claVision: Visual Automatic Piano Music Transcription

Mohammad Akbari
School of Engineering Science
Simon Fraser University
8888 University Drive
Burnaby, British Columbia, Canada V5A 1S6
akbari@sfu.ca

Howard Cheng
Department of Mathematics and Computer Science
University of Lethbridge
4401 University Drive
Lethbridge, Alberta, Canada T1K 3M4
howard.cheng@uleth.ca

ABSTRACT

One important problem in Musical Information Retrieval is Automatic Music Transcription, which is an automated conversion process from played music to a symbolic notation such as sheet music. Since the accuracy of previous audio-based transcription systems is not satisfactory, we propose an innovative visual-based automatic music transcription system named claVision to perform piano music transcription. Instead of processing the music audio, the system performs the transcription only from the video performance captured by a camera mounted over the piano keyboard. claVision can be used as a transcription tool, but it also has other applications such as music education. The claVision software has a very high accuracy (over 95%) and a very low latency in real-time music transcription, even under different illumination conditions.

Author Keywords

claVision, automatic music transcription, piano, deaf musician, computer vision

ACM Classification

J.5 [Arts and Humanities] — Music, I.4.9 [Image Processing and Computer Vision] Applications.

1. INTRODUCTION

Automatic Music Transcription (AMT) is the process of extracting the musical sounds from a piece of music and converting them to a symbolic notation such as a music score or a Musical Instrument Digital Interface (MIDI) file using a computer. The main input is usually audio or sound, and most previous AMT methods are based on audio processing techniques [1, 2, 3]. Almost none of these methods have satisfactory accuracy because of numerous difficulties resulting from the use of audio signals such as the presence of multiple lines of music (e.g, simultaneous played notes or instruments), audio noise, exact note duration extraction, incorrect recorded pitches, etc. Thus, new technologies are required to deal with this problem.

A new vision-based system named **claVision** is introduced in this paper to perform visual automatic transcription of piano music. A camera is located at the top of the piano keyboard to capture the video of a performance. The software visually analyzes the music played on the piano based on the pressed keys and the pianist's hands. Finally, the transcription of the played music is automatically produced from the video analysis without listening to the audio of the music. Our idea was inspired by the famous composer and pianist, Ludwig van Beethoven. By the last decade of his life, he was almost totally deaf, forcing him to rely more heavily on the visual relationship between his fingers and piano keys. claVision mimics what a deaf musician does.

All musical keyboards such as piano, harpsichord, electronic organs, etc. can be used with our software. Even if only a portion of the keyboard is captured by the camera, claVision still transcribes music correctly, as long as all the keys played are visible. It has a high accuracy (over 95%) in transcribing piano music over a wide range of speed and complexity of the played music. Both live (real-time) and recorded video processing can be handled. The real-time transcription has low latency, and results are obtained as the music is being played. Despite the sensitivity of image and video processing techniques to illumination issues, claVision can deal with many different lighting conditions.

2. SYSTEM ARCHITECTURE

The requirements for performing real-time music transcription are a digital camera and any kind of stable mounts holding the camera at the top of the piano. The ideal camara location is demonstrated in Figure 1, though the software can adjust for variations in camera positions. The angled view allows the pressed keys to be seen more clearly. claVision can produce three different types of outputs including highlight of the pressed keys, music score, and the synthesized MIDI sound of the transcribed music.

Figure 1: Ideal location of the camera over the piano and electronic keyboards.

3. APPROACH

In claVision, there are four main stages to perform music transcription: keyboard registration, illumination normalization, pressed keys detection, and note transcription.

(1) Keyboard registration is achieved by performing three tasks. Keyboard detection is done at first, which includes

NIME'15, May 31-June 3, 2015, Louisiana State Univ., Baton Rouge, LA.

locating, transforming, and cropping the piano keyboard image. Next, the background update procedure is performed to identify the best background image without any hand covering the keyboard. Finally, the locations of the black and white keys as well as their musical features are calculated to be used in the next stages.

(2) To deal with varying illumination, noise, and shadows, an image smoothing approach is used to reduce the differences between the background image and the other video frames.

(3) Given the background image and the corrected image in each video frame, pressed keys detection is done in four steps. First, the difference image between the background image and the corrected video image in each working frame is computed. Next, the location of the pianist's hands is determined to identify the candidate keys potentially pressed. From the pixels identified in the difference image, the pressed keys are detected and matched with their related musical features. The detected keys are finally highlighted.

(4) At the final stage, the obtained musical information (e.g., note, octave, etc.) is transcribed to the MIDI structure and sheet music.

Most of the algorithms used in claVision are relatively simple. The design philosophy is to use the simplest possible methods in order to reduce the latency in real-time processing, while maintaining a high transcription accuracy.

4. EVALUATION

claVision was installed and tested on a laptop with an Intel Core i7-4510U CPU (2.00 GHz) and 16 GB DDR3 RAM. Different videos captured using two digital cameras, an SD 240p webcam (with resolution and frame rate of 320×240 pixels and 24 FPS) and an HD 720p webcam (with resolution and frame rate of 640×360 pixels and 30 FPS) were used to evaluate the effectiveness of claVision on different piano and electronic keyboard performances with a variety of tempo, from 40 beats per minute (BPM) to 240 BPM.

The effectiveness of our system in different steps was evaluated based on accuracy and processing time. The locations of the keyboard in all sample images were successfully detected. The keys detection process had a very high accuracy of 95.2%. The processing times in the keyboard registration stage including keyboard detection, background update, and keys detection did not affect the real-time processing because it was done in a separate thread. According to our experimental results, issues such as varying illumination conditions and noise were satisfactorily dealt with using the image correction algorithm. This algorithm has a very low processing time of 1.8 ms, and did not cause any significant latency for real-time processing in claVision. Our system has a very high accuracy in pressed keys detection with recall and precision rates of 97.5% and 97.4%. There was a very low latency of 6.6 ms for the pressed keys detection stage. The synthesized MIDI file was accurately produced based on the given musical information and it sounded the same as the played music.

In order to evaluate the correctness of the transcribed music produced by claVision, the sheet music of one of the sample video performances called "Twinkle Twinkle Little Star" is analyzed (Figure 2). Based on the test results, 10 keys are incorrectly detected as pressed in this piece of music, which are highlighted in the sheet music with the color red. All errors are related to the adjacent white keys with no black key in between (e.g., E and F). This is because

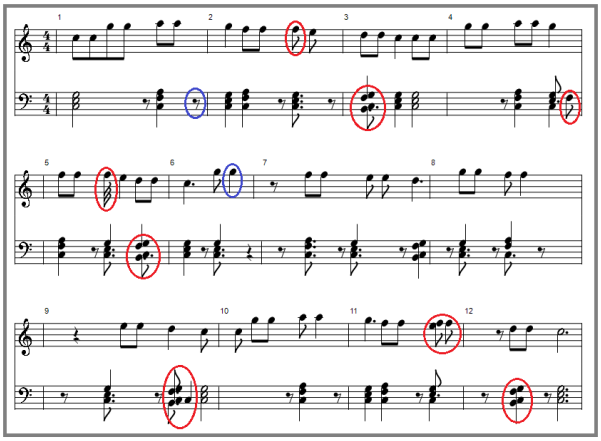

Figure 2: The produced sheet music of the song "Twinkle Twinkle Little Star".

they are more difficult to be distinguished from each other in the difference images. Our software produces the appropriate note shapes and measures in the sheet music based on the note durations in milliseconds. However, pianists usually apply tempo rubato to their performances (flexibility in time and irregularity of rhythm and/or tempo), which causes some notation errors in the duration and location of the notes and rests in the measures (e.g., the note and rest highlighted in Figure 2 with the color blue). In this case, the transcription is in fact accurate with respect to the actual performance captured in the video.

Unlike other similar products that perform automatic music transcription by "listening" to the music, in claVision, the audio of the played music is ignored. As a result, the drawbacks of existing transcription techniques from audio are no longer present. However, there is a number of limitations in our software caused by drastic lighting changes, inappropriate camera views, hands coverage of the keyboard (more than 60%), and vibrations of the camera or the piano. In addition, there are some aspects in music that are related to the expressive side of music such as tempo rubato, dynamics, and articulations that cannot be dealt with using claVision. One potential direction of claVision is to combine it with audio processing technologies to provide a robust multi-modal system for music transcription.

5. ACKNOWLEDGMENT

The work was done with funding support from the Natural Sciences and Engineering Research Council Discovery Grant Program and the Alberta Innovate Technology Futures. The authors would also like to thank the Microsoft Imagine Cup competition for the opportunity to showcase our software, as well as Bill Buxton for his encouragement on our work.

6. REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41:407–434, 2013.

[2] A. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282, 2004.

[3] C. Yeh. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, Universite Paris VI - Pierre et Marie Curie, France, 2008.