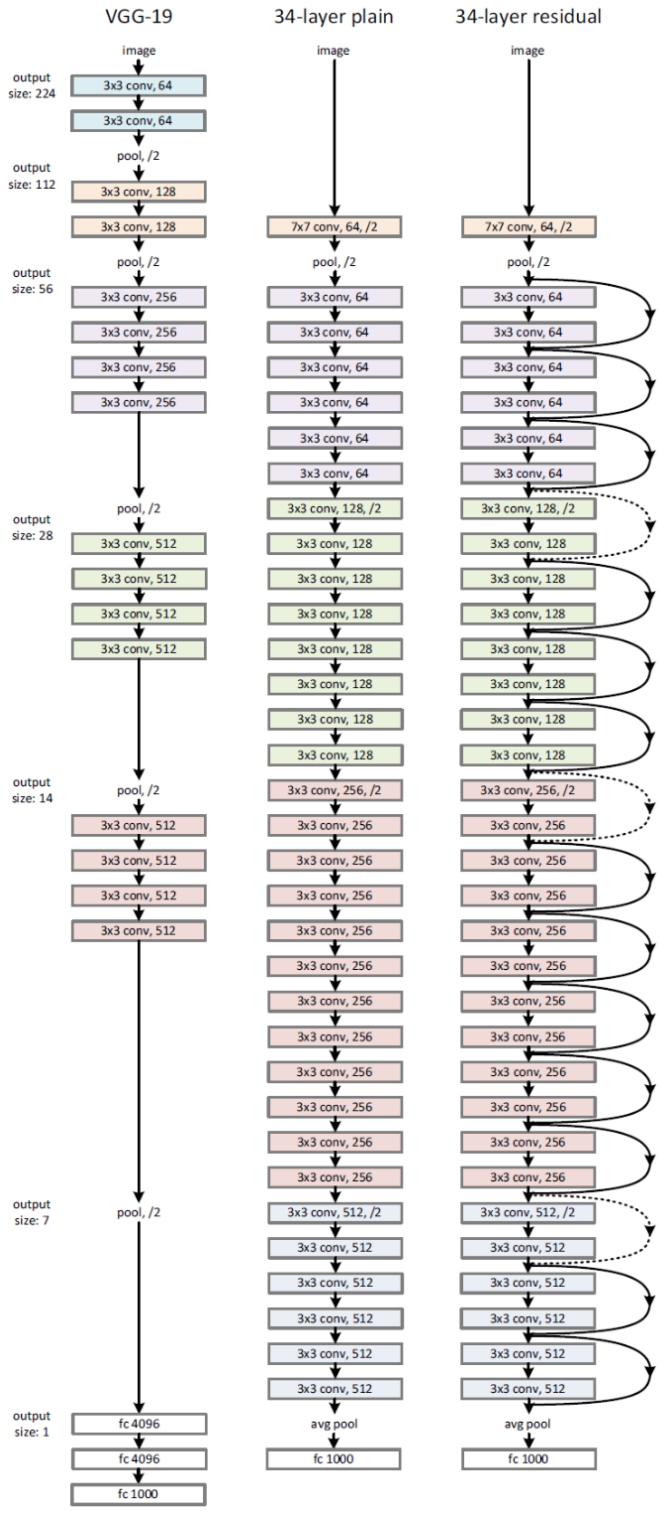

Figure 1. ResNet101 Architecture (ResNet, 2015).

3. Musical Notation Basics

One of the common systems that are used to identifying a musical score is Western musical notation and a complex one too. It can describe uncountable aspects of any music piece, understand such system is a field of study by its own, and describing it comprehensively would take time and a lot of resources which would go outside the scope of this paper. Therefore, we will only cover the aspects that is related to the classification process. In western musical notation staff represents the five horizontal lines and the four spaces between, each one of them represents a different pitch. Notes represents a musical sound and rests represents the absence of sound. Pitch represents the frequency of vibration produced by a sound wave. The vertical placement of a note on the staff shows the pitch of that note.

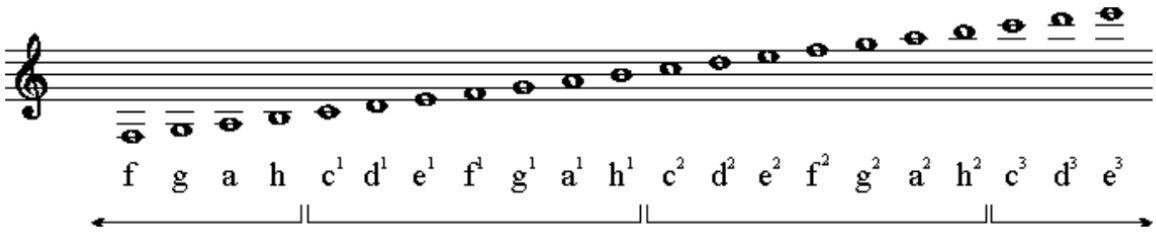

Figure 2. Musical notes with different pitches.

There are many systems to define a tone. In the work we used Helmholtz designation where the Middle C (261.6 Hz) is represented by the name c1. However, we should note that the English system call the same tone C4 which is used in MusicXML. The duration of a note or a rest is represented by its symbol as shown in Figure 3 and Figure 4 below. As noted, we have seven durations for a note and six for a rest combing it with the pitch (vertical placement) we would be able to read that score and play it.

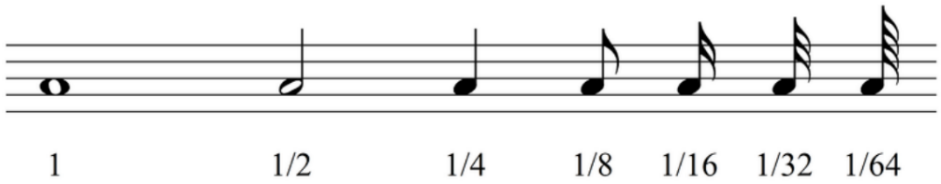

Figure 3. Note durations expressed as fractions.

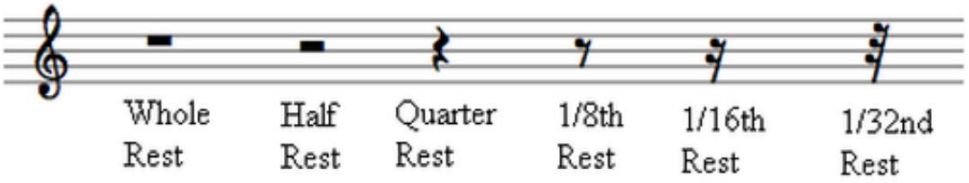

Figure 4. Rest durations.

4. Dataset Preparation and Implementations

The dataset in our experiment was taken from Attwenger, P RecordLabel (Attwenger, 2015) and OMR-dataset and then labeled manually by music theory. In our approach, first pre-processing is applied to the dataset. Then, we handle distortions. Finally, the network is built using ResNet101 and trained for the recognition of musical notation. This section explains the details.

4.1. Image Pre-processing

OpenCV Python script was used to normalize the images. A size of 75×75 pixels was applied. Skew Correction, Thinning and Skeletonization and noise removal are also applied to the dataset.

4.2. Distortions

When we are dealing with image recognition, it is more robust resultings when applying distortions to images before bulding the classifier. However, distortions should be naturally happened due to normal cases. here are the three types of distortions we considered in our data:

One technique of achieving more robust recognition is to apply distortions to the images before using them. However, those distortions should be like to the ones that also happen naturally in the data. There are three types of distortions were considered in musical data sheets:

1. Rotation: In such documents rotation due to scaning is common. As a result, some notes near edges may not appear.

2. Skewing: such problem rises when the lid of the scanner is not closed correctly. For example, if a book was scanned the edge reagions of the papper could suffer from this distortion.

3. Noise: such document is more likely to have noise raise from scaning devices, cameras, lights, etc. In our example, it may be visible as ink blisters in white spaces and vice versa. In addition to that the scanners could sometimes make wrong configuration,

resulting in gray or black areas that should be white, for example scanning a page with cut off edges. We applied all distortions to the images in the training set.

Rotation was applied through affine transformation. “salt and pepper” noise with variable degrees of white and black was implemented. Skewing was approached through shearing.

Layer (type)	Output Shape	Param #
conv2d_70 (Conv2D)	(None, 73, 73, 32)	896
activation_63 (Activation)	(None, 73, 73, 32)	0
max_pooling2d_63 (MaxPooling	(None, 36, 36, 32)	0
batch_normalization_101 (Bat	(None, 36, 36, 32)	128
dropout_101 (Dropout)	(None, 36, 36, 32)	0
conv2d_71 (Conv2D)	(None, 34, 34, 32)	9248
activation_64 (Activation)	(None, 34, 34, 32)	0
max_pooling2d_64 (MaxPooling	(None, 17, 17, 32)	0
batch_normalization_102 (Bat	(None, 17, 17, 32)	128
dropout_102 (Dropout)	(None, 17, 17, 32)	0
conv2d_72 (Conv2D)	(None, 15, 15, 32)	9248
activation_65 (Activation)	(None, 15, 15, 32)	0
max_pooling2d_65 (MaxPooling	(None, 7, 7, 32)	0
batch_normalization_103 (Bat	(None, 7, 7, 32)	128
dropout_103 (Dropout)	(None, 7, 7, 32)	0
conv2d_72 (Conv2D)	(None, 15, 15, 32)	9248
activation_65 (Activation)	(None, 15, 15, 32)	0
max_pooling2d_65 (MaxPooling	(None, 7, 7, 32)	0
batch_normalization_103 (Bat	(None, 7, 7, 32)	128
dropout_103 (Dropout)	(None, 7, 7, 32)	0
conv2d_73 (Conv2D)	(None, 5, 5, 32)	9248
activation_66 (Activation)	(None, 5, 5, 32)	0
max_pooling2d_66 (MaxPooling	(None, 2, 2, 32)	0
batch_normalization_104 (Bat	(None, 2, 2, 32)	128
dropout_104 (Dropout)	(None, 2, 2, 32)	0
flatten_11 (Flatten)	(None, 128)	0
dense_50 (Dense)	(None, 128)	16512
batch_normalization_105 (Bat	(None, 128)	512
dropout_105 (Dropout)	(None, 128)	0
dense_51 (Dense)	(None, 82)	10578

Total params: 56,754
Trainable params: 56,242
Non-trainable params: 512

Figure 5. Layers of the Network.

4.3. Building the CNN Network

The network was built by using Resnet101 and by CNN python code using OpenCV and TensorFlow. TensorFlow is preferred because of high speed computation and availability of open-source libraries. It has flexible architecture that enables processing on various platforms such as CPUs, TPUs, GPUs, and on different devices such as computers, servers, mobile devices, embedded systems and edge devices. ResNet101 was developed by a team of engineers and researchers from Google Brain as a part of Google’s Artificial Intelligence (AI) organization. As a result, Google Brain team had a strong contribution to the fields of machine learning and

deep learning. Figure 5 shows one of the ResNet101 models used in our work. We set 'ObservationsIn' to 'columns' to match the arrangements used for training features then train multiclass SVM classifiers using a fast-linear solver.

5. Evaluation and Results

The evaluation setup was as follows. We extracted notes features from ResNet101 and trained them with SVM. The input was single note image with 80:20 train to test split. But we are able to write a code that can split the paper data sheet into single tones and recognize then reconstruct the paper with computer readable notes. Total data was 44,124 images that consisting of 37,500 notes and 6,624 rests split as 80:20 in python code with random selection.

5.1. Results of Note and Rest Classification (Two Classes)

Tables below show the results comes from evaluation test sets for each classification. Table 1 shows classification results for two classes, notes and rests. We used 80% training and 20% testing of the dataset. Results are very accurate, in particular notes are classified with an accuracy of 99.15% and rests are classified with an accuracy of 98.29%. Figure 6 shows the confusion matrix of the results.

Table 1. Experimental results for two classes; note and rest

Class	**Accuracy**
Note	99.15%
rest	98.29%
***Total accuracy* = 99.02%**	

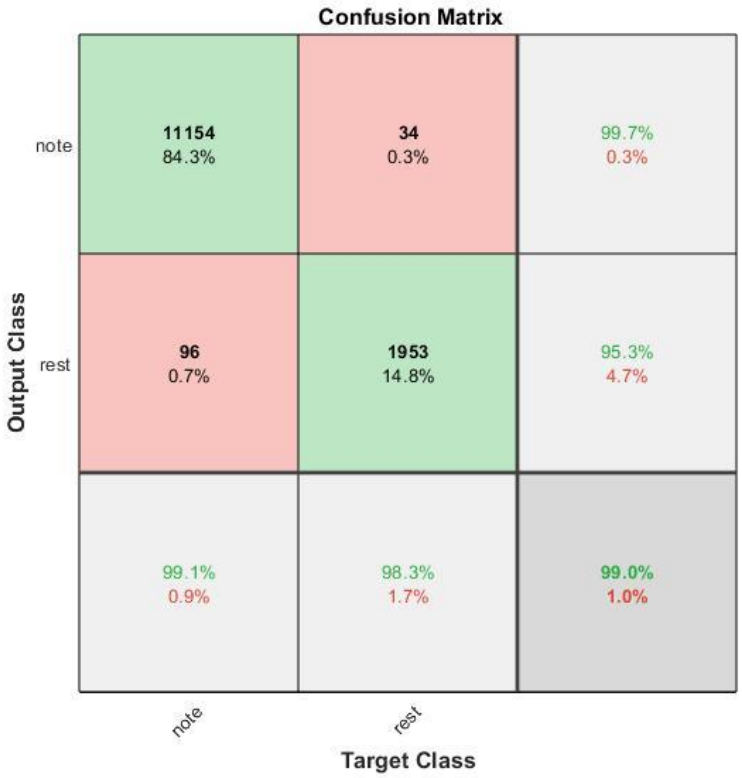

Figure 6. Confusion matrix for two class classification.

5.2. Results of Five Note Classes Classification

Table 2 shows classification results for five note classes. We used 80% training and 20% testing of the dataset. Results show that half, quarter and sixteenth have higher accuracy than others. In particular, half note has a classification accuracy of 99.57%, quarter note has an accuracy of 93.58% and sixteenth has an accuracy of 92.15%. Among the all classes, eighth has the lowest accuracy since it is confused with the sixteenth as shown by the confusion matrix in Figure 7. In overall, the total accuracy is 81.47% in our dataset. The confusion matrix of all five classes are shown in Figure 7.

Table 2. Experimental results for five notation classes

Class	Accuracy
whole	81.25%
half	99.57%
quarter	93.58%
eighth	40.80%
sixteenth	92.15%
Total accuracy = 81.47%	

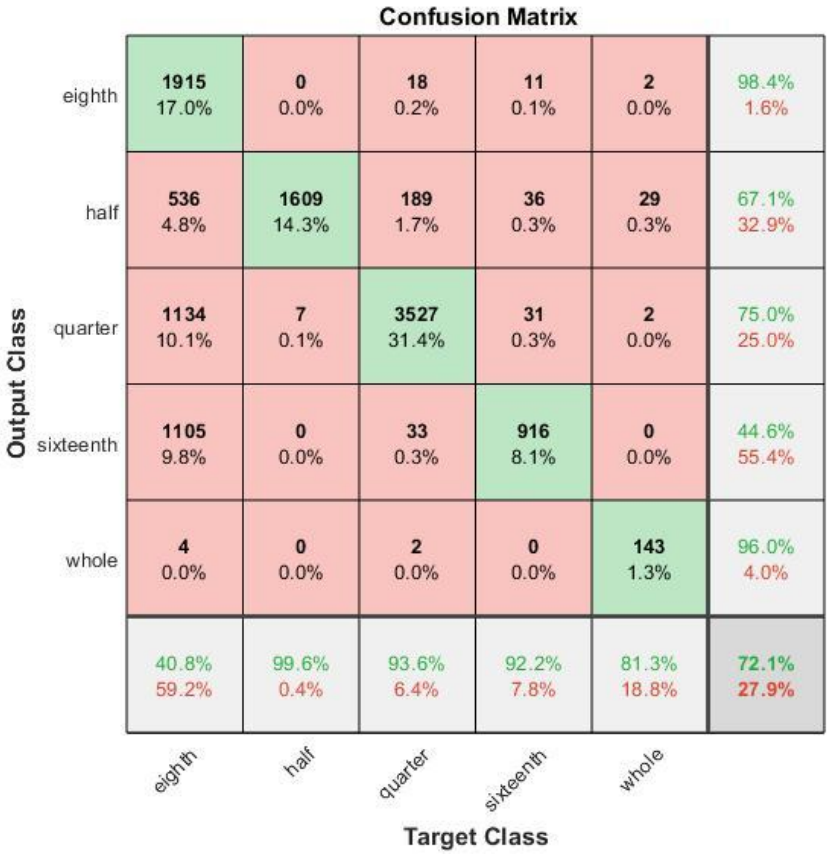

Figure 7. Confusion matrix for five notes classification.

6. Conclusions

We have presented a new approach for musical notation recognition in images. ResNet-101 is employed for rich feature extraction of notations, and then multiclass SVM is used for training and classification. We described the basic concept of musical notations. We explained how to prepare a realistic dataset for musical notations. For evaluation, first we performed classification for rest and note images. Second, we performed classification for five different notation types. Results are very promising. This is the first time that Resnet-101 and a SVM is combined together to perform musical notation recognition.

References

Attwenger, P. (2015). RecordLabel, http://homepage. univie.ac.at/a1200595/recordlabel/

Bainbridge, D., & Bell, T. (2001). The challenge of optical music recognition. Comput. Humanit, 35, 95–121, doi:10.1023/A:1002485918032.

Calvo-Zaragoza, J., & Rizo, D. (2018). End-to-End Neural Optical Music Recognition of Monophonic Scores, Appl. Sci, 8, 606, doi:10.3390/app8040606.

Casey, M., & Veltkamp, R., & Goto, M., & Leman, M., & Rhodes, C., & Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. In Proc. of IEEE, 668–696, doi:10.1109/JPROC.2008.916370.

Cho, K., & van Merrienboer, B., & Gulcehre, C., & Bahdanau, D., & Bougares, F., & Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, arXiv 2014, arXiv:1406.1078

Dai, J., & Li, Y., & He, K., & Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks, arXiv 2016, arXiv:1605.06409.

Girshick, R. (2015). Fast R-CNN. arXiv 2015, arXiv:1504.08083.

Good, M., & Actor, G. (2003). Using MusicXML for file interchange. International Conference on WEB Delivering of Music, 15–17, doi:10.1109/WDM.2003.1233890.

Hajiˇc, J., & Pecina, P. (2017). The MUSCIMA++ Dataset for Handwritten Optical Music Recognition. IAPR International Conference on Document Analysis and Recognition (ICDAR), 39–46, doi:10.1109/ICDAR.2017.16.

Hajiˇc, J., & Dorfer, M., & Widmer, G., Pecina, P. (2018). Towards Full-Pipeline Handwritten OMR with Musical Symbol Detection by U-Nets. International Society for Music Information Retrieval Conference, 23–27.

He, K., & Zhang, X., & Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition, arXiv 2015, arXiv:1512.03385.

LeCun, Y., & Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444, doi:10.1038/nature14539.

Lin, T.Y., & Goyal, P., & Girshick, R., & He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection, arXiv 2017, arXiv:1708.02002.

Liu, W., & Anguelov, D., & Erhan, D., & Szegedy, C., & Reed, S., & Fu, C.Y., & Berg, A.C. (2016). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision; Springer: Cham, Switzerland, 21–37, doi:10.1007/978-3-319-46448-0_2.

Pacha, A., & Hajiˇc, J., & Calvo-Zaragoza, J. (2018). A Baseline for General Music Object Detection with Deep Learning, Appl. Sci., 8, 1488, doi:10.3390/app8091488.

Rebelo, A., & Fujinaga, I., & Paszkiewicz, F., & Marcal, A.R.S., & Guedes, C., & Cardoso, J.S. (2012). Optical music recognition: State-of-the-art and open issues. Int. J. Multimed. Inf. Retr, 1, 173–190, doi:10.1007/s13735-012-0004-6.

Redmon, J., & Divvala, S., & Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection, arXiv 2015, arXiv:1506.02640.

Ren, S., & He, K., & Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv 2015, arXiv:1506.01497.

ResNet, (2015). https://towardsdatascience.com/review-resnet-winner-of-ilsvrc-2015-image-classification-localization-detection-e39402bfa5d8

Ronneberger, O., & Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv 2015, arXiv:1505.04597.

Sutskever, I., & Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. In Advances in Neural Information Processing Systems 27; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds., 3104–3112.

Tuggener, L., & Elezi, I., & Schmidhuber, J., & Pelillo, M., & Stadelmann, T. (2018). DeepScores—A Dataset for Segmentation, Detection and Classification of Tiny Objects. arXiv 2018, arXiv:1804.00525.

Tuggener, L., & Elezi, I., & Schmidhuber, J., & Stadelmann, T. (2018B). Deep Watershed Detector for Music Object Recognition, arXiv 2018, arXiv:1805.10548.

Van der Wel, E., & Ullrich, K. (2017). Optical Music Recognition with Convolutional Sequence-to-Sequence Models, arXiv 2017, arXiv:1707.04877.