Applied Mechanics and Materials Vols. 333-335 (2013) pp 742-748
© (2013) Trans Tech Publications, Switzerland
doi:10.4028/www.scientific.net/AMM.333-335.742

Automatic transcription of piano music using audio-vision fusion

Yulong Wan[1, a], Zhigang Wu[2, b], Ruohua Zhou[3, c] and Yonghong Yan[4, d]

[1, 3, 4]DSP Building, 21 Beisihuan West Road, Haidian District, Beijing, P. R. China

[2]58 Huanghe Road, Urumqi, P. R. China

{[a]wanyulong, [c]zhouruohua, [d]yanyonghong}@hccl.ioa.ac.cn, [b]893912446@qq.com

Keywords: music transcription, polyphonic, onset detection, matched filter, audio-vision fusion

Abstract. Over the last decade many sophisticated and application-specific methods have been proposed for transcription of polyphonic music. However, the performance seems to have reached a limit. This paper describes a high-performance piano transcription system with two main contributions. Firstly, a new onset detection method is proposed using a specific energy envelope matched filter, which has been proved very suitable for piano music. Secondly, a computer-vision method is proposed to enhance audio-only piano music transcription, using the recognition of the player's hands on the piano keyboard. We carried out comparable experiments respectively for onset detection and overall system based on the *MAPS*[5] database and the video database. The results were compared with the best piano transcription system in *MIREX 2008*, which still kept the best performance in piano subset till *MIREX 2012*. The results show that the system outperforms the state-of-art method substantially.

Introduction

Music transcription is to create symbolic representation using musical notations from audio recordings. For monophonic music, this task is considered to be solved as many mature and easy-understood methods have been proposed. Varieties of attempts have also been presented for the polyphonic transcription, such as the method based on novel onset/offset detection in [4] and the one with recurrent neural networks in [3]. But due to the spectral overlapping in polyphonic music, the performance in accuracy and flexibility is still significantly below that of human experts, and the reported accuracies in recent years seem to have reached a limit. For an overview of current methods and identification of directions for future research, the readers are referred to [2] and [6].

The previous works related to the current approach include the rule-based system in [1], which employed the Resonator Time-Frequency Image (*RTFI*) as the time-frequency representation and performed best on the piano subset in *MIREX 2008*, the onset detection algorithm in [7] based on a fractional power envelope matched filter and the hand recognition systems in [8] and [9].

Here, we propose a polyphonic piano transcription system combined with computer vision recognition and a new onset detection method using specific energy envelope matched filter. Experiments were carried out on recordings from the *MAPS* database and video set recorded by a simple setup described below, and the performances were competitive compared to the transcription system proposed by Zhou and Reiss in [1].

System description

In our system, a video camera was mounted above the piano with its field of view covering the whole keyboard, and a microphone was fixed beside the piano. While the pianist was playing, the camera captured the video and the audio was recorded simultaneously. Then we carried out processing according to the overall flow in Fig. 1.

The outline of this paper is as follows: Section 3 describes the *RTFI* representation and Section 4 details the proposed onset detection method; in Section 5, multiple pitch estimation and preliminary transcription are presented; Section 6 details the hand recognition and the fusion method; finally, we describe the experiments and evaluations in Section 7 and draw conclusions in Section 8.

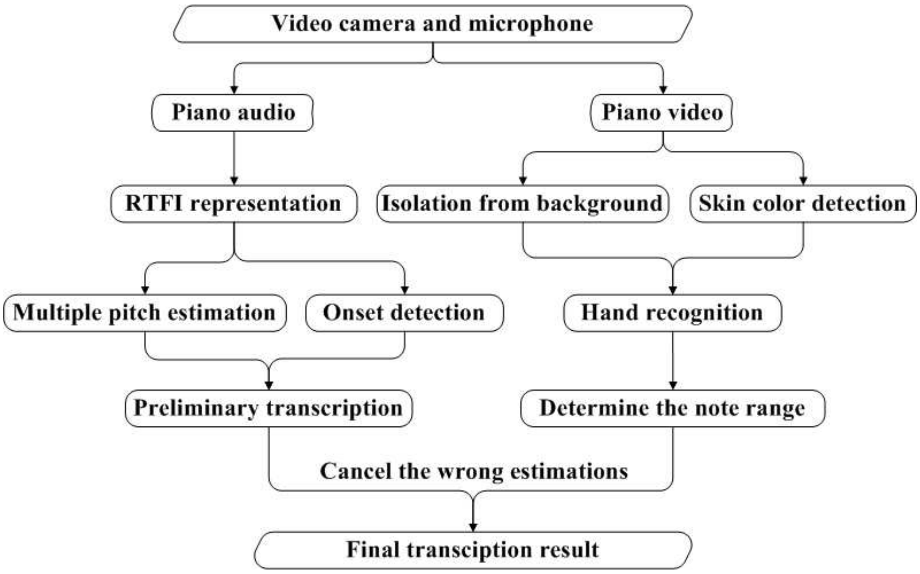
Video camera and microphone
Piano audio
Piano video
RTFI representation
Isolation from background
Skin color detection
Multiple pitch estimation
Onset detection
Hand recognition
Preliminary transcription
Determine the note range
Cancel the wrong estimations
Final transciption result

Fig. 1 Overall flow of the piano transcription system

Time-frequency Representation

We employed the *constant-Q RTFI* in Eq. 1 for our time-frequency representation, which was proposed in [10]. Its main feature is that it selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis.

$$RTFI(t,w)=s(t)*I_R(t,w)=r(w)\int_0^t s(\tau)e^{r(w)(\tau-t)}e^{-jw(\tau-t)}d\tau \tag{1}$$

Where $I_R(t,w)=r(w)e^{-(r(w)+jw)t}, t>0$, I_R denotes the impulse response of the first-order complex resonator filter with oscillation frequency *w*, and the factor $r(w)$ is used to normalize the gain of the frequency response.

We chose it due to the flexibility with regards to time-frequency resolution, the simplicity and computational efficiency of the implementation based on first-order filters. The more details of the *RTFI* can be found in [10].

In this paper, the analyzed frequency range was from 25.96 Hz to 4.43k Hz, and 890 filters were used to cover the analyzed frequency range. The center frequencies were set in logarithmic scale and the frequency interval between two neighboring filters was equal to 0.1 semitones. Then the time interval between two successive frames was set to 10ms, correspondingly the number of samples for calculating the Average Energy Spectrum (*AES*) was set to 441.

According to the description above, we defined the *AES* in Eq. 2:

$$AES(n,f)=db\left(\frac{1}{441}\sum_{i=(f-1)\times 441+1}^{f\times 441}|RTFI(n,i)|^2\right) \tag{2}$$

Where *n* denotes the index of the log-frequency bin, $RTFI(n,i)$ denotes the value of discrete *RTFI* at sampling point *i*, and *f* is the frame index. The AES is then used as the only input for the piano transcription system.

Onset Detection

Preprocessing. According to the different sensitivities of the human auditory system in the different frequency bands, which is often described by tracing equal-loudness contours, we preprocessed the audio signal before detecting the piano onsets. In our method, the *AES* was adjusted following the *Robinson and Dadson equal-loudness contours*, which have been standardized in the international standard *ISO-226* [11]. This standard provides equal-loudness contours limited to 29 frequency bins, and we only used the contour corresponding to 70DB, getting the equal-loudness contours (*EC*) of 890 frequency bins by cubic-spline interpolation in the logarithmic frequency scale. The Equal-Loudness Average Energy Spectrum (*EAES*) was calculated in Eq. 3:

$$EAES(n,f) = AES(n,f) - EC(n) \tag{3}$$

Where n denotes the index of the frequency bin and f represents the frame index.

Onset detection. The concept of the matched filter (*MF*) is to use the reversal of the pattern as the filter to find the desired object. To detect the onsets, we applied a very simple matched filter to the average energy envelope for each pitch, and both the accuracy and the computation efficiency were significantly improved comparing to the methods before.

Firstly, the average energy values were calculated around each pitch for all frames based on the *EAES* in Eq. 4:

$$Mean(n_{midi}, f) = \frac{1}{13}\sum_{l=n_{midi}-6}^{n_{midi}+6} EAES(l,f) \tag{4}$$

Where f denotes the index of the frame, and the number n_{midi} represents the index of the frequency bin.

As the width of the time slots we chose was 10ms, it could be observed from the statistics that the values of the energy envelope in the onset regions were near to:

[-9.6258, -6.9773, -5.6138, -4.6522, -3.968, -4.7485, -1.6418, 9.9056, 13.3167, 11.7671, 9.3814, 9.6014]

Therefore, an average energy envelope *MF* was proposed according to the definition above:

MF[12]=[9.6014, 9.3814, 11.7671, 13.3167, 9.9056, -1.6418, -4.7485, -3.968, -4.6522, -5.6138, -6.9773, -9.6258]

Afterwards, we performed the convolution of the mean energy array (obtained in (4)) for each pitch in Eq. 5 using this specific *MF*:

$$CV(n,f) = \sum_{\tau=1}^{12}(Mean(n, f-\tau)\times MF[\tau]) \tag{5}$$

Where n denotes the index of the frequency bin, and f is the frame index.

Finally, onsets of each pitch were detected by simply peak picking on $CV(n,\ldots)$, and it was also treated as non-onset if the mean value was below -70 or the convolution result was below 150 at the peak. The last procedure of the proposed onset detection method was to summarize all the onsets detected with a minimum 100ms distance. Fig. 2 shows the details of detecting the onsets of MIDI number 30.

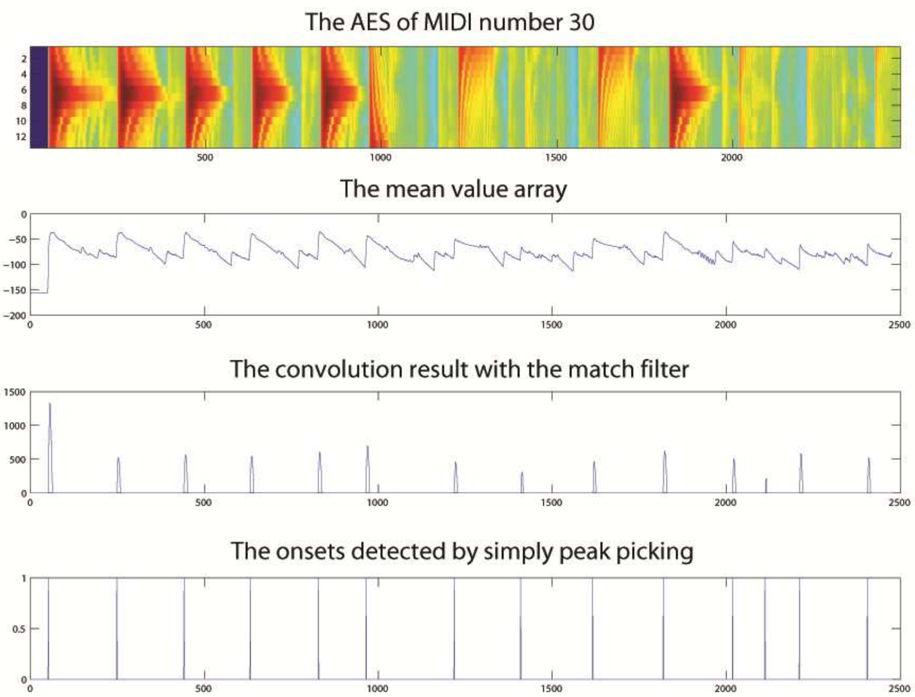

Fig. 2 Details of detecting the onsets of MIDI number 30

Multiple Pitch Estimation

To estimate multiple pitches in each frame, the input *AES* was firstly transformed into the Pitch Energy Spectrum (*PES*) based on the harmonic grouping principle, and then into the Relative Pitch Energy Spectrum (*RPES*), which represented the distribution of the pitches existing. As the frequency indexes were in the logarithmic scale, the Harmonics Deviations (*HD*) could be approximated as in Eq. 6:

$$\mathrm{HD}[8] = [0, 120, 190, 240, 279, 310, 337, 360] \tag{6}$$

Accordingly, the *PES* and *RPES* could be easily approximated in the logarithmic scale by the following calculations in Eq. 7 and Eq. 8:

$$PES(n,f) = \frac{1}{N_h}\sum_{i=1}^{N_h} AES(n + HD[i], f) \tag{7}$$

$$RPES(n,f) = PES(n,f) - \frac{1}{2N_r + 1}\sum_{i=n-N_r}^{n+N_r} PES(i,f) \tag{8}$$

Where f denotes the frame index, and N_h denotes the number of the *n*th pitch's harmonic components. In our experiment, N_h was set to 4, and N_r was set to 30. Afterwards, the *AES* was transformed to the Relative Energy Spectrum (*RES*) according to the following expression in Eq. 9:

$$RES(n,f) = AES(n,f) - meanV(AES(...,f)) \tag{9}$$

Where $meanV(AES(...,f))$ denotes the average value of the *f*th frame. The peaks of *RES* can be used to mark the harmonic components' existence.

Based on the *RPES* and *RES*, we estimated the possible pitches in each frame following the rules detailed in [1], which have been proved very useful by a great number of experiments.

Although these rules have improved the performance of the multiple pitch estimation a lot, there are still many extra incorrect estimations focus on the pitches whose note intervals are 120 or 190 higher than the true pitches in the logarithmic scale. And these extra pitches can be cancelled by judging the positions of pianist's hands. So we considered the recognition of hands as a method to rectify the transcription results.

Hand Recognition and Combination

The equipment setup. Considering that the recognition algorithm was applied in every video frame, we preferred the 640×480 resolution of the camera and the frame rate was defined as 25fps in order to achieve a detailed illustration of hand movements. Besides, we chose the white light which also played a crucial role in the hand recognition via skin detection. Bad lighting or the one that creates a visual illusion of different shades of the skin will lead to the rejection of certain point of the skin area or even to the non-detection of a skin area. Finally, we imported the video signal into the computer and used the Open Source Computer Vision Library (*OpenCV*) to process it.

Hand recognition. In our system, we chose the combination of the two methods as follows to detect the hands: isolation from a background and skin color detection with edge merging techniques.

Firstly, we delineated the piano keyboard as only the part of the image containing the piano keyboard was used in further processing and stored the background image in gray scale at the same time.

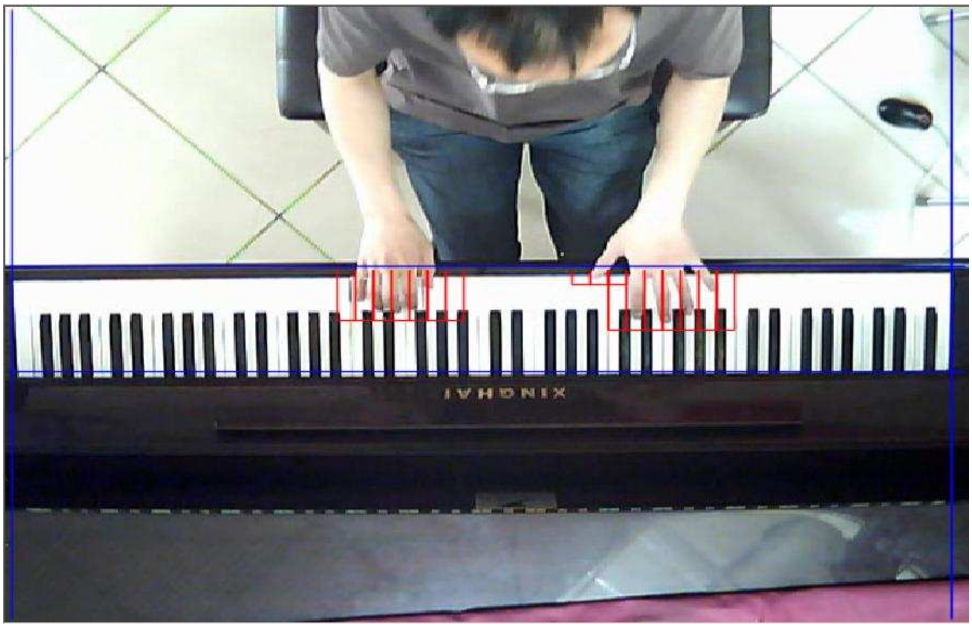

Fig. 3 Hands' positions detected on the keyboard

As the playing session started, the comparison between the foreground and the background gray image was performed to detect the moving objects. Then, we used the skin detection to isolate the hands from the motion parts.

Human skin color's specific chromatic distribution is totally different from the one of the other objects in the background, and the differentiations in the skin colors of various tribes are characterized only by the differences in brightness but not chrominance. As a result, we performed image analysis in non-linearly transformed color spaces that separate brightness values [9]. The *YCrCb* and the *HSV* color space were both employed, in which the psychophysical thresholds of human color perception are preset:

For pixel $p_{ij} = (Y_{ij}, Cr_{ij}, Cb_{ij})$ in *YCrCb*, if $Cr_{ij} \in (133,173)$ & $Cb_{ij} \in (77,128)$, then it is a skin pixel. While for pixel $p_{ij} = (H_{ij}, S_{ij}, V_{ij})$ in *HSV*, if $H_{ij} \in (0,18)$ & $S_{ij} \in (0,191)$ & $V_{ij} \in (80,255)$ or $H_{ij} \in (135,180)$ & $S_{ij} \in (0,178)$ & $V_{ij} \in (80,255)$, then it is a skin pixel. In our experiment, we confirmed a pixel as a skin pixel only if it was a skin pixel both in the *HSV* and the *YCrCb* color spaces.

The result of this procedure was a binary exportation including the hands' contours only. We finally carried out morphology skinning techniques based on dilation and erosion.

Fig. 3 shows the result of hand tracking (as red boxes circumscribed around the hands). Experiments showed that the described technique is sufficient for detecting the ranges of the pianist's hands' positions.

Finally, we used the hands' relative positions on the keyboard to determine the range of the notes played according to the distribution of the notes on the piano keyboard, and then cancelled the incorrect ones, which were not in the range, to improve the estimation precision.

Experiments and Evaluations

Experiment databases. We carried out experiments on two databases, and compared the results with the system proposed by Zhou and Reiss in *MIREX 2008* which still kept the best performance on the piano subset till 2012. The first one was the *MAPS* database to evaluate the proposed method of piano onset detection. Here, we chose 62 pieces sampled at 44.1k Hz, 16 bits, from the *MUS* (Pieces of Piano Music) subset as the test set, which consisted of stereo recordings in 9 different recording conditions. The note locations and durations had been adjusted by hand by the creator of the *MIDI* database. The second one was composed of 5 videos and corresponding audio files to evaluate the advancement of the audio-vision fusion system, whose annotations were also adjusted by the pianist.

Results and evaluations. To evaluate the performance, the detected onsets of the estimated notes must be compared with the reference ones. For a given reference onset at time *t*, a detection within a tolerance time-window [*t-60ms, t+60ms*] was considered to be correct, and the result was considered to be correct only if the note was the same with the reference and its onset was correct.

We used three frame-based evaluation measures: Precision, Recall, and F-Measure. And the Running Time (*RT*) (ms) was averaged for processing every recording of 1000*ms*:

Table 1 Comparison results of piano onset detections on MAPS database

	Precision	Recall	F-Measure	RT (ms)
Zhou and Reiss'	0.993	0.809	0.887	3.78
Ours	0.978	0.942	0.959	0.21

Table 1 details the comparison results for piano onset detection. It should be noted that the proposed onset detection method demonstrates exceptional advantage compared to the Zhou and Reiss's by improving the computation efficiency and performance especially in Recall.

Table 2 Comparison results of overall systems

	Precision	Recall	F-Measure
Zhou and Reiss'	0.640	0.686	0.662
Ours	0.710	0.786	0.746

Table 2 details the advancement of the overall system combined with computer vision. We can note that the Precision of the result is additionally improved, which indicates that the incorrect pitches are cancelled efficiently by the computer vision method.

Conclusions

In this paper, we propose an automatic transcription system for polyphonic piano music, which employs a specific onset detection method, multiple pitch estimation, and computer vision recognition of pianist's hands. Experiments were performed on piano recordings from the MAPS database and 5 videos we recorded, and the results outperformed the best piano transcription system in MIREX.

In order to reduce the number of missed estimations, our future research will focus on the specific harmonic structure of polyphonic piano pitches. Finally, our system also proves that computer vision can be a useful extension method for the transcription of polyphonic piano.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

References

[1] Ruohua Zhou and JD Reiss, "A real-time polyphonic music transcription system," in *Proceedings of the Fourth Music Information Retrieval Evaluation eXchange*, 2008.

[2] A. P. Klapuri, "Automatic Music Transcription as We Know it Today," *Journal of New Music Research*, vol. 33, pp. 269-282, 2004.

[3] S. Bock and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 121-124.

[4] E Benetos and S Dixon, "Polyphonic music transcription using note onset and offset detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 37-40.

[5] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[6] E. Benetos, S. Dixon, and D. Giannoulis, "Automatic music transcription: breaking the glass ceiling," in *13th International Society for Music Information Retrieval Conference(ISMIR)*, 2012, pp. 379-384.

[7] J. J. Ding and C. Tseng, "Improved onset detection algorithm based on fractional power envelope match filter," in *European signal processing conference*, 2011.

[8] M. Sotirios, "Computer Vision Method in Music Interaction," in *International Conference on Advances in Multimedia*, 2009.

[9] D. O. Gorodnichy, "Detection and tracking of pianist hands and fingers," in *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision*, 2006.

[10] R. Zhou, "Feature extraction of musical content for automatic music transcription," Infoscience | Ecole Polytechnique Federale de Lausanne[http://infoscience.epfl.ch/oai2d.py] (Switzerland), Ph.D. thesis 2006.

[11] D. W. Robinson and R. S. Dadson, "A re-determination of the equal-loudness relations for pure tones," *Br. J. Appl. Phys*, vol. 7, pp. 166-181, 1956.