Recognition of Drum Music using Sound and Video

Kodai Hara and Hiroyoshi MIWA

Abstract Drums are one of the important instruments in music, especially in rock and pop music. At present, many videos of playing drums have been posted on video sites. These performance videos are viewed not only for the enjoyment of watching, but also for the benefit of improving performance skills, especially for intermediate and advanced drummers. However, for beginners of drumming, it is often difficult to refer to the performance video because the performance movement is too fast or it is invisible behind instruments or other objects. Therefore, there is a great need for musical scores corresponding to performances in videos. The musical scoring was usually done by people with experience and ability, but in recent years, research has been conducted on the automation of the musical scoring, especially from sound sources, for various musical instruments such as pianos. However, there are few studies on musical scoring for a drum set. This is because it is difficult to identify drum sounds by frequency analysis. In this paper, we propose a method to increase the accuracy of the musical scoring by using both music and video of a drum performance. In addition, we conduct performance evaluation experiments on actual drum performances using the system implementing the proposed method, and we show the effectiveness of the proposed method.

1 Introduction

Drums are one of the important instruments in music, especially in rock and pop music. In addition to setting the beat and rhythm of a musical piece, it is also used

Kodai Hara
Kwansei Gakuin University, 1 Gakuen Uegahara, Sanda, Hyogo, 669–1330 Japan, e-mail: iul01481@kwansei.ac.jp

Hiroyoshi Miwa
Kwansei Gakuin University, 1 Gakuen Uegahara, Sanda, Hyogo, 669–1330 Japan, e-mail: miwa@kwansei.ac.jp

to add accents and drum fill-in (improvising over a few bars to create variations). At present, many videos of playing drums have been posted on video sites. These performance videos are viewed not only for the enjoyment of watching, but also for the benefit of improving performance skills, especially for intermediate and advanced drummers. However, for beginners of drumming, it is often difficult to refer to the performance video because the performance movement is too fast or it is invisible behind instruments or other objects. Therefore, there is a great need for musical scores corresponding to performances in videos.

Musical scoring is the generation of musical scores from recorded music and recorded performance videos. The musical scoring was usually done by people with experience and ability, but in recent years, research has been conducted on the automation of the musical scoring, especially from sound sources, for various musical instruments such as pianos. However, there are few studies on musical scoring for a drum set. Here, the musical scoring of a drum set is to determine the drums and hitting times of all musical notes in a drum performance. Most of the previous studies have used electronic drum sound sources or limited to music of a drum set containing only three drums; the hi-hat, snare, and bass. This is because it is difficult to identify drum sounds by frequency analysis.

In this paper, we propose a method to increase the accuracy of the musical scoring by using both music and video of a drum performance.In order to propose a system that can be easily used by the general public, the system does not use costly sensors that detect movement or vibration. In addition, we conduct performance evaluation experiments on actual drum performances using the system implementing the proposed method, and we show the effectiveness of the proposed method.

2 Related Works

There are many studies on drum performance recognition from acoustic signals [1, 2]. Methods based on Non-negative Matrix Fractrization (NMF) [3, 4, 5, 6, 7], and Recurrent Neural Networks (RNN) [8, 9, 10], are representative for drum music recognition and have achieved high accuracy [1]. NMF is an algorithm that decomposes a nonnegative matrix into two other nonnegative matrices. When NMF is applied to acoustic signals, NMF can be applied to a matrix represented by frequency $\times$ time and decomposed into the basis of the sound and its activation. In other words, when it is applied to the voice of a drum, it can be decomposed into the average spectrum and activation of each percussion instrument [6, 7], and the classification and onset estimation of percussion instruments can also be performed using this information. In [7], Dittmar et. al. applied NMF to mono drum sound sources and classified them into 3 classes: hi-hat, snare drum, and kick drum, and obtained F values of more than 95%. NMF-based drum music recognition is also suitable when large learning data sets are not available because it performs well with only a small amount of learning data. Reliable performance can be expected when the only voice data to be analyzed is drum sounds, but there is still plenty of room for improvement in music

recognition in music containing multiple percussion sounds or mixed with musical instrument tones containing melodies [1].

RNN is a neural network often used for data analysis of time series data. In a conventional Deep Neural Network (DNN), data input from an input layer is sequentially propagated to multiple intermediate layers. However, RNN is composed of only an input layer, one intermediate layer, and an output layer, and the final output can be said to be influenced by the past output because the calculation result in the middle layer is output and the same intermediate layer is input again at the same time. For this RNN reason, it is often used in the analysis of time series data, and its application to melodic recognition of drums has been reported in [8]. In addition, Southall et al. [9], have also introduced Bidirectional RNNs (BRNN), a system based on bidirectionality in music recognition of drums. RNNs are neural networks that are only unidirectional from past to future in forward propagation, whereas BRNNs are neural networks that also combine backward propagation from future to past. As a result, better results were obtained for solo drum data sets than for RNN-based music recognition.

In acoustic signal analysis, RNN-based methods are the most promising and boast high accuracy, but only when large data sets are available. On the other hand, the NMF-based method is suitable for cases where large learning data sets are not available because it performs well with only a small amount of learning data.

3 Drum Music Recognition Method

3.1 Overview of proposed method

In the basic idea of the proposed method, drum music recognition is performed by using acoustic signal processing and video processing to estimate onset time and classify each type of percussion instrument. The acoustic signal processing uses a monaural sound source with a sampling rate of 22050 Hz and a resolution of 16 bits as input. The frequency characteristics of each percussion instrument are then used to classify which percussion instrument was played at the onset time. For video processing, the input video is a 30 fps MP4 file with a resolution of 3840×2160. The color of the stick tip marker is binarized with other colors, and blob analysis is performed to classify the percussion instrument type based on the overlap information between the marker and the region of each percussion instrument. Finally, the percussion instruments that are estimated to have been played by both acoustic signal analysis and video processing are used as the output results at each time.

3.2 Acoustic signal processing

We describe the acoustic signal processing method for drumming. The input signal has a sampling rate of 22050 Hz, a quantization bit number of 16 bits, and a WAVE file in which the sound of the recorded video is converted to mono.

Fourier transform is a technique to convert functions related to time into functions related to frequency, but it loses information related to time. Therefore, the short-time Fourier transform is to maintain the function related to time by Fourier transform it by multiplying the window function while shifting the function related to time in a short time.

The window function is a function that yields zero outside a certain interval. Waveforms separated by a finite number of samples are unlikely to become periodic functions, discontinuities occur when the values at both ends of the separated intervals are different, and Fourier transform disperses the spectrum over a wide range from low to high frequencies. The function to prevent this is the window function. Frequently used window functions are Hamming window, Hanning window and Blackman window. In this paper, in the short-time Fourier transform, 2048 points are divided and Fourier transform is carried out in each interval using 512 as a frame period. The frequency point is 2048 points, and the Hanning window is used for the window function.

The onset time is the time when any percussion instrument is sounded, and the onset time is estimated by acoustic signal processing. First, the function of the time change in the power spectrum is obtained by squaring the absolute value of the short-time Fourier-transformed function, and then the function of the time change in signal intensity (decibel) is obtained.

The onset intensity of each frequency point at time t can be obtained by subtracting the power of each frequency point at time t from the power of each frequency point at time t by applying maximum filtering [11] to the power of each frequency point at time $t-1$. The onset time can be determined by examining the peak from the onset intensity at each time. Figure 1 shows the temporal change in the onset intensity of the sound produced by actually sounding the snare drum twice apart and the onset time estimated from the peak.

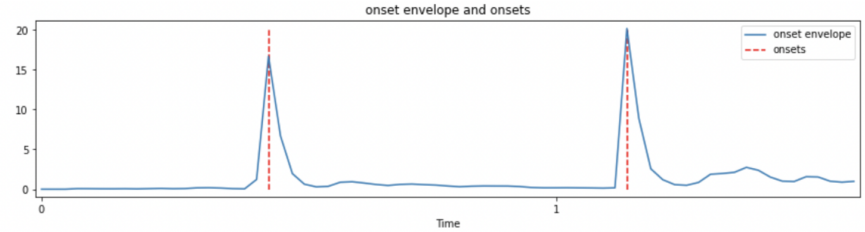

Fig. 1: Time variation of onset intensity

We describe the method to classify each percussion instrument. Figure 2 shows the power spectrum of each percussion instrument when each percussion instrument is struck only once. The horizontal axis is frequency and the vertical axis is power. Table 1 shows the frequency peaks, or frequency characteristics, of each percussion instrument. The proposed method uses these characteristics to classify each percussion instrument. Finally, the percussion instrument is identified from the frequency characteristics of the acoustic signal at the estimated onset time.

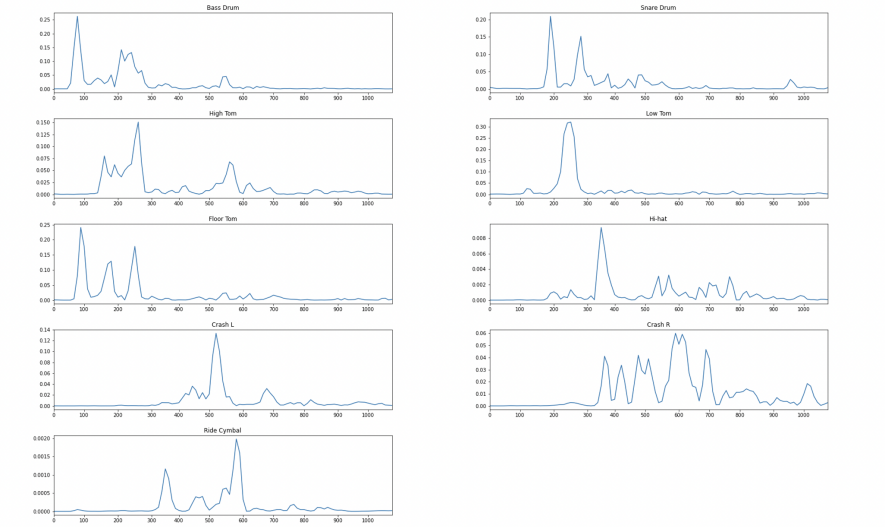

Fig. 2: FFT result for each percussion instrument

Table 1: Peak frequency of each percussion instrument

Types of percussion instruments	Peak frequency (Hz)
Bass Drum	75
Snare Drum	194
High Tom	269
Low Tom	258
Floor Tom	86
Hi-hat	355
Crash L	516
Crash R	613
Ride Cymbal	839

3.3 Video analysis

We describe the method for analyzing the moving image of drums. The video is shot directly above the drumming. It also has 30 fps in mp4 format, and the video size is 3840×2158. The aim is to estimate which percussion instrument has been struck by tracking the blue tape at the end of the stick (hereafter referred to as the marker) and examining the overlap of the marker with the area of each preset percussion instrument. Figure 3 shows a drum performance being filmed.

Fig. 3: A frame in Video filming of Drum Performance

First, each frame of the image is binarized to track the marker at the end of the stick. Binarization is an image processing method that converts an image to be analyzed into only two colors, white and black, and the process speed is improved by clarifying the boundary between the image and the background through binarization processing, making analysis easy. In this study, the color space is converted from RGB to HSV in order to binarize with blue, which is the color of the marker, and other colors. RGB and HSV are both methods of representing colors, and while RGB expresses colors using a combination of red, green, and blue elements, HSV expresses colors using Hue, Saturation, and Value elements, making it easier to intuitively understand colors and to set parameters when binarizing colors compared to the RGB color space. In addition, the HSV color space is adopted in this study because the adjustment of the parameters becomes complicated when the brightness is unstable in RGB.

The median filter is applied to the binarized frame of the performance video. The median filter is one of the denoising methods, and is a filter that outputs the middle pixel as the middle value when odd $k \times k$ pixels are sorted from smallest to smallest.

Figure 4 is an image in which the median filter is applied after the actual binarization. In Figure 4, a red circle surrounds the marker at the end of the stick in the image of Figure 5.

Fig. 4: Position of marker

Then, blob analysis is performed. Blob analysis is a technique to analyze images after binarization. Generally, the white part of the binarized image is analyzed from the position, size and area of the object. We estimate when each percussion instrument is likely to be struck by examining the hit judgment of the object and the area of each percussion instrument. The blue circles drawn in Figure 5 represent the area of each percussion instrument. If the center of gravity of a marker obtained by binarization processing, median filter, and blob analysis overlaps the area of any percussion instrument, it is judged that the percussion instrument may have been struck. If both acoustic signal analysis and video analysis determine that a given percussion instrument has been struck, it is determined that the instrument has been struck.

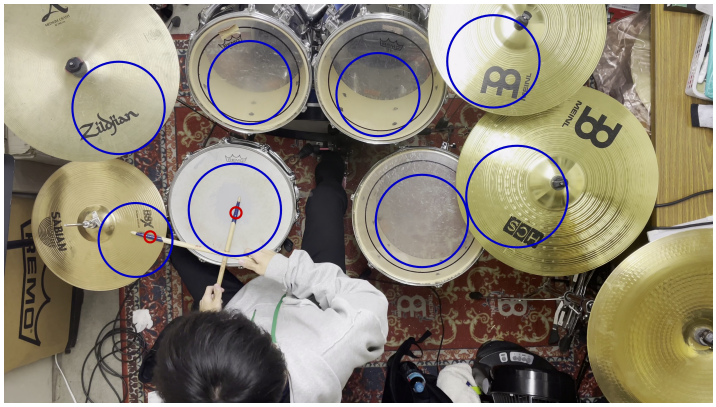

Fig. 5: Positions of each percussion instrument

4 Performance Evaluation

We evaluate the proposed method in this section. An original piece of music using nine percussion instruments is played at five different tempos, and the estimated result only from acoustic signal analysis is compared with the estimated result by combining the proposed method of acoustic signal analysis and video processing.

The original musical composition shown in the score of Figure 6 is played and evaluated in five tempos of BPM 80, 100, 120, 140, 160. The experimental results is evaluated in terms of the reproducibility, the suitability ratio and the F-value, and shall be calculated by the following equations.

$$\text{reproducibility} = \frac{\text{Number of correct detected sound}}{\text{Number of times the sound actually sounded}}$$

$$\text{suitability ratio} = \frac{\text{Number of correct detected sound}}{\text{Number of detected sound}}$$

$$\text{F-value} = \frac{2 \times \text{reproducibility} \times \text{suitability ratio}}{\text{reproducibility+suitability ratio}}$$

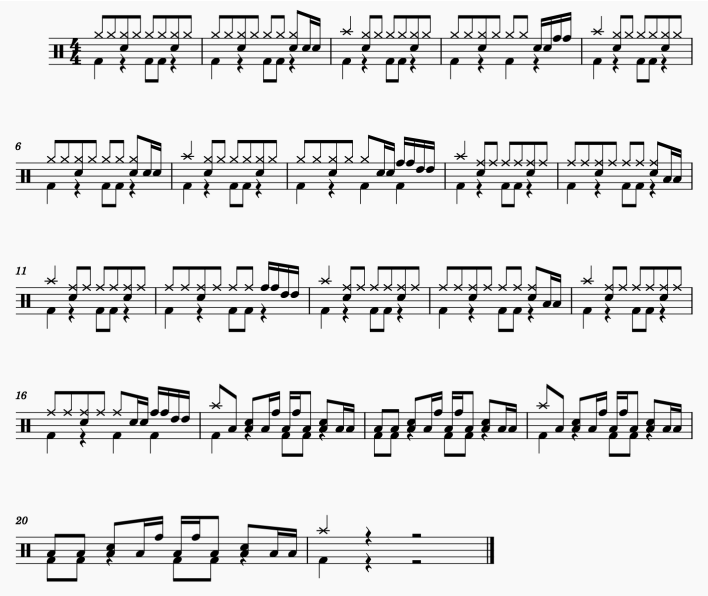

Fig. 6: Musical score

Table 2 shows the results of the evaluation for each percussion instrument by acoustic signal analysis alone, and Table 3 shows the results of the evaluation by

the proposed method that also uses video processing. In addition, the experimental results of the proposed method, in which number of correct detected sound, number of undetected sound, and number of detected sound are respectively shown in musical scores, are shown in Figure 7. As for the colors used in these scores, black is the correct detected sound, red is the undetected sound when it is actually played but not detected, and blue is the detected sound but not actually played.

The results show that the F-value of the proposed method exceeded that of the method based on acoustic signal analysis alone for all BPM and percussion instruments. The results of the proposed method show that the F-values of the toms (High tom, Low tom, Floor tom) and the bass drum are lower than the others. As for the bass drum and the floor tom, it is difficult to classify these two drums by acoustic signal analysis because the peak frequencies are similar. Another reason for the low F-value is that the bass drum is not analyzed by video. In fact, many false positives are noticeable in the parts that uses a lot of floor toms in the latter half of the piece of music. It is also difficult to classify high tom and low tom, because they had almost the same peak frequency. In addition, these toms are sometimes not detected, even if they are actually played, because their peak frequencies change depending on how hard or where they are struck. Crash cymbals are also sometimes not detected, because the peak frequency varies depending on how they are struck. Another problem is that the onset times are not detected in the experiments at BPM 120 or higher, because the tam performance is too fast.

Table 2: Results when analyzed only by acoustic signals

Types of percussion instruments	BPM	reproducibility	suitability ratio	F-value
Bass Drum	80	95.2%(59/62)	68.6%(59/86)	0.80
	100	95.2%(59/62)	67.8%(59/87)	0.79
	120	96.8%(60/62)	72.3%(60/83)	0.83
	140	90.3%(55/62)	76.7%(55/71)	0.83
	160	90.3%(56/62)	76.7%(56/73)	0.83
Snare Drum	80	100%(46/46)	100%(46/46)	1.00
	100	100%(46/46)	97.9%(46/47)	0.99
	120	97.8%(45/46)	95.7%(45/47)	0.97
	140	89.1%(41/46)	95.3%(41/43)	0.92
	160	91.3%(42/46)	95.5%(42/44)	0.93
High Tom	80	87.5%(14/16)	22.6%(14/62)	0.36
	100	100%(16/16)	25.8%(16/62)	0.41
	120	56.3%(9/16)	23.7%(9/38)	0.33
	140	18.8%(3/16)	10.3%(3/29)	0.13
	160	18.8%(3/16)	10.3%(3/29)	0.13
Low Tom	80	100%(6/6)	13.3%(6/45)	0.24
	100	100%(6/6)	16.7%(6/36)	0.29
	120	100%(6/6)	16.7%(6/36)	0.29
	140	100%(6/6)	26.1%(6/23)	0.41
	160	33.3%(2/6)	13.3%(2/15)	0.19
Floor Tom	80	94.7%(36/38)	39.1%(36/92)	0.55
	100	94.7%(36/38)	38.7%(36/93)	0.55
	120	86.8%(33/38)	39.8%(33/83)	0.55

Table continues on next page

Continued from previous page

Types of percussion instruments	BPM	reproducibility	suitability ratio	F-value
	140	92.1%(35/38)	42.7%(35/82)	0.58
	160	89.5%(34/38)	43.0%(34/79)	0.58
Hi-hat	80	100%(51/51)	28.3%(51/180)	0.44
	100	100%(51/51)	28.3%(51/180)	0.44
	120	100%(51/51)	29.5%(51/173)	0.46
	140	100%(51/51)	32.1%(51/159)	0.49
	160	92.2%(47/51)	32.0%(47/147)	0.47
Crash L	80	100%(3/3)	5.77%(3/52)	0.11
	100	66.7%(2/3)	5.00%(2/40)	0.09
	120	100%(3/3)	9.68%(3/31)	0.18
	140	66.7%(2/3)	8.70%(2/23)	0.15
	160	33.3%(1/3)	5.56%(1/18)	0.10
Crash R	80	100%(7/7)	11.3%(7/62)	0.20
	100	100%(7/7)	13.2%(7/53)	0.23
	120	100%(7/7)	12.5%(7/56)	0.22
	140	100%(7/7)	13.7%(7/51)	0.24
	160	100%(7/7)	14.0%(7/50)	0.25
Ride Cymbal	80	100%(49/49)	27.4%(49/179)	0.43
	100	98.0%(48/49)	26.8%(48/179)	0.42
	120	95.9%(47/49)	27.8%(47/169)	0.43
	140	89.8%(44/49)	29.0%(44/152)	0.44
	160	91.8%(45/49)	30.8%(45/146)	0.46

Ends here

Table 3: Results by proposed method

Types of percussion instruments	BPM	reproducibility	suitability ratio	F-value
Bass Drum	80	95.2%(59/62)	68.6%(59/86)	0.80
	100	95.2%(59/62)	67.8%(59/87)	0.79
	120	96.8%(60/62)	72.3%(60/83)	0.83
	140	90.3%(55/62)	76.7%(55/71)	0.83
	160	90.3%(56/62)	76.7%(56/73)	0.83
Snare Drum	80	100%(46/46)	100%(46/46)	1.00
	100	100%(46/46)	100%(46/46)	1.00
	120	97.8%(45/46)	100%(45/45)	0.99
	140	87.0%(40/46)	100%(40/40)	0.93
	160	91.3%(42/46)	100%(42/42)	0.95
High Tom	80	87.5%(14/16)	77.8%(14/18)	0.82
	100	93.8%(15/16)	83.3%(15/18)	0.88
	120	56.3%(9/16)	90.0%(9/10)	0.69
	140	18.8%(3/16)	75.0%(3/4)	0.30
	160	18.8%(3/16)	75.0%(3/4)	0.30
Low Tom	80	100%(6/6)	100%(6/6)	1.00
	100	83.3%(5/6)	83.3%(5/6)	0.83
	120	83.3%(5/6)	83.3%(5/6)	0.83
	140	100%(6/6)	66.7%(6/8)	0.86
	160	33.3%(2/6)	50.0%(2/4)	0.40
Floor Tom	80	94.7%(36/38)	72.0%(36/50)	0.82
	100	94.7%(36/38)	76.6%(36/47)	0.85

Table continues on next page

Continued from previous page

Types of percussion instruments	BPM	reproducibility	suitability ratio	F-value
	120	89.5%(34/38)	87.1%(34/39)	0.88
	140	81.6%(31/38)	83.8%(31/37)	0.83
	160	86.8%(33/38)	91.7%(33/36)	0.89
Hi-hat	80	100%(51/51)	98.1%(51/52)	0.99
	100	100%(51/51)	98.1%(51/52)	0.99
	120	100%(51/51)	100%(51/51)	1.00
	140	100%(51/51)	94.4%(51/54)	0.97
	160	92.2%(47/51)	100%(47/47)	0.96
Crash L	80	100%(3/3)	100%(3/3)	1.00
	100	66.7%(2/3)	66.7%(2/3)	0.67
	120	100%(3/3)	100%(3/3)	1.00
	140	66.7%(2/3)	100%(2/2)	0.80
	160	33.3%(1/3)	33.3%(1/3)	0.33
Crash R	80	100%(7/7)	100%(7/7)	1.00
	100	100%(7/7)	100%(7/7)	1.00
	120	71.3%(5/7)	83.3%(5/6)	0.77
	140	85.7%(6/7)	85.7%(6/7)	0.86
	160	85.7%(6/7)	85.7%(6/7)	0.86
Ride Cymbal	80	100%(49/49)	100%(49/49)	1.00
	100	98.0%(48/49)	100%(48/48)	0.99
	120	95.9%(47/49)	100%(47/47)	0.98
	140	89.8%(44/49)	100%(44/44)	0.95
	160	91.8%(45/49)	100%(45/45)	0.96

Ends here

5 Conclusions

In this paper, we proposed the method to increase the accuracy of the musical scoring by using both music and video of a drum performance. The basic idea of the proposed drum music recognition method is the combination of the acoustic signal processing and the video processing. The method estimates the onset times and classifies the percussion instruments for all musical notes.

We compared the proposed method with the estimated result only from acoustic signal analysis by the reproducibility, the suitability ratio and the F-value. The experimental results showed that the F-value of the music recognition by the proposed method exceeded that of the music recognition by acoustic signal analysis alone; therefore, it can be said that it is effective to use video analysis in music recognition of drums. In the proposed method, the classification of snare drum, hi-hat, and ride cymbal yielded high F-values, but it was difficult to classify high tom and low tom, and floor tom and bass drum because they had similar frequency characteristics. In addition, high toms, low toms, and crash cymbals are sometimes not recognized because the peak frequency changes depending on the position and the way of hitting, and in some quick tom performances, onset detection was not detected.

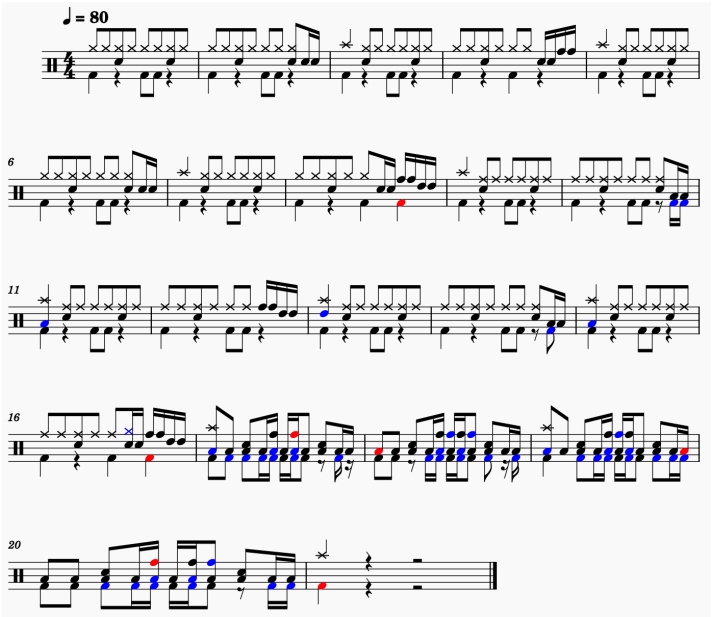

Fig. 7: Result of music recognition by proposed method at BPM 80

The results in this paper demonstrate that the accuracy of the musical scoring is increased by using both music and video of a drum performance. The improvement of the accuracy is a future work.

References

1. C. Wu, C. Dittmar, C. Southall, G. Widmer, J. Hockman, M. Mller, A. Lerch, “A Review of Automatic Drum Transcription,” IEEE/ACM Transactions on Audio, Speech, Language Processing, Vol. 26, No. 9, pp. 1457-1483, 2018.
2. K. Yoshii, M. Goto, H. Okuno, “Automatic drum sound description for real-world music using template adaptation and matching methods,” Proc. ISMIR, Barcelona, Spain, pp. 184-191, 2004.
3. J. Paulus, “Signal processing methods for drum transcription and music structure analysis,” Ph.D. dissertation, Tampere Univercity of Technology, Tampere, Finland, 2009.
4. J. Paulus, T. Virtanen, “Drum transcription with non-negative spectrogram factorisation,” Proc. EUSIPCO, Antalya, Turkey, 2005.
5. P. Roy, F. Pachet, S. Krakowski, “Improving the classification of percussive sounds with analytical features: A case study,” Proc. ISMIR, Vienna, Austria, pp. 229-232, 2007
6. E. Battenberg, “Techniques for machine understanding of live drum performances,” ph. D. dissertation, Univercity of California at Berkeley, 2012.
7. C. Dittmar, D. Gärtner, “Real-time transcription and separation of drum recordings based on NMF decomposition,” Proc. Intl.Conf. onDigital Audio Effects(DAFx), Erlangen, Germany, pp. 187-194, 2014.

8. R. Vogl, M. Dorfer, P. Knees, "Recurrent neural networks for drum transcription," Proc.ISMIR, New York City, United Dtates, pp. 730-736, 2016.
9. C. Southall, R. Stables, J. Hockman, "Automatic drum transcription using bi-direcrional recurrent neural networks," Proc. ISMIR, New York City, United States, pp. 591-597, 2016.
10. R. Vogl, M. Dorfer, P. Knees, "Drum transcription from polyphonic music with reccurent neural networks," Proc. ICASSP, New Orleans, Louisiana, USA, pp. 201-205, 2017.
11. S. Böck, G. Widmer, "Maximum filter vibrato suppression for onset detection," 16th International Conference on Digital Audio Effects, Maynooth, Ireland, pp. 1-7, 2013.