Chinese Journal of Electronics
Vol.24, No.3, July 2015

Automatic Piano Music Transcription Using Audio-Visual Features*

WAN Yulong, WANG Xianliang, ZHOU Ruohua and YAN Yonghong

(*The Key Laboratory of Speech Acoustics and Content Understanding,Chinese Academy of Sciences, Beijing 100190, China*)

Abstract — The performance of automatic music transcription seems to have reached a limit over the last decade, and a promising direction of improvements could be to incorporate music instruments' specific parameters. We propose a novel piano-specific transcription system, using both audio and visual features for the first time. Contribution of the paper mainly includes two parts: A new onset detection method is proposed using a specific spectrum envelope matched filter on multiple frequency bands. A computer-vision method is proposed to enhance audio-only piano music transcription, through tracking the positions of the pianist's hands on the piano keyboard. Based on the MIDI Aligned piano sounds (MAPS) database and a self-recorded video database, we carried out comparable experiments for audio-only onset detection and overall system, respectively. The performance was compared with the best piano transcription system in Music information retrieval evaluation exchange (MIREX), and the results showed that the proposed system outperforms the state-of-art method substantially.

Key words — Music transcription, Polyphonic piano, Onset detection, Matched filtering, Multimedia fusion.

I. Introduction

Music transcription is a process to transform an acoustic signal into a symbolic representation, which comprises music notes, their pitches, timings, and instrument types. This is a complicated cognitive task performed routinely by human musicians. But to date it has not been conquered by Automatic music transcription (AMT) systems. The symbolic representation produced by AMT is needed for content-based music retrieval and music analysis tasks, such as melody extraction, music summarization[1] and rhythm tracking. AMT can also aid musicologist in analyzing music that has never been written down, such as improvised or ethnical music.

Conventional AMT systems are based on a wide range of different technologies. However, all of them have to deal with two problems: the onset detection and the fundamental frequency (F0) estimation of music notes. Wherein the detected onsets are used to divide the input signal into several audio segments, and then the pitches obtained in each audio segment are then merged into one whole transcription.

State-of-the-art onset detection algorithms typically employ band-wise processing. Scheirer[2] was the first to clearly point out that an onset detection algorithm should follow the human auditory system by treating frequency bands separately and combining results in the end. Goto[3] utilized the sudden energy changes in seven frequency ranges to detect onsets. Klapuri[4] divided the signal into 21 frequency bands and then used amplitude envelopes to find onsets across these bands. Duxbury[5] proposed a hybird multi-band onset detection approach, using an energy-based detector to find hard onsets in the upper bands, and a frequency based distance measure in the lower bands to improve the detection of soft onsets. For more onset detection methods, readers can refer to Ref.[6].

The F0 estimation algorithms can be divided into mono- and multi-F0 ones according to the number of concurrent music notes. For mono-F0 estimation, solutions have matured with many well-understood algorithms, which are largely unsuitable for multi-F0 estimation. Multi-F0 estimation is one of the most complex and still outstanding tasks to be solved, as there exist many musical instruments, each having a unique characteristic temporal and spectral structure. An additional problem stems from the fact that western music is based on harmonic relations, which give rise to spectral overlapping and even complete masking of certain notes.

Attempts towards polyphonic AMT date back to the 1970's, with the pioneering work of Moorer from Standford University[7]. Although limited to duets, it has inspired many subsequent works. One of the continuing effort results was described in Ref.[8], introducing the constant-Q discrete Fourier transform for the multi-resolution audio analysis. Research at M.I.T. began with Stautner's work[9], which used a filterbank analysis based upon 1/3 octave filters, mirroring the audi-

*Manuscript Received Nov. 14, 2013; Accepted Feb. 21, 2014. This work is partially supported by the National Natural Science Foundation of China (No.10925419, No.90920302, No.61072124, No.11074275, No.11161140319, No.91120001, No.61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA06030100, No.XDA06030500), the National 863 Program (No.2012AA012503), and the CAS Priority Deployment Project (No.KGZD-EW-103-2).

DOI:10.1049/cje.2015.07.027

tory system. Then F0s were captured using the peaks in a constructed pitch periodogram. Another approach, presented by Martin[10], is based on blackboard frameworks, integrating front-end based on sinusoidal analysis with musical knowledge.

Since 2005, when the first MIREX competition[11] took place, varieties of AMT algorithms have been proposed and compared against worldwide accepted test cases. These approaches are based on techniques such as: pitch trajectory analysis, harmonic clustering, bispectral analysis[12], Nonnegative matrix factorization[13], and Hidden Markov model[14]. For an overview of AMT approaches, readers can refer to Ref.[15].

Despite significant progress in current AMT research, there still exists no end-user application that can accurately and reliably transcribe polyphonic music. Even the most recent systems' performance is still clearly below a human expert's. As pointed out in Ref.[15], two promising directions could be considered to improve the AMT performance. One is restricting the employed instrument models to specific types. The other is to fuse information across the aspects of music.

In the first improvement direction, several examples of instrument-specific transcription can be found, such as Ref.[16] for violin, Ref.[17] for bell, Ref.[18] for tabla, Ref.[19] for guitar, and Ref.[20] for piano. Depending on the sound production mechanism of these instruments, different sets of instrument-specific parameters and constraints are defined in these AMT approaches.

In the second improvement direction, several multimodal-based works have been proposed. In Ref.[21], an application of independent component analysis was presented to extract audiovisual features from video streams, giving a simplified musical example of fingers on piano keyboard. In Ref.[22], a drum AMT system was described exploiting both the video and audio modalities. Several early- and late-fusion techniques were evaluated on drum-solos and showed that feature-level fusion by simple concatenation of audio and video features can achieve significant improvements compared to either of the monomodal systems. Other multimodal-based AMTs can also be found, such as Ref.[23] for violin, and Ref.[24] for guitar.

In this paper, we choose the piano as the single instrument, which is one of the instruments where the problems due to polyphony are the most challenging. Also, there exists a large corpus of solo piano music that can be used for testing and evaluation. Since 2007, the evaluation of AMT performance on piano subtask has been considered in MIREX competition and the test dataset remains unchanged. Fig.1 gives the best results for the note tracking task of piano subset based on onset only over the past years. It is worth mentioning that the best system was proposed by Zhou in 2008[25], which has gone unimproved since 2009 and is employed here as the baseline system to evaluate our proposed method's performance.

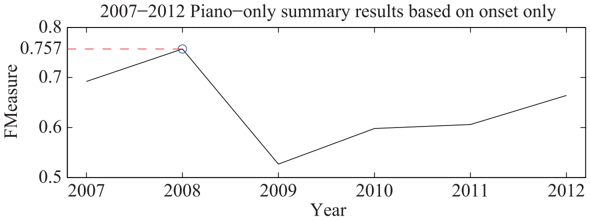

Fig. 1. 2007-2012 piano-only summary results based on onset only in MIREX

Motivated by the observation that the positions of the pianist's hands are associated with the piano notes played, we enhance the piano AMT performance by fusing detection results from multi-media streams. In our system, a microphone is fixed beside the piano and a video camera is mounted above the piano with its field of view covering the whole keyboard. As shown in Fig.2, while the pianist is playing, the audio stream is recorded and the camera captures the video stream simultaneously. Then the audio processing unit detailed in Sections III and IV extracts the preliminary transcription result from the audio stream, and the video processing unit detailed in Section V extracts the notes' range by detecting the hands' positions from the video stream. Subsequently, we use the notes' range to cancel the wrong pitch estimations for more accurate results. Here, we only concentrate on the two most important aspects of piano notes, their pitches and onset times.

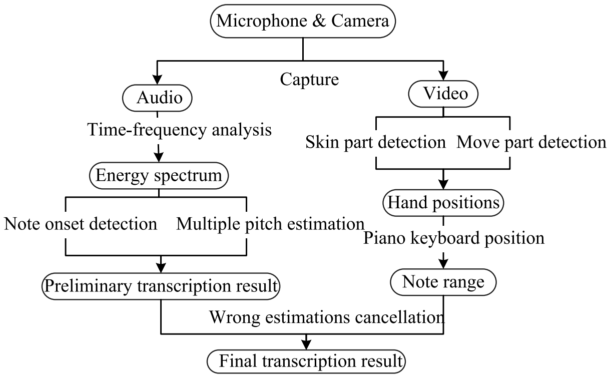

Fig. 2. The overall diagram of the proposed system

The remainder of this paper is as follows: Section II describes the time-frequency representation and Section III details the proposed onset detection. In Section IV, multi-F0 estimation and preliminary transcription method are presented; Section V details the hand recognition and the audio-visual fusion method. Finally, we evaluate the performance of our system in Section VI and draw conclusions in Section VII.

II. Resonator Time-Frequency Image

Here, we employ a computationally efficient time-frequency representation named Resonator time-frequency image (RTFI), which was proposed by Zhou in Ref.[26] and has been proved very fit for music signal analysis in Ref.[27]. It selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis as Eq.(1).

$$\begin{aligned} RTFI(t,\omega) &= s(t) * I_R(t,\omega) \\ &= r(\omega)\int_0^t s(\tau)\mathrm{e}^{r(\omega)(\tau-t)}\mathrm{e}^{-\mathrm{j}\omega(\tau-t)}d\tau \end{aligned} \qquad (1)$$

where $I_R(t,\omega) = r(\omega)\mathrm{e}^{(-r(\omega)+\mathrm{j}\omega)t}$, $t > 0$, denotes the impulse response of the first-order complex resonator filter with oscillation frequency ω. The factor $r(\omega)$ is used to normalized the gain of the frequency response when the resonator filter's input frequency is the oscillation frequency. It also determines the Equivalent rectangular bandwidth (ERB) of the implementing filter (*i.e.*, the frequency resolution of time-frequency analysis). According to the discussion in Ref.[26], the ERB can be approximated as follows in Eq.(2):

$$ERB(\omega) = d + c\omega \approx r(\omega) \cdot \pi \tag{2}$$

$$r(\omega) \approx ERB(\omega)/\pi = (d + c\omega)/\pi \tag{3}$$

where $d + c > 0$, $c \geq 0$ and $d \geq 0$. The commonly used frequency resolutions for music analysis are the specific cases of the Eq.(3), and different time-frequency resolutions can be selected by simply setting d and c.

To reduce the memory usage of storing the RTFI values, the RTFI is separated into time frames, and a RTFI Average energy spectrum (AES) is calculated in each frame as Eq.(4):

$$AES(k, \omega_m) = db(\frac{1}{M} \sum_{n=(k-1)\times M+1}^{k\times M} |RTFI(n, \omega_m)|^2) \tag{4}$$

where k denotes the frame index, $db(\cdot)$ converts the value to decibels, and M is an integer equal to the number of samples in each frame. Thus the ratio of M to the sampling rate is the duration time of each frame. $RTFI(n, \omega_m)$ represents the value of the discrete RTFI at sampling point n and frequency ω_m. The detailed description of the discrete RTFI can be found in Ref.[26]. The AES is then used as the input for the audio-only part of our piano AMT system.

In this paper, we employed the constant-Q RTFI ($d = 0$ and $c = 0.0058$), as the inter-harmonic spacings are the same for any periodic sounds. Piano recordings sampled at 44.1 kHz, 16-bit are considered as the inputs. The time interval between two successive frames is set to 10ms, correspondingly the number of samples in each frame is 441. Just as the parameter setting in Ref.[27], we use 10 filters to cover the frequency band of one semitone. And the number of bins per octave is 120, for the reason that each octave has 12 semitones. A total of 890 filters are used to cover the analyzed frequency range for the entire 88 music notes of piano (MIDI = 21–108), extending from 25.96Hz (MIDI = 20) to 4.43kHz (MIDI = 109).

III. Proposed Onset Detection Method

1. Equal-loudness pre-processing

With the consideration that the human auditory system reacts with different sensitivities in the different frequency bands,, we first pre-process the AES following the Robinson and Dadson equal-loudness contour[28] before onset detection. This equal-loudness contour has been standardized in the international standard ISO-226, which provides equal-loudness contours limited to 29 frequency bins and we only use the one corresponding to 70dB relative to the reference amplitude for 16-bit audio files[4]. The equal-loudness contour of 890 frequency bins is obtained by cubic-spline interpolation in the logarithmic frequency scale.

Here we refer to this equal-loudness contour as $ELC(\omega_m)$ in dB. Thus, the Equal-loudness AES (EAES) can be calculated as follows in Eq.(5):

$$EAES(k, \omega_m) = AES(k, \omega_m) - ELC(\omega_m) \tag{5}$$

where ω_m represents the angle frequency of the mth frequency bin and k represents the index of the frame.

2. Matched filtering

In signal processing, matched filtering is obtained by correlating a template signal, with an unknown signal to detect the presence of the template in the unknown signal[29]. This is equivalent to convolving the unknown signal with a conjugated time-reversed version of the template. For a discrete-time system, this processing can be expressed as Eq.(6).

$$y[n] = \sum_{k=-\infty}^{\infty} h[n-k]x[k] \tag{6}$$

where $x[n]$ is the observed signal, $h[n]$ is a linear matched filter, and $y[n]$ is the output response. The intuition behind this relies on correlating the observed signal with a filter that is parallel with the signal, maximizing the inner product.

As a result, to detect the onsets of notes, we just need to seek a filter h to maximize the onset parts in the output y, which is the inner product of h and the observed array x.

3. Onset detection method

The overview of our onset detection method is presented in Fig.3. We utilize the band-wise processing principle as motivated above. First, the overall AES of the signal is normalized to EAES at 70dB level tracing the equal-loudness contour. Then the EAES is divided into 89 non-overlapped bands, each covering one octave. At each band, we detect onset components and determine their time and intensity. In final phase, the onset components are combined to yield onsets.

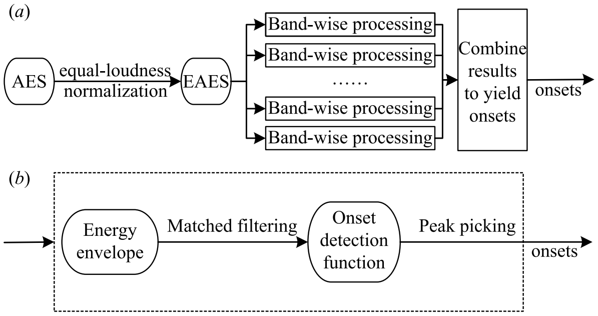

Fig. 3. (a) Overview of onset detection method; (b) Processing at each frequency band

Firstly, we calculate the mean value array of EAES for each band as follows in Eq.(7):

$$MA(k, \omega_m) = \frac{1}{N_m} \sum_{n=-\frac{N_m}{2}}^{\frac{N_m}{2}} EAES(k, \omega_{m+n}) \tag{7}$$

where $MA(k, \omega_m)$ denotes the mean value of the mth frequency band in the kth frame and N_m represents the number of frequency bins in the mth band.

The frame number in each onset region is fixed to 12 just as in Ref.[30]. From the statistics of preliminary experiments, it can be observed that the envelopes of the mean values in the onset regions can be approximated as Eq.(8).

$$\begin{aligned} \mu \times [&-9.6258, -6.9773, -5.6138, -4.6522, \\ &-3.968, -4.7485, -1.6418, 9.9056, \\ &13,3167, 11.7671, 9.3814, 9.6014] \end{aligned} \tag{8}$$

where μ is some constant. Therefore, according to the definition of matched filter, we use an envelope matched filter as follows in Eq.(9):

$$MF[12] = [9.6014, 9.3814, 11.7671, 13.3167, 9.9056, -1.6418, -4.7485, -3.968, -4.6522, -5.6138, -6.9773, -9.6258] \tag{9}$$

Afterwards, we calculate the convolution result of the mean value array (obtained in Eq.(7)) of each band using this specific matched filter as follows in Eq.(10):

$$CR(k, \omega_m) = \sum_{\tau=1}^{12}(MA(k-\tau, \omega_m) \times MF[\tau]) \tag{10}$$

where CR denotes the convolution result and k is the frame index. Finally, onsets in each band are detected by simply peak picking on CR, and a peak is also treated as non-onset if the corresponding MA value is below –70 or the CR value is below 150, determined by previous experiments.

The last phrase of the proposed onset detection method is to combine all the onsets detected from separate bands to yield onsets of the overall signal. First the onset components from different bands are all sorted in time order, and regarded as onset candidates hereafter. Then each onset candidate is assigned a loudness value, which is calculated by summing up the EAES values in the corresponding frequency band. Then we drop out candidates that are too close ($<$ 100ms) to a louder candidate. Among equally loud but too close candidates, the middle one is chosen and the others are abandoned. The remaining onset candidates are accepted as true ones. Fig.4 shows the details of detecting the onsets in the 32th band related to the piano MIDI number 52.

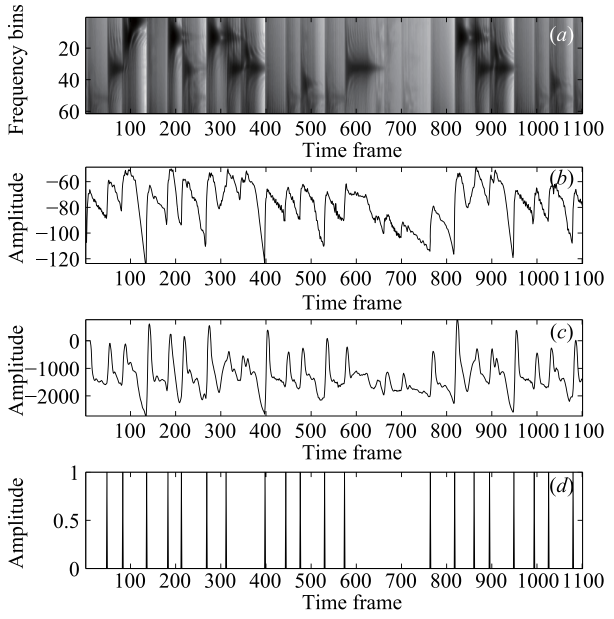

Fig. 4. The details of detecting the onsets in the 32th band. (*a*) EAES of the 32th band; (*b*) Mean value array of the EAES; (*c*) The convolution result; (*d*) The onsets detected in the 32th band

IV. Multiple Pitch Estimation

To estimate the multiple pitches in each frame, the input AES is first transformed into the Pitch energy spectrum (PES) based on the harmonic grouping principle, and then into the Relative pitch energy spectrum (RPES), which represents the distribution of the pitches existing. As the frequency indexes are in the logarithmic scale and each octave is represented by 120 frequency bins as mentioned in Sec. II, the deviations[26] between the harmonics and the pitches can be approximated as follows in Eq.(11):

$$HD[8] = [0, 120, 190, 240, 279, 310, 337, 360] \tag{11}$$

Accordingly, the PES and RPES can be easily approximated in the logarithmic scale by the following calculations in Eqs.(12) and (13):

$$PES(k, \omega_m) = \frac{1}{N_h}\sum_{i=1}^{N_h} AES(k, \omega_{m+HD[i]}) \tag{12}$$

where k is the frame index, and N_h denotes the number of harmonic components considered to be the pitch m's. Supposing that the energy of each music note mainly distributes over the first 4 partials, we set the N_h to 4 in our experiment.

$$RPES(k,\omega_m) = PES(k,\omega_m) - \frac{1}{2N_r+1}\sum_{i=m-N_r}^{m+N_r} PES(k,\omega_i) \tag{13}$$

where N_r denotes the number of frequency bins around the frequency bin m used to calculate the relative value. In this paper, it is set to 30.

Afterwards, the AES is transformed to the Relative energy spectrum (RES) according to the following Eq.(14):

$$RES(k, \omega_m) = AES(k, \omega_m) - Mean(AES(k)) \tag{14}$$

where $Mean(AES(k))$ denotes the mean value of the kth frame in AES. The peaks of RES can be the signs of the harmonic components' existence.

Based on the RPES and RES, we estimate the candidate pitches in each frame following the rules detailed in Ref.[25].

Although these rules have improved the performance of the multi-F0 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. If we can get the range of notes played at each moment, these extra pitches would be canceled by simply judging whether in the limitation range. As the positions of the pianist's hands have the most direct connection with the notes played, we consider the recognition of pianist's hands' positions as a method to rectify the preliminary transcription results.

V. Hand Recognition

1. The equipment setup

This method is achieved with the installation of one camera above a piano, so as to fully record hand movements on the keyboard. Considering that the recognition algorithm is applied in every video frame, we prefer the 640 × 480 resolution of the camera and the frame rate is defined as 25 fps in order to achieve a detailed illustration of hand movements.

Besides, we choose the white light which also plays 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 import the video signal into the computer and use the Open source computer vision library (OpenCV)[31] to process it.

2. Hand recognition algorithm

In our system, we combine two methods to detect the hands: the motion detection and the skin color detection.

Firstly, we delineate the piano keyboard and only the part of the image containing the piano keyboard will be used in further processing. The background image without hands is stored in gray scale at the very beginning.

As the playing session starts, the comparison between the foreground image and the background image in the gray scale is performed to detect the motion part. Then, we can use the skin color detection to isolate the hands from the motion parts.

It is observed that the 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. Therefore, it is generally a good idea to perform image analysis in a non-linearly transformed color space that separates brightness values of the signal[32]. The YCrCb and the HSV color space are both employed, in which the psychophysical thresholds of human color perception are preset. The skin color detection algorithms in these two color space are detailed as follows in Algorithm 1 and Algorithm 2. The skin color ranges used here can be found in Ref.[33] and have been adjusted according to our experimental results. If a pixel p_{ij} is a skin pixel both in the HSV and the YCrCb color space, we confirm it as a skin pixel.

Algorithm 1 Skin color detection in HSV color space

```
// $p_{ij} = (H_{ij}, S_{ij}, V_{ij})$
for $i \leftarrow 480$
    for $j \leftarrow 640$
        if $V_{ij} \in (80, 255)$
            if $S_{ij} \in (0, 191) \& H_{ij} \in (0, 18)$
                $p_{ij}$ is a skin pixel.
            else if $S_{ij} \in (0, 178) \& H_{ij} \in (135, 180)$
                $p_{ij}$ is a skin pixel.
            else
                $p_{ij}$ is not a skin pixel.
        else
            $p_{ij}$ is not a skin pixel.
```

Algorithm 2 Skin color detection in YCrCb color space

```
// $p_{ij} = (Y_{ij}, Cr_{ij}, Cb_{ij})$
for $i \leftarrow 480$
    for $j \leftarrow 640$
        if $Cr_{ij} \in (133, 173) \& Cb_{ij} \in (77, 128)$
            $p_{ij}$ is a skin pixel.
        else
            $p_{ij}$ is not a skin pixel.
```

The result of this procedure is a binary exportation including the hand contour only. Finally, morphology skinning techniques based on dilation and erosion are carried out on the result. Fig.5 shows the result of hand tracking (as boxes circumscribed around the hands).

Finally, using the hands' relative positions on the keyboard, we determine the range of the notes played at each moments according to the distribution of the notes on the piano keyboard, and then cancel the incorrect notes, which are not in the range, to improve the estimation accuracy.

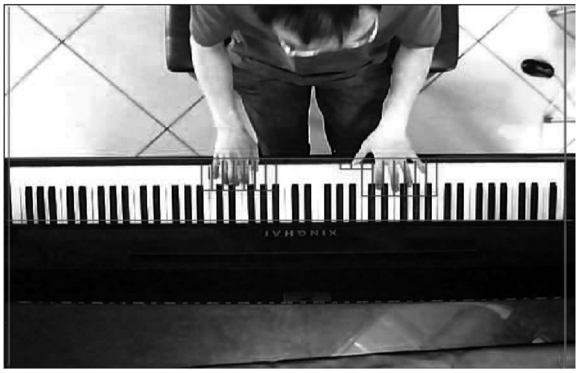

Fig. 5. The hands' positions detected on the keyboard

VI. System Performance Evaluations

1. Experiment databases

We carried out experiments on two databases, and compared the results with the baseline system mentioned in Section1.

The first database is the MAPS database[34]. Here, we chose 9 pieces sampled at 44.1kHz, 16-bit, from the MUS (pieces of real piano music) subset as the test set, which consists of stereo recordings under 9 different recording conditions named as "AkPnBcht", "AkPnBsdf", "AkPnCGdD", "AkPnStgb", "ENSTDkAm", "ENSTDkCl", "SptkBGAm", "SptkBGCl", and "StbgTGd2". And we use "1–9" to stand for the conditions, respectively. Details of these conditions can be found in Ref.[34]. The note locations and durations have been adjusted by the creator of the MIDI database and there are totally 11891 piano notes and 5440 onsets in this test set.

The second database is composed of 4 piano videos and corresponding audio recordings sampled at 44.1kHz, 16-bit to evaluate the performance advancement of the proposed audiovisual fusion method. The reference annotations are also adjusted by the pianist, and there are totally 10925 video frames, 2695 piano notes and 1218 onsets in this test set.

2. Evaluation Metrics

To evaluate the performance of onset detection, the detected onsets must be compared with the reference ones. For a onset detected at time t, if there is a reference one within a tolerance time-window $[t - 50\text{ms}, t + 50\text{ms}]$, it is considered to be True positive (TP), otherwise, it is False positive (FP). The reference onsets outside all the tolerance windows are counted as False negative (FN).

When evaluating the performance of the note-tracking, a result note is considered to be correct (or TP) only if the note is same with a reference one and the corresponding onset is in the tolerance time-window, otherwise, it is considered as FP. All the undetected reference notes are counted as FN.

Here, we use the Precision, Recall and F-measure to summarize the results, which can be expressed as Eq.(15):

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad Recall = \frac{N_{TP}}{N_{TP} + N_{FN}}$$
$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{15}$$

where N_{TP} is the number of TP onsets (notes), N_{FP} is the number of FP ones, and N_{FN} is the number of FN ones.

3. Audio-only performance comparison

Table 1 details the comparison results of onset detection in 9 different recording conditions, in which we can see that the proposed onset detection method lead to higher Recall in all conditions. Fig.6(a) gives the onset detection performance comparison over the whole test set. It should be noted that the proposed onset detection method achieved 94.68% F-measure, which outperformed the baseline system by 4.38% F-measure relatively. This is mainly because our method considers the onsets in all frequency bands related to piano's pitches, which results in 9.32% relatively higher Recall. It should also be pointed out that a little decrease in the precision is caused by making a trade-off for higher F-measure and Recall.

Table 2 details the comparison results of note-tracking in all the conditions and Fig.6(b) gives the note-tracking performance comparison over the whole test set. We can tell that the audio-only performance tendency kept approximate and unanimous with the onset detection's, which was improved by 10.58% F-measure, especially 17.75% Recall.

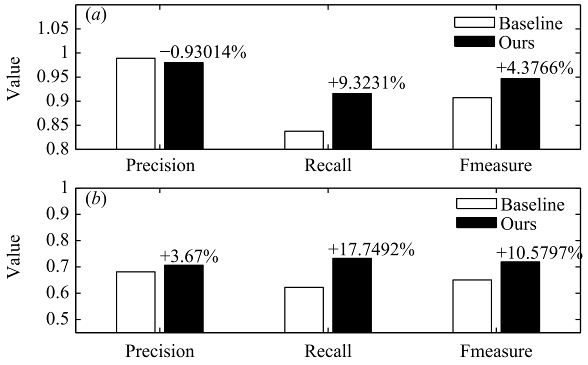

Fig. 6. Audio-only note-tracking performance comparison. (a) Audio-only onset detection performance comparison; (b) audio-only note-tracking performance comparison

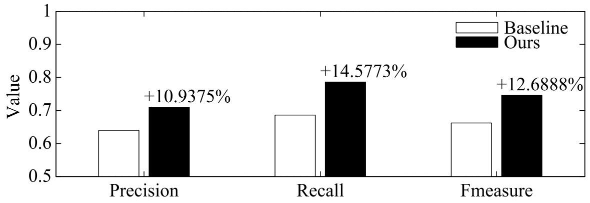

Fig. 7. Overall transcription performance comparison

Table 1. Performance comparisons of audio-only onset detection in 9 different recording conditions

Conditions		1	2	3	4	5	6	7	8	9
Precision	Baseline	0.999	1.000	0.997	1.000	0.955	0.951	1.000	1.000	0.998
	Proposed	0.989	0.964	0.980	0.995	0.939	0.946	0.987	0.998	1.000
Recall	Baseline	0.884	0.910	0.949	0.829	0.869	0.829	0.678	0.850	0.723
	Proposed	0.954	0.977	0.972	0.957	0.963	0.847	0.914	0.888	0.768
F-measure	Baseline	0.938	0.953	0.973	0.907	0.910	0.886	0.808	0.919	0.838
	Proposed	0.972	0.971	0.976	0.975	0.951	0.894	0.949	0.940	0.869

Table 2. Performance comparisons of audio-only note tracking in 9 different recording conditions

Conditions		1	2	3	4	5	6	7	8	9
Precision	Baseline	0.533	0.742	0.732	0.627	0.580	0.750	0.812	0.818	0.826
	Proposed	0.673	0.705	0.765	0.614	0.616	0.729	0.832	0.840	0.804
Recall	Baseline	0.598	0.759	0.759	0.630	0.521	0.659	0.479	0.630	0.609
	Proposed	0.818	0.801	0.843	0.821	0.601	0.671	0.732	0.695	0.626
F-measure	Baseline	0.563	0.751	0.745	0.628	0.549	0.702	0.603	0.712	0.701
	Proposed	0.739	0.750	0.802	0.702	0.608	0.699	0.779	0.761	0.704

4. Overall performance comparison

The hand tracking performance is evaluated by counting the number of video frames in which all the piano keys covered by hands are completely detected. The detection accuracy is up to 99.3%, as 10848 video frames are correct without keys missing, which means that the hand tracking method is sufficient for notes' range detection.

Fig.7 illustrates the overall performance comparison of the baseline system and ours. Clearly we can see that the overall performance was improved with about 12.69% F-measure increase and about 19.44% error reduction with the fusion of audio-visual features. This proves that the multi-modal fusion of audio and video cues is very promising in improving multi-F0 estimation accuracy and transcription performance, which results from efficiently canceling of incorrect estimations by tracking the hands' positions.

VII. Conclusion

AMT is a rapidly developing research area where many approaches have been investigated. However, the performance of current systems is still not sufficient for applications which require a great degree of accuracy. A promising direction could be to adapt the instruments' specific parameters. Many existing piano AMT methods are not so much tailored to piano music. In this paper, we build an audio-visual fusion based AMT system for polyphonic piano music, which employs a onset detection method using a specific matched filter on multiple frequency bands, multi-F0 estimation based harmonic rules, and a note range limitation method by tracking the pianist's hands' positions. Compared with the best piano AMT system in MIREX, experimental results clearly verify our initial hypothesis that performance can be improved significantly by fusing audio and visual cues. However, there are still many aspects that deserve in depth investigations to further enhance the transcription performance.

Piano AMT has many applications such as piano education tutoring, which requires very high transcription speed and accuracy. We have taken the first step to enhance piano AMT by fusing multimedia streams and future research will focus on analyzing specific harmonic structure of piano pitches.

References

[1] X. Shao, M.C. Maddage, C. Xu, and M.S. Kankanhalli, "Automatic music summarization based on music structure analysis", *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings*, Philadelphia, Pennsylvania, USA, pp.1169–1172, 2005.

[2] E.D. Scheirer, "Tempo and beat analysis of acoustic musical signals", *The Journal of the Acoustical Society of America*, Vol.103, No.1, pp.588–601, 1998.

[3] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds", *Journal of New Music Research*, Vol.30, No.2, pp.159–171, 2001.

[4] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge", *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings*, Phoenix, Arizona, USA, pp.3089–3092, 1999.

[5] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection", *Proc. Digital Audio Effects Conf. (DAFX,02)*, Hamburg, Germany, pp.33–38, 2002.

[6] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A tutorial on onset detection in music signals", *IEEE Transactions on Speech and Audio Processing*, Vol.13, No.5, pp.1035–1047, 2005.

[7] J.A. Moorer, "On the transcription of musical sound by computer", *Computer Music Journal*, Vol.1, No.4, pp.32–38, 1977.

[8] C. Chafe, D. Jaffe, K. Kashima, B. Mont-Reynaud and J. Smith, *Techniques for Note Identification in Polyphonic Music*, CCRMA, Department of Music, Stanford University, 1985.

[9] J.P. Stautner, "Analysis and synthesis of music using the auditory transform", *Ph.D. Thesis*, Massachusetts Institute of Technology, 1983.

[10] K.D. Martin, "A blackboard system for automatic transcription of simple polyphonic music", *Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report*, pp.385, 1996.

[11] Music information retrieval evaluation exchange (MIREX), available at *http://music-ir.org/mirexwiki/*, 2008.

[12] F. Argenti, P. Nesi and G. Pantaleo, "Automatic transcription of polyphonic music based on the constant-q bispectral analysis", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.19, No.6, pp.1610–1630, 2011.

[13] B. Niedermayer, "Non-negative matrix division for the automatic transcription of polyphonic music", *Proc. ISMIR*, Drexel University, Philadelphia, PA, USA, pp.544–549, 2008.

[14] V. Emiya, R. Badeau and B. David, "Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches", *Proc. Eur. Conf. Sig. Proces.(EUSIPCO)*, Lausanne, Switzerland, 2008.

[15] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff and A. Klapuri, "Automatic music transcription: Challenges and future directions", *Journal of Intelligent Information Systems*, Vol.41, No.3, pp.407–434, 2013.

[16] I. Barbancho, C. de la Bandera, A.M. Barbancho and L.J. Tardon, "Transcription and expressiveness detection system for violin music", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, pp.189–192, 2009.

[17] M. Marolt, "Automatic transcription of bell chiming recordings", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.20, No.3, pp.844–853, 2012.

[18] O. Gillet and G. Richard, "Automatic labelling of tabla signals", *Proc. ISMIR*, Baltimore, Maryland, USA, 2003.

[19] A.M. Barbancho, A. Klapuri, L.J. Tardon and I. Barbancho, "Automatic transcription of guitar chords and fingering from audio", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.20, No.3, pp.915–921, 2012.

[20] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music", *IEEE Transactions on Multimedia*, Vol.6, No.3, pp.439–449, 2004.

[21] P. Smaragdis and M. Casey, "Audio/visual independent components", *Proc. ICA*, Nara, Japan, pp.709–714, 2003.

[22] O. Gillet and G. Richard, "Automatic transcription of drum sequences using audiovisual features", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Pennsylvania, USA, pp.iii–205, 2005.

[23] Y. Wang, B. Zhang and O. Schleusing, "Educational violin transcription by fusing multimedia streams", *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*, Augsburg, Bavaria, Germany, pp.57–66, 2007.

[24] M. Paleari, B. Huet, A. Schutz and D. Slock, "A multimodal approach to music transcription", *IEEE International Conference on Image Processing*, San Diego, California, USA, pp.93–96, 2008.

[25] R. Zhou and J.D. Reiss, "A real-time polyphonic music transcription system", *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, Philadelphia, Pennsylvania, USA, 2008.

[26] R. Zhou, "Feature extraction of musical content for automatic music transcription", *Ph.D. Thesis*, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2006.

[27] E. Benetos, "Automatic transcription of polyphonic music exploiting temporal evolution", *Ph.D. Thesis*, Queen Mary University of London, 2012.

[28] D.W. Robinson and R.S. Dadson, "A re-determination of the equal-loudness relations for pure tones", *British Journal of Applied Physics*, Vol.7, No.5, pp.166, 1956.

[29] G.L. Turin, "An introduction to matched filters", *IRE Transactions on Information Theory*, Vol.6, No.3, pp.311–329, 1960.

[30] J.J. Ding, C.J. Tseng, C.M. Hu and T. Hsien, "Improved onset detection algorithm based on fractional power envelope match filter", *European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, pp.709–713, 2011.

[31] "Open source computer vision library (Opencv)", available at *http://opencv.org*, 2013.

[32] W. Westerman, "Hand tracking, finger identification, and chordic manipulation on a multi-touch surface", *Ph.D. Thesis*, University of Delaware, 1999.

[33] D. Chai and K.N. Ngan, "Face segmentation using skin-color map in videophone applications", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.9, No.4, pp.551–564, 1999.

[34] 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*, Vol.18, No.6, pp.1643–1654, 2010.

WAN Yulong received his B.S. degree from Electronic Engineering Department of Peking University in 2009. Currently he is a Ph.D. candidate in the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests include voice activity detection, speaker diarization and music signal processing. (Email: wanyulong@hccl.ioa.ac.cn)

WANG Xianliang received B.E degree from College of information science and technology, Nankai University in 2010. Now he is a Ph.D. candidate in the Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS. His research interests include speaker recognition and language recognition. (Email: wangxianliang@hccl.ioa.ac.cn)

ZHOU Ruohua received his B.S. degree from the Electronics Engineering Department, Beijing Institute of Technology in 1994, the M.S. degree of engineering in microelectronics and semiconductor devices from Microelectronics R&D Center, CAS, in 1997, and the Ph.D. degree from the Signal Processing Laboratory, Swiss Federal Institute of Technology in 2006. Currently he is a professor at the Key Laboratory of Speech Acoustics and Content Understanding. His research interests include language/speaker recognition, and music signal processing. (Email: zhouruohua@hccl.ioa.ac.cn)

YAN Yonghong received his B.E. from Tsinghua University in 1990, and the Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as assistant professor (1995), associate professor (1998) and associate director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired Human Computer Interface Research Council, worked as principal engineer of Microprocessor Research Lab and Director of Intel China Research Center. Currently he is a professor and director of the Key Laboratory of Speech Acoustics and Content Understanding. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. (Email: yanyonghong@hccl.ioa.ac.cn)