Handwritten Music Symbol Classification Using Deep Convolutional Neural Networks

Sangkuk Lee*, Sung Joon Son*, Jiyong Oh† and Nojun Kwak*
*Graduate School of Convergence Science and Technology, Seoul National University, Korea
Email: {sangkuklee, sjson718, nojunk}@snu.ac.kr
Homepage: http://mipal.snu.ac.kr
†Electronics and Telecommunications Research Institue (ETRI), Korea
Email: jiyongoh@etri.re.kr

Abstract—**In this paper, we utilize deep Convolutional Neural Networks (CNNs) [1] to classify handwritten music symbols in HOMUS [2] data set. HOMUS data set is made up of various types of strokes which contain time information and it is expected that online techniques are more appropriate for classification. However, experimental results show that CNN which does not use time information achieved classification accuracy around 94.6% which is way higher than 82% of dynamic time warping (DTW) [3], the prior state-of-the-art online technique. Finally, we achieved the best accuracy around 95.6% with the ensemble of CNNs.**

I. Introduction

Information technology has been drastically advanced in the last two decades so that digitization is deeply related to various fields of music such as preservation, duplication and distribution [4]. Also in composition and songwriting areas, a lot of music experts use digital devices and computer softwares nowadays. Most of those programs provide the function of making and editing music scores by musical instruments and computer input devices. Nonetheless, pen and paper still occupy an important position among various composition tools. For this reason, it is required to automatically recognize handwritten musical symbols. In particular, the recent wide use of pen-based digital devices such as smartphones and tablet PCs increases the demand more and more.

However, since 2000s, only a small number of related studies such as [5], [6], [7] have been presented in the literature. Recently, a large data set for handwritten music symbol recognition, which is named HOMUS, was released and typical methods for the problem were evaluated on the data set in [2]. Although the authors of the paper verified both online and offline techniques, they did not consider deep network-based algorithms. However, deep convolutional neural networks (CNNs) have been shown to be very effective in many applications such as large scale image classification [8].

In this paper, we utilize deep CNNs to classify handwritten music symbols in HOMUS data set. To our best knowledge, this is the first trial to use CNN to recognize music notations. Experimental results show that a CNN can reduce classification error by more than 10% compared to dynamic time warping (DTW) [3], which provided the best performance in [2]. In addition, ensemble of CNNs leads another 1% of accuracy on HOMUS data set compared to the single CNN. This agrees with the fact that deep CNN is a very useful tool for image classification.

This paper is organized as follow. Related works are briefly reviewed in Sec. II. In Section III, the database and methods are described. Then, experimental results are shown in Sec. IV and the paper is concluded in Sec. V

II. RELATED WORK

There are mainly two types of music symbol data, the strokes and the images. The former data which contains time information of the strokes are normally collected by digital pens in smart devices such as smartphones and tablet PCs. On the other hand, the latter is relatively easier to acquire but does not contain any time information, thus is considered more challenging to classify.

While there have been many techniques suggested for music symbol classification, different methods are needed for classification of online stroke data and offline image data. The online techniques try to classify musical symbols by utilizing time information made by the pen strokes. In this case, the number of points generated by a stroke vary according to the pace of user's pen stroke. For this kind of stroke data sets, DTW [3] [9] [10] and hidden Markov models (HMM) [11] are used as online techniques.

On the other hand, offline techniques are for the image-type music symbol data which include general classification algorithms to classify an image into a category. The well-known classifiers such as k-nearest neighbor (k-NN) [4], multi-layer Perceptron (MLP) [5], support vector machines (SVM) [12] [13], and ensemble Neural Networks [14] can all be used to classify an image-type music symbol.

III. Data and Methods

In this paper, deep CNNs and ensemble of CNNs are compared with more shallow networks such as MLP and ensemble MLP. We also show that CNNs [1] not only perform well for the image-type data as in [15] or [16], but also are highly efficient for the stroke-type data compared to the state-of-the-art DTW algorithm using time information.

This work was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2015.

A. Database

For all classification experiments performed in this paper, we use the Handwritten Online Musical Symbols (HOMUS) dataset [2]. The HOMUS dataset was collected from 100 musicians whose data are divided in the corresponding 100 folders. There are 32 classes and 15,200 sample images in total. In [2], classification performance was measured using 10-fold cross validation. In other words, among the 100 folders, 90 folders are used for training while the other 10 folders are used as a test set. In this paper, we also use the same strategy to report the performance of each method.

HOMUS data set is basically for online techniques because each sample of a dataset contains the strokes consisting of a set of points. For visualization, we transformed strokes into images using MATLAB codes. Figure 1 shows some example strokes. These are sample datasets of 3 different musicians for 32 classes. The images in Figure 1 are obtained by firstly connecting original points of a stroke with a line of one pixel wide, then applying a square dilation kernel of 4×4 pixels. As can be seen in this figure, sample images include a considerable amount of deviation.

B. Compared Methods

Dynamic time warping (DTW) is a technique for measuring the dissimilarity between two timed samples which may be of different durations. Simply, it is composed of Euclidean distance between two points. The details can be found in [3] [9] [10].

Multi-layer perceptron (MLP) is the most typical type of conventional neural networks. MLP typically have one or two hidden layers. This kind of networks can be trained with the backpropagation algorithm [17].

Ensemble neural network is made up of several neural networks. Each network is trained independently and once a test sample is inserted, the outputs of different networks are merged to predict the final class. Typically, majority vote is used as a combination method. More details are in [14].

Convolutional neural network (CNN) is an algorithm for learning by using a larger and deeper neural network architecture [1]. CNN has achieved great success for various image-based machine learning problems [18]. Because music symbol is a kind of image, we expect that CNN will be successfully applied to music symbol classification, too. In this paper, we implemented CNN architecture of CifarNet [19], AlexNet[15], and GoogleNet[16].

CifarNet uses 3 layers with input size of 32×32×3. In our application, because the stroke image is binary, we modified the network with input size of 32×32. In addition, because various input sizes are tested for CifarNet, the tunable parameters such as number of padding, convolution filter size and strides are tuned appropriately as will be described in Section IV. Alexnet and GoogleNet use 8 layers with input size of 227×227×3 and 22 layers with input size of 224×224×3, respectively. In the same manner as CifarNet, the parameters are tuned appropriately.

In this paper, as an objective function of MLPs and CNNs, cross entropy which is based on softmax probability is used with regularization as follows.

$$J(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\mathbf{1}\{y^{(i)}=j\}\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{K} e^{\theta_l^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{K}\sum_{j=0}^{n}\theta_{ij}^2$$

Here, θ is the parameter to be learned, $(x^{(i)}, y^{(i)})$ is the i-th training sample with input $x^{(i)}$ and target $y^{(i)}$, n and K are the number of samples and classes respectively, and $\mathbf{1}(\cdot)$ denotes the indicator function.

The regulation factor (λ) was obtained by parameter tuning through many comparative experiments on other data sets. Additionally, we verified that activation function in the form of Rectified linear unit (ReLU) shows better performance than sigmoid through many comparative experiments on other data sets. As a result, we use ReLU and set $\lambda = 0.35$ in all the experiments with MLP. For the experiments with CNNs, the default value of λ for each network was used, which are 0.004, 0.0005, and 0.0002 for CifarNet, AlexNet, and GoogleNet, respectively.

IV. EXPERIMENTAL RESULTS

A. Conventional Methods

We studied prior state-of-the-art techniques for HOMUS data and executed comparative experiments. DTW is used for an online classification, while MLP, ensemble neural networks are used as offline classifiers.

1) DTW: For comparison, we performed 10 fold cross validation for HOMUS data set. The classification accuracies on the test data of the 10 different cross validation trials are in between the lowest accuracy of 77.83% and highest accuracy of 85.46%. The average accuracy is 82.73% Since we implemented DTW similar to that of [2], the mean accuracy of our implementation of DTW is very similar to that of [2] which is around 82%.

2) MLP: In [2], DTW performed the best, while MLP showed relatively poor accuracy of around 62%. But according to the recent researches, there are many ways to achieve better performance in learning neural networks [18], [20]. Therefore, as a baseline of CNN, we implemented new MLP networks for HOMUS dataset by changing the size of the input image (square, thin, and fat), number of layers (1 and 2), number of hidden nodes (1000, 2000, and 3000) based on [18] and [20]. We also used different sizes of dilation kernel in transforming a stroke to an image (1x1 to 4x4). The detailed numbers can be found in Table I. For all the experiments, ReLU [21] was used with regularization factor of 0.35. The performances in the table are the averages of 10 fold cross validation.

As a result, when we use 60×30 input size (thin), 4×4 dilation kernel, 1 layers with 3,000 nodes, the highest mean accuracy of 79.43% is obtained. This is an improvement of almost 20% compared to the result of MLP in [2]. Furthermore, this performance is comparable to that of DTW. If we can run a network having bigger input size or more nodes and layers, it can be expected that the accuracy would even more

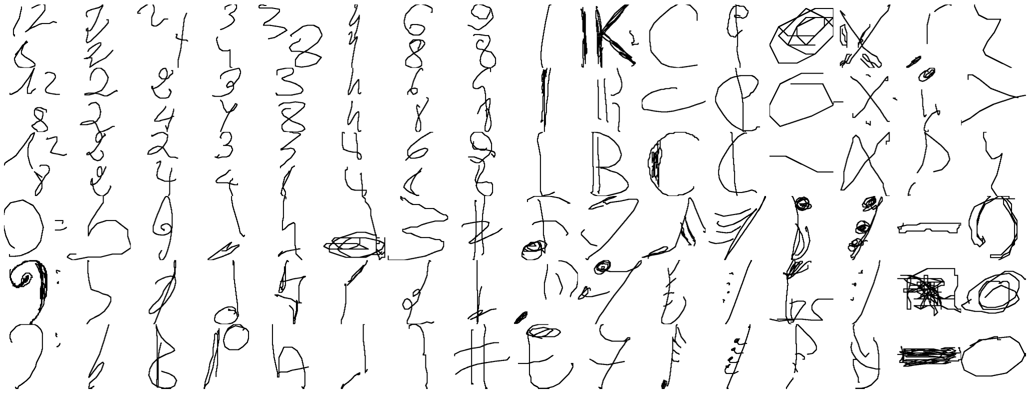

Fig. 1. Examples of images in HOMUS data set.

TABLE I
MLP - EVALUATION WITH CHANGING ARCHITECTURES

Experiment	Input Size	Number of Nodes	Dilation Kernel	Accuracy
Image Size 1 (Square)	28 × 28	1000	3 × 3	0.7638
	40 × 40	1000	3 × 3	0.7782
	60 × 60	1000	3 × 3	0.7757
Image Size 2 (Thin)	35 × 20	1000	3 × 3	0.7815
	60 × 30	1000	3 × 3	**0.7817**
Image Size 3 (Fat)	20 × 35	1000	3 × 3	0.7153
	30 × 60	1000	3 × 3	0.7664
Nodes	60 × 30	1000	3 × 3	0.7817
	60 × 30	2000	3 × 3	0.7865
	60 × 30	3000	3 × 3	**0.7878**
Layer	60 × 30	1000	3 × 3	**0.7817**
	60 × 30	1000,1000	3 × 3	0.7726
Dilation Kernel Size	60 × 30	1000	1 × 1	0.6921
	60 × 30	1000	2 × 2	0.7628
	60 × 30	1000	3 × 3	0.7817
	60 × 30	1000	4 × 4	**0.7918**
best-MLP	**60 × 30**	**3000**	**4 × 4**	**0.7943**

TABLE II
ENSEMBLE NEURAL NETWORK - COMPARISON ON MONO-MLP AND TWO VERSIONS OF ENSEMBLE NETWORKS WITH THREE DIFFERENT COMBINATION METHODS

	Experiment	Accuracy
mono-MLP	MLP1(60 × 60)	0.7757
	MLP2(35 × 20)	0.7832
	MLP3(60 × 30)	0.7828
	MLP4(20 × 35)	0.7153
	MLP5(30 × 60)	0.7664
Ensemble of 3 MLP (1,2,3)	Combination Method 1	0.7995
	Combination Method 2	0.8006
	Combination Method 3	**0.8020**
Ensemble of 5 MLP (1,2,3,4,5)	Combination Method 1	0.7963
	Combination Method 2	0.7947
	Combination Method 3	**0.8049**

increase. This is one of the reasons why we have experimented deep CNN. Also, thin input images result in more successful performance. Perhaps, it is because most of the music symbols have thin shapes rather than fat.

3) Ensemble Neural Network: In this part, we executed experiments using ensemble neural networks with 3 MLPs and 5 MLPs. Three different rules were used for combining different networks. First, we used the classification results of individual MLPs for majority vote. This method was already suggested in [14]. The following two are what are newly suggested in this paper. The second combining method is based on the softmax output (class probability) of individual MLPs. That is, most high softmax output among different networks was chosen as the class of the stroke in question. Lastly, we used the sum of softmax outputs from different MLPs for each class. That is, the highest sum of softmax outputs of an individual class was chosen. For all experiments, we fixed parameters to be $\lambda = 0.35$, 3×3 dilation kernel, 1 layer, 1000 nodes and ReLU. Then, we evaluated the mean of 10-fold cross validation for HOMUS data set. Table II shows the results.

In the table, we can see that the ensemble neural networks show slightly improved performances, especially third method of combination is most successful.

B. Convolutional neural networks

From the result in Section IV-A3, we came to reason that more sophisticated and deeper architecture will perform better. In this chapter, we implemented various CNN networks such as CifarNet, AlexNet, and GoogleNet with properly modified input sizes. CAFFE was used as an implementation tool [22].

TABLE III
COMPARISON OF VARIOUS INPUT SIZES WITH CIFARNET

Network	input size	Filter Size	Padding	Stride	Accuracy
CifarNet	32×32	5x5	2	1	0.8494
CifarNet	64×64	2x2	0	2	0.8807
CifarNet	224×224	7x7	0	7	**0.8976**

TABLE IV
COMPARISON OF DIFFERENT CNNS

Network	input size	Filter Size	Padding	Stride	Accuracy
CifarNet	224×224	7x7	0	7	0.8976
AlexNet	224×224	8x8	0	4	0.9368
GoogleNet	224×224	7x7	3	2	**0.9461**

Firstly, we tested various input image sizes with simple CifarNet. As Table III shows, larger input size results in higher accuracy and input size of 224 × 224 performed best. Then, various networks were tested with fixed input size. The first three rows of Table IV shows that as the depth of the network increases, the performance increases accordingly. Lastly, we tested ensemble of CNNs made up of two different GoogleNets with input image size of 224 × 224 (square) and 224 × 112 (thin), respectively. As a result, ensemble of CNNs achieved best average accuracy of 95.55%. Table V shows that ensemble of networks performs better than the two basic networks in every 10 trials of cross validation. In this experiment, we used the combination rule that chooses the highest sum of softmax outputs of an individual class because this combination method showed the best performance in Section IV-A3.

Figure 2 summarises the experimental results performed in this Section. In Figure 2, 'best-MLP' is the MLP showing the best performance in Table I – 60×30 input size(thin), 0.35 λ, 4×4 dilation kernel, 1 layer, 3000 nodes and ReLU. Also, 'best-ComNN' is the Combined Neural Network showing the best performance in Table II -using 5 MLP and third method for majority vote.

TABLE V
ENSEMBLE OF CNNS

Cross Validation	GoogleNet-1 (224×224)	GoogleNet-2 (224×112)	Ensemble of 1, 2
1-10	0.9283	0.9322	0.9447
11-20	0.9539	0.9533	0.9645
21-30	0.9414	0.9368	0.9434
31-40	0.9467	0.9342	0.9493
41-50	0.9283	0.9211	0.9401
51-60	0.9461	0.9362	0.9553
61-70	0.9434	0.9224	0.9533
71-80	0.9566	0.9566	0.9678
81-90	0.9553	0.9507	0.9671
91-100	0.9605	0.9605	0.9691
Average	0.9461	0.9404	**0.9555**

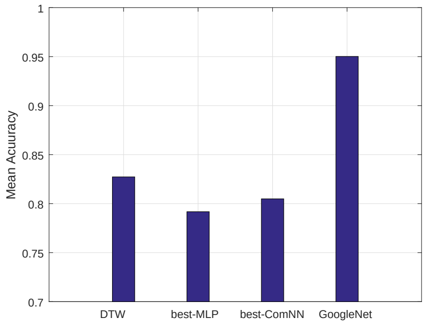

Fig. 2. Performance comparison

V. CONCLUSIONS

In this work, we studied a handwritten music symbol classification problem. HOMUS data set which consists of online stroke information was used. For this dataset, prior state-of-the-art techniques are DTW for online classification and MLP and ensemble neural network are used for offline classification. We implemented these algorithms and compared these with further parameter tuning. In addition, CNN was also used for the classification of this data set.

A comprehensive experiments with different versions of MLPs, CNNs and ensemble networks were executed. As CNN architectures, CifarNet, AlexNet, and GoogleNet were used. Through the experiments, it was shown that larger inputs and deeper architectures are more successful to this dataset. With 224×224 of input size, GoogleNet reached 94.61% average accuracy. Finally, we achieved the best accuracy of 95.55% with the ensemble of CNNs. As a conclusion, we verified that CNN can be very successful in handwritten music symbol classification problems.

For future works, we will study the ways to integrate online and offline techinques for further improvement of the classification accuracy. Because there are many pros and cons between the online techniques and offline techniques, we think that there can be a significant room for improvement in this synergized data features.

REFERENCES

[1] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[2] J. Calvo-Zaragoza and J. Oncina, "Recognition of Pen-Based Music Notation: The HOMUS Dataset," in *2014 22nd International Conference on Pattern Recognition (ICPR)*, Aug 2014, pp. 3038–3043.

[3] Alex Waibel, *Readings in speech recognition*, Morgan Kaufmann, 1990.

[4] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre R. S. Marcal, Carlos Guedes, and Jaime S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.

[5] Susan E. George, “Online Pen-Based Recognition of Music Notation with Artificial Neural Networks,” *Computer Music Journal*, vol. 27, no. 2, pp. 70–79, June 2003.

[6] Y. Mitobe, H. Miyao, and M. Maruyama, “A fast HMM Algorithm Based on Stroke Lengths for On-line Recognition of Handwritten Music Scores,” in *Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 521–526.

[7] Hidetoshi Miyao and Minoru Maruyama, “An online handwritten music symbol recognition system,” *International Journal of Document Analysis and Recognition*, vol. 9, no. 1, pp. 49–58, 2007.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] Marcos Faundez-Zanuy, “On-line signature recognition based on vq-dtw,” *Pattern Recognition*, vol. 40, no. 3, pp. 981–992, 2007.

[10] Bastian Hartmann and Norbert Link, “Gesture recognition with inertial sensors and optimized dtw prototypes,” in *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2102–2109.

[11] Kian Chin Lee, Somnuk Phon-Amnuaisuk, and Choo Yee Ting, “Handwritten music notation recognition using hmma non-gestural approach,” in *Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on*. IEEE, 2010, pp. 255–259.

[12] Vladimir Naumovich Vapnik and Vlamimir Vapnik, *Statistical learning theory*, vol. 1, Wiley New York, 1998.

[13] Bernhard Schölkopf and Christopher JC Burges, *Advances in kernel methods: support vector learning*, MIT press, 1999.

[14] Cuihong Wen, Ana Rebelo, Jing Zhang, and Jaime Cardoso, “Classification of optical music symbols based on combined neural network,” in *Mechatronics and Control (ICMC), 2014 International Conference on*. IEEE, 2014, pp. 419–423.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[17] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, “Neurocomputing: foundations of research,” *JA Anderson and E. Rosenfeld (Eds.)*, pp. 696–699, 1988.

[18] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[19] Alex Krizhevsky and Geoffrey Hinton, “Learning multiple layers of features from tiny images,” 2009.

[20] Yoshua Bengio, “Practical recommendations for gradient-based training of deep architectures,” in *Neural Networks: Tricks of the Trade*, pp. 437–478. Springer, 2012.

[21] Vinod Nair and Geoffrey E Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.