Guitar Chord Recognition Based on Finger Patterns with Deep Learning

Takumi Ooaku
Faculty of Science and Engineering, Teikyo Univ.[1]
1461069u@stu.teikyo-u.ac.jp

Tran Duy Linh
Graduate School of Science and Engineering, Teikyo Univ. [1]
duylinh161287@gmail.com

Masayuki Arai
Graduate School of Science and Engineering, Teikyo Univ. [1]
arai@ics.teikyo-u.ac.jp

Tsukasa Maekawa
Graduate School of Science and Engineering, Teikyo Univ. [1]
hello.navi.local@gmail.com

Kozo Mizutani
Graduate School of Science and Engineering, Teikyo Univ. [1]
mizutani@ics.teikyo-u.ac.jp

ABSTRACT
Many guitar players use video contents such as Youtube to practice. If the content contains noise or background sounds, then the player must watch the videos repeatedly, which is very troublesome. In order to solve this problem, we attempt to build a system that can recognize the finger patterns of guitar players in video and can automatically generate a corresponding musical score. The present paper introduces a method to recognize finger patterns with deep learning. Experimental results reveal that a three-chord classifier can achieve a recognition rate of approximately 90% and a five-chord classifier can achieve a recognition rate of approximately 70%.

CCS Concepts
•CCS → Computing methodologies → Machine learning Machine learning approaches → Neural networks.

Keywords
Generating musical score from video, Guitar chord recognition, Finger pattern, Deep learning, GoogLeNet

1. INTRODUCTION
Many guitar players practice by viewing video content on Youtube[1] or SoundCloud[2]. However, such content sometimes includes noises and background sounds, so that the players cannot correctly distinguish chords. In order to solve this problem, we propose a musical score generation system using images of guitar playing videos. In the present paper, we introduce a method by which to recognize guitar chords from a player's finger patterns using deep learning.

2. SYSTEM OUTLINE
We are planning to establish a musical score generation system from images of guitar playing videos, as shown in Figure 1. The procedure used to create this system is as follows: (1) detect areas in an input guitar playing video of the fret hand, (2) recognize chords from the cropped images, and (3) generate and output a musical score.

The most important and critical function in the system is to recognize a chord from an image. In the present study, we propose a method by which to recognize chords with deep learning.

3. METHOD TO RECOGNIZE GUITAR CHORDS
A number of machine learning methods have been proposed for classification. We used deep learning for the guitar chord classification task. Since deep learning has recently become very popular in many computer vision tasks, most state-of-the-art image classifiers use deep learning [3]–[7].

3.1 Network Model
We use a pre-trained network based on GoogLeNet that includes numerous hand patterns [8]–[10]. Figure 2 shows the network topology of GoogLeNet.

3.2 Data Collection
Guitar chords are identified based on root notes and are expressed alphabetically as A, B, C, D, E, F, and G, which correspond to the seven musical notes "la", "ti", "do", "re", "mi", "fa", and "sol", respectively. Figure 3 shows examples of finger patterns used to fret the A, B, D, E, and G chords.

Deep learning requires a large amount of data for training. We used an application called vatic.js [11] to crop images that include clear finger patterns from still images of guitar playing videos. Figure 4 shows the procedure used for data collection. First, we select the cropping areas manually, and then vatic.js outputs a still image and XML data for each cropped area. Finally, the proposed program generates cropped finger pattern images. We use still images of every fifth frame in order to exclude similar finger patterns.

Figure 5 shows some examples of cropped images. The resolution of the cropped images is 250 pixel * 250 pixel.

ICCIP 2018, November 2–4, 2018, Qingdao, China

ACM ISBN 978-1-4503-6534-5/18/11…$15.00
http://doi.org/10.1145/3290420.3290422

[1] 1-1 Toyosatodai, Utsunomiya, Tochigi, Japan.
Postal code: 320-8551, Phone: +81-28-627-7225

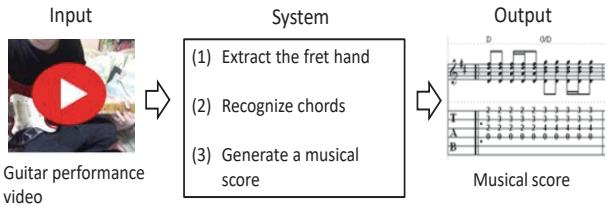

Figure 1. System configuration.

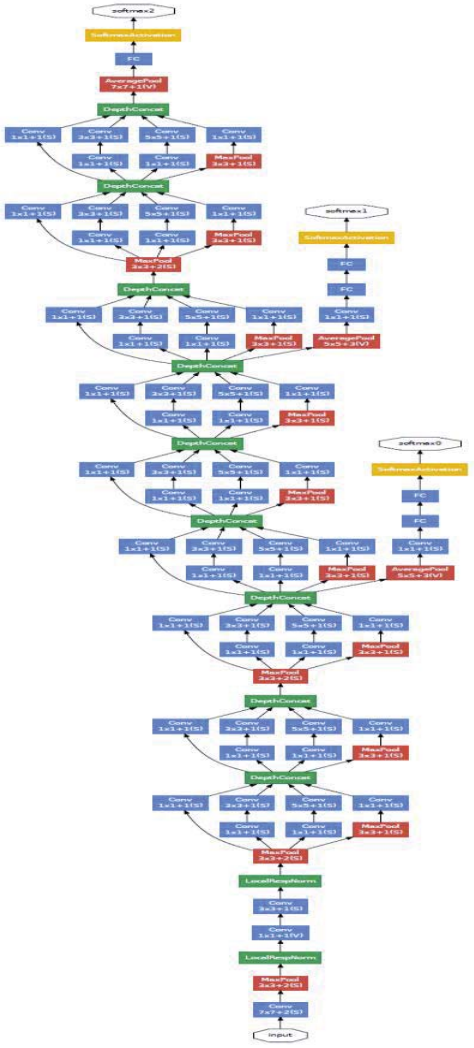

Figure 2. Network topology of GoogLeNet [11]

3.3 Developing Environment

The network is implemented within the Caffe framework [12], and the computation is run on a GeForce GTX 1080 Ti GPU.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We increase the number of categories step by step so that data collection is quite time-consuming.

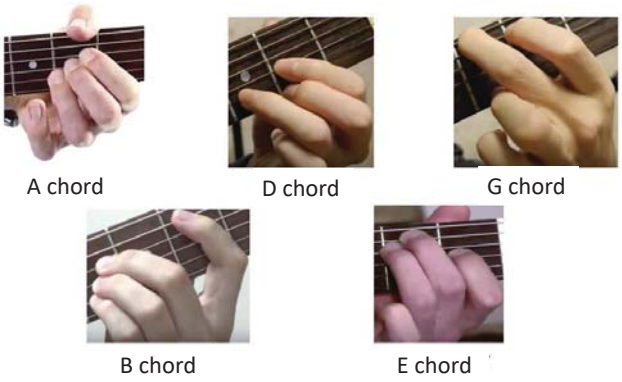

Figure 3. Examples of finger patterns used to fret the A, B, D, E, and G chords.

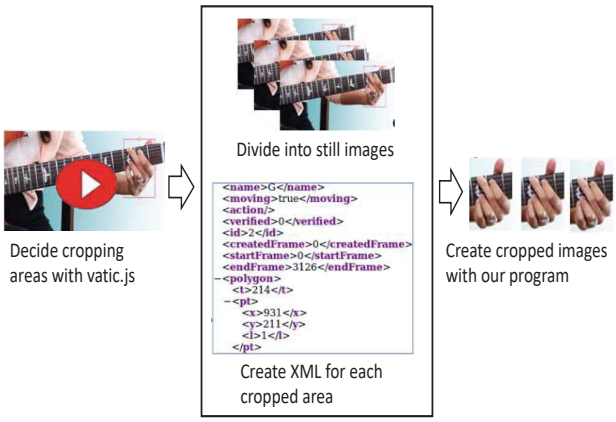

Figure 4. Procedure for collecting finger patterns.

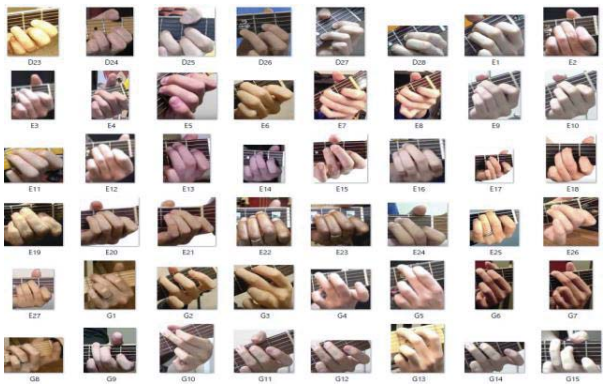

Figure 5. Examples of cropped images.

4.1 Three Categories

A number of very simple musical scores comprise only three chords. As such, we performed a three-chord experiment involving the A, D, and G chords. Table 1 shows the numbers of images data used in the experiment.

Table 1. Numbers of training and test data used in the three-chord experiment

Chord	Training data	Test data
A	302	163
D	284	153
G	183	165

We fine-tuned many of the parameters of the network, in particular, the learning rate and the number of training iterations. Finally, we achieved a recognition rate of approximately 90%, and the training time was approximately 5 hours and 50 minutes.

The G chord test data had the highest recognition rate, followed by the D and A chords, in that order. We believe that recognition rates of the D and A chords were lower than that of the G chord because the finger patterns used to fret the D and A chords are similar.

Figure 6 shows examples of classification results using a trained network of three chords. The D and G chords are classified correctly, whereas the A chord is misclassified. We believe that the A chord is classified incorrectly because the background image and the guitar are similar, as shown in Figure 6.

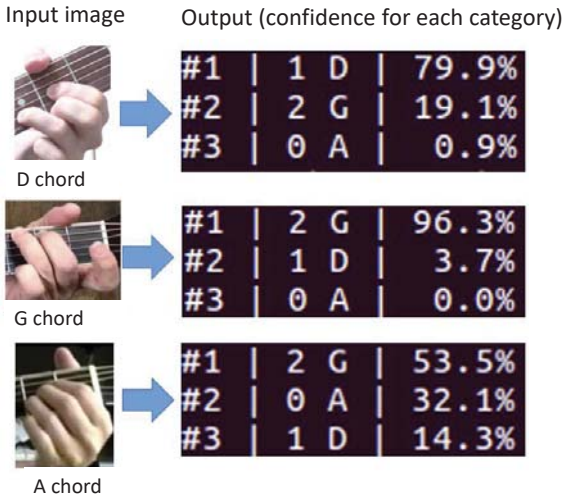

Figure 6. Examples of classification results using the three-chord trained network.

4.2 Five Categories

We also performed a five-chord experiment involving the A, B, D, E, and G chords, which are often used in guitar music. We first used the same number of data used the three-chord experiment. However, a lack of image data caused overfitting. As such, we augmented the number of image data, as shown in Table 2.

After fine-tuning the parameters of the neural network, we achieved a recognition rate of approximately 70%.

Table 2. Number of training and test data used in the five-chord experiment

Chord	Training data	Test data
A	342	172
D	324	162
G	223	174
B	336	132
E	253	169

4.3 Discussion

Predicting the recognition rate as the number of categories is increased is very difficult [13]. Figure 7 shows the relationship between the number of categories and the recognition rate.

Under this experimental condition, we could only achieve a recognition rate of approximately 55% for the A, B, C, D, E, F, and G chords. We must increase the number of image data and reconsider the network architecture in order to achieve a higher recognition rate.

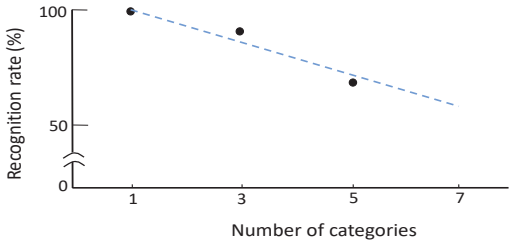

Figure 7. Number of categories vs. recognition rate.

In actual guitar playing videos, since the player's fret hand moves from chord to chord, the player may not always be fretting a chord. As such, a "no-chord" category should also be defined.

The processing time per image of the proposed network was a few seconds, and the time complexity must be reduced if the proposed system is to be used with real videos.

5. RELATED RESEARCH

5.1 Music Score Generation System

Most musical score generation systems use only sound to generate scores. For instance, Chordify [14] is a website that can generate a score from the soundtrack of a video. Chord Tracker [15] is a smartphone application that uses music in the phone. Finally, MIDI Musical Score [16] is an application that converts a musical instrument digital interface (MIDI) file into a musical score.

5.2 Hand Gesture Recognition

Hand gesture recognition is a very significant technology for human computer interaction (HCI) and sign language recognition [10]. The histogram of oriented gradients (HOG) [17], the scale-invariant feature transform (SIFT) [18], binary robust independent elementary features (BRIEF) [19], oriented FAST and rotated BRIEF (ORB) [20], and the Gabor filter response [21] have often been used for feature extraction. On the other hand, the support vector machine (SVM) [22], the hidden Markov model (HMM), the conditional random field (CRF) [23], and adapted boosting (AdaBoost) [24] have been used as classifiers. At present, most state-of-the-art research uses deep learning for not only feature extraction but also classification.

6. CONCLUSION

In the present study, we established a musical score generation system from images of guitar playing videos. The most important function in the system is recognizing a chord from an image. In the present paper, we proposed a method by which to recognize guitar chords with deep learning using a pre-trained network based on GoogLeNet.

Experimental results reveal that a three-chord classifier can achieve a recognition rate of approximately 90% and a five-chord classifier can achieve a recognition rate of approximately 70%.

In the future, we intend to augment the number of training data in order to improve the recognition rate and to increase the number of chords recognized by the system in order to expand the range of music to which the proposed system can be applied. Eventually we plan to build a system generating guitar chords from guitar playing videos automatically.

7. REFERENCES

[1] https://www.youtube.com Feb. 21, 2018.

[2] https://soundcloud.com/ Feb. 21, 2018.

[3] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[4] X. Zhang, Z. Li, C. C. Loy, and D. Lin. Polynet, "A Pursuit of Structural Diversity in Very Deep Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[5] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual Path Networks," arXiv preprint arXiv:1707.01083, 2017.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation Networks," ArXiv preprint arXiv:1709.01507, 2017.

[7] Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." arXiv preprint arXiv:1707.07012, 2017.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Proceeding of the IEEE onference on Computer Vision and Pattern Recognition, 2015.

[9] https://www-i6.informatik.rwth-aachen.de/~koller/1miohands/ Feb. 19, 2018.

[10] Ayan Sinha, Chiho Choi, and Karthik Ramani, "DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition 2016.

[11] https://dbolkensteyn.github.io/vatic.js/ Feb. 19, 2018.

[12] Jia, Yangqing, et al., "Caffe: Convolutional architecture for fast feature embedding," Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp. 675-678, 2014.

[13] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei, "What Does Classifying More Than 10,000 Image Categories Tell Us?," European Conference on Computer Vision (ECCV) 2010, pp.71-84, 2010.

[14] https://noizmoon.com/2014/04/chordify.html Feb. 21, 2018.

[15] https://app-liv.jp/975438908/ Feb. 21, 2018.

[16] https://app-liv.jp/797590911/ Feb. 21, 2018.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[18] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," International Journal of Computer Vision, vol. 60, no.2, pp. 1150-1157, 2004.

[19] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "BRIEF: Binary Robust Independent Elementary Features," 11th European Conference on Computer Vision (ECCV), Heraklion, Crete. LNCS Springer, September 2010.

[20] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski, "ORB: An efficient alternative to SIFT or SURF," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2564-2571, 2011.

[21] R. Sandler and M. Lindenbaum, "Gabor Filter Analysis for Texture Segmentation," Computer Vision and Pattern Recognition Workshop, 2006.

[22] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large Margin Methods for Structured and Interdependent Output Variables," The Journal of Machine Learning Research, vol. 6, no.9, pp.1453-1484, 2005.

[23] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," ICML'01 Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289, 2001.

[24] Y. Freund and R. E. Schapire, "A Decision Theoretic Generalization of On-line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.