Automatic Music Transcription using Audio-Visual Fusion for Violin Practice in Home Environment

Bingjun Zhang and Ye Wang
School of Computing, National University of Singapore
{bingjun, wangye}@comp.nus.edu.sg

ABSTRACT

Violin practice in a home environment, where there is often no teacher available, can benefit from automatic music transcription to provide feedback to the student. This paper describes a high performance violin transcription system with three main contributions. First, as onset detection is an important but challenging task for automatic transcription of pitched non-percussive music, such as from the violin, we propose an effective audio-only onset detection approach based on supervised learning. The proposed approach outperforms the state-of-the-art methods substantially. Second, we introduce the visual modality, i.e., bowing and fingering of the violin playing, to infer onsets, thus enhancing the audio-only onset detection. We devise automatic and real-time video processing algorithms to extract indicative features of onsets from bowing and fingering videos. Third, we evaluate state-of-the-art multimodal fusion techniques to fuse audio and visual modalities and show this improves onset detection and transcription performance significantly. The audio-visual fusion based violin transcription system provides more accurate transcribed results as learning feedback even in acoustically inferior environments. With efficient and fully automatic audio-visual analysis components, the system can be easily deployed in a home environment.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing, Systems; I.4.8 [**Scene Analysis**]: Motion, Tracking, Sensor fusion

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Music Transcription, Onset Detection, Hand Tracking, Fingering Analysis, Multimodal Fusion

1. INTRODUCTION

Automatic music transcription (AMT), which converts music audio into MIDI (piano-roll) notation, can provide effective feedback when students practice violin at home [27]. AMT includes two basic steps: onset detection and pitch estimation. Onset detection finds the note boundaries (onsets) in time domain to segment a violin piece into individual notes. Subsequently, pitch estimation estimates pitch values of each note segment. In violin music, most of the notes are monophonic and very few are double-stops (notes with two pitches) or triple-stops (notes with three pitches). Due to this characteristic, pitch estimation in violin music is largely a monophonic pitch estimation problem, which is considered solved. However, because of the soft transient around note onsets in pitched non-percussive (PNP) sounds, such as from the violin, onset detection is recognized as a difficult task. State-of-the-art audio-only onset detection approaches in PNP sounds reveal poor performance [8].

In this paper, we address the onset detection problem in violin music and show how to build a high performance audio-visual music transcription system to assist violin practice at home. The main contributions include the following:

- We propose an audio-only onset detection approach based on supervised learning. Gaussian Mixture Models (GMM) are used to classify onset and non-onset frames based on Mel-Frequency Cepstral Coefficients (MFCCs). The proposed onset detection approach outperforms the state-of-the-art methods [4, 9, 14, 18] by about 10% F-measure in less noisy conditions.
- To enhance the audio-only onset detection, we introduce the visual modality of the violin playing, including bowing and fingering, to infer onsets. We devise real-time and fully automatic algorithms to extract indicative features of onsets from bowing and fingering videos captured in a home environment.
- We evaluate state-of-the-art multimodal fusion techniques, including feature level (early) fusion and decision level (late) fusion, to combine the audio-visual modalities for onset detection and violin transcription. For violin onset detection, the audio-visual fusion based approach outperforms the proposed audio-only approach by 5% to 18% F-measure in different noisy conditions. Thus, the overall transcription accuracy is improved by 14% to 20%.

The remainder of this paper is organized as follows. In Section 2, the system framework and methodology are outlined. Section 3 and 4 detail the audio and video processing

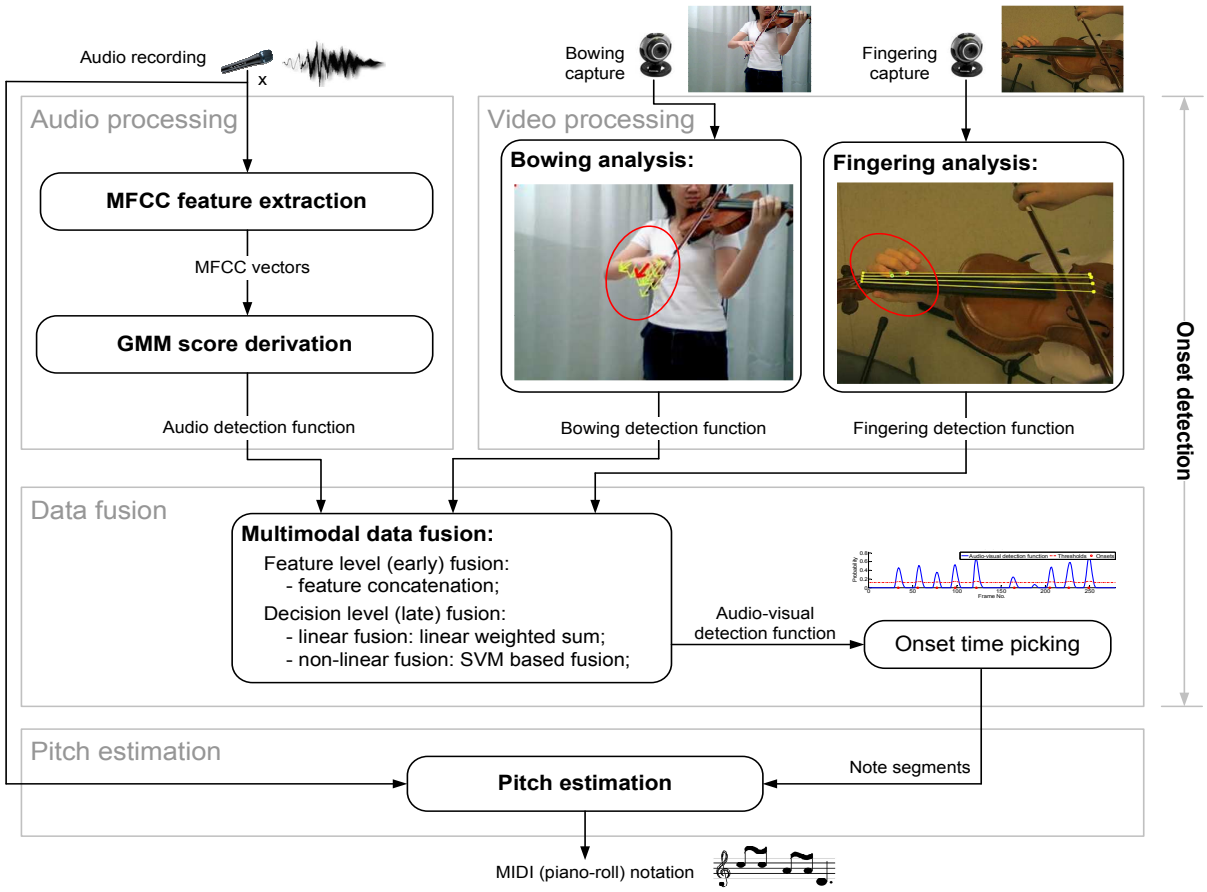


Figure 1: System diagram of audio-visual music transcription for violin practice at home.

components. We then discuss multimodal fusion techniques for audio-visual fusion in Section 5. Section 6 evaluates our system performance which is followed by comments on related work (Section 7) and a closing summary (Section 8).

2. SYSTEM DESCRIPTION

As illustrated in Fig. 1, when students practice a violin piece by following a reference notation, the audio-visual transcription system records the audio input stream by a microphone, and captures two video input streams by two ordinary webcams. Then its audio and video processing units extract indicative features of onsets (detection functions) from the input audio-visual streams. Subsequently, multimodal data fusion techniques are applied to fuse audio and visual modalities for more accurate onset detection. Pitch estimation is conducted at last to produce the MIDI (piano-roll) notation of the played violin music. The comparison of the transcribed results and the reference notation shows the violin students which notes are played correctly/wrongly. Our preliminary evaluations have shown that such feedback is important for beginning violin learners [27].

3. AUDIO PROCESSING

This section describes the audio-only violin transcription sub-system, which is to be integrated with the visual modality, described in Section 4. In the audio-only sub-system, we propose a supervised learning based onset detection method, which outperforms the state-of-the-art methods [4, 9, 14, 18] significantly. We employ a violin specific audio-only pitch estimator to conduct pitch estimation [18].

3.1 Audio-only Onset Detection

3.1.1 Existing Onset Detection Methods

As recognized in the literature, onset detection in violin sounds is a difficult problem, because of the soft transient around note boundaries in audio in terms of energy, pitch, etc. Existing methods for onset detection rely on certain characteristic features of the audio signal to derive a detection function, which is a one-dimensional function with peaks indicating sudden changes (onset times) in an audio signal. The detection function can also be viewed as decision scores at different time instances where larger values indicate higher probability of onsets. Onsets are detected as the local maxima of a detection function by a peak-picking algorithm [8]. Some features that have been used for deriving a detection function include pitch change [9], equal loudness change [14], phase change [4], spectrum correlation change [18]. Nevertheless, as observed in [8], those feature based onset detection methods produced poor accuracy for PNP sounds, e.g., violin music.

In [15], Lacoste and Eck proposed a supervised learning approach with Feed-Forward Neural Networks to classify onset and non-onset times based on raw spectrogram features. This approach is superior to existing feature based methods,

because it uses supervised learning to separate the distribution of audio spectra of onset and non-onset frames. It is capable of modeling general characteristics in various audio dynamics if enough training data is available. However, Lacoste and Eck employed raw spectrograms as audio features, which have a very high number of dimensions (about 800). According to the curse of dimensionality, exponentially more training data are needed to fully train a Feed-Forward Neural Network. What's worse, with high dimensional input features and many hidden neurons, training such a neural network and using it for classification are very time consuming. All these drawbacks prevent this method from being employed in practical applications.

3.1.2 *The Proposed Onset Detection Approach*

In this paper, we propose a supervised learning approach for onset detection by using Gaussian Mixture Models (GMM) to classify onset and non-onset frames based on Mel-Frequency Cepstral Coefficients (MFCCs).

MFCC features can reflect the difference between onset and non-onset frames of violin audio signals. MFCCs model the spectrum envelop in a perceptual and concise way [17]. As can be seen around onsets A in Fig. 2(a), the spectrograms around onset times are more noisy or less harmonic than the spectrograms within a note. This difference between onset and non-onset frames is reflected in the corresponding MFCCs. On the other hand, if the transient between certain two notes is noiseless (see onset B in Fig. 2(a)), derivatives of MFCCs still reflect the harmonic change at the note boundary, hence indicating the onset times. Clear difference between MFCC features of onset and non-onset times can be observed in Fig. 2(b), where MFCCs and their first, second order derivatives are drawn for each audio frame.

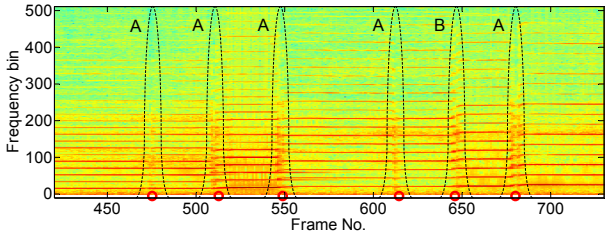
The probability of the onset or non-onset label, $l \in \{l_o, l_n\}$, for each frame can be modeled as a random variable drawn from a probability distribution of the MFCC features $\mathbf{f}_t$. As GMM is known to work well in modeling the distribution of MFCCs [22], we propose to use GMM to model the random variable of onset or non-onset label for each frame, which is defined as:

$$P^l(\mathbf{f}_t|\Theta^l) = \sum_{i=1}^{M^l} w_i^l p^l(\mathbf{f}_t|\mu_i^l, \mathbf{\Sigma}_i^l) \quad (1)$$

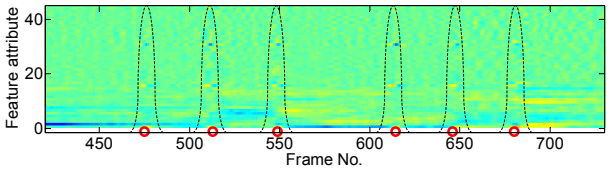
where $\mathbf{f}_t$ is MFCC feature vector at time t; $\Theta^l = \{w_i^l, \mu_i^l, \mathbf{\Sigma}_i^l \mid 1 \le i \le M^l\}$ is the parameter set for the GMM with label l; w_i^l, μ_i^l and $\mathbf{\Sigma}_i^l$ are the weight, mean vector and covariance matrix of each Gaussian mixture respectively; M^l is the number of mixtures; $P^l(\mathbf{f}_t|\Theta^l)$ is the probability that $\mathbf{f}_t$ is drawn from feature set with label l given a GMM characterized by the parameter set Θ^l.

In the system, we use Expectation-Maximization (EM) algorithm [21] to train two GMMs to profile the distribution of onset and non-onset features. To assemble the training data, we select features $\{\mathbf{f}_t|t_o - \tau \le t \le t_o + \tau\}$ around each onset frame at t_o to form the onset feature set F^{l_o}, and select the rest to form the non-onset feature set F^{l_n}.

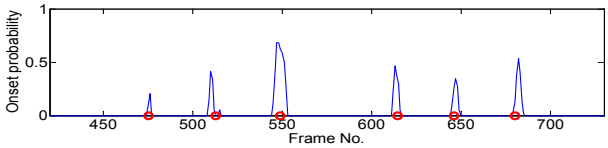
To derive a detection function based on onset and non-onset GMMs, a series of MFCC features in time with unknown labels, $F = \{\mathbf{f}_t|0 \le t \le T\}$, is extracted from the incoming audio piece with the same settings as the training MFCC features. The detection function from audio is then



(a) Spectrograms of violin audio signals



(b) MFCC features of violin audio signals



(c) MFCC and GMM based audio-only detection function

Figure 2: An onset detection approach by MFCCs and GMM. Onsets are human annotated as circles.

calculated as:

$$D_a(t) = Rect(P^{l_o}(\mathbf{f}_t|\Theta^{l_o}) - P^{l_n}(\mathbf{f}_t|\Theta^{l_n})) \quad (2)$$

where $Rect(x) = x$ if $x \ge 0$, and $Rect(x) = 0$ otherwise. As shown in Fig. 2(c), peaks in the audio detection function, $D_a(t)$, indicate onset times.

In our system, a Mel-scale filter bank with 81 filters is applied and 15 DCT coefficients are truncated as MFCCs. The first order and second order derivatives in time are appended to form a 45-dimensional MFCC feature vector for a frame. Based on experiments, the sampling rate of 22.05 kHz, frame length of 1024 and hop size of 341 samples for the audio signal are found optimal to extract MFCC features. Under this setting, the sampling rate of MFCCs or the audio detection function is about 64.7 Hz. A diagonal covariance matrix is used for each Gaussian mixture. The parameter settings, with τ as 2 frames, M^{l_o} and M^{l_n} as 256 mixtures, yields the best results for onset detection.

As shown in the experiments, this approach outperforms state-of-the-art onset detection methods. In addition, due to the low dimensional feature space of MFCCs (45 dimensions) and the efficient modeling approach by GMM, the proposed onset detection method is computationally effic-net, thus suitable for practical applications, such as the automatic violin transcription system built in this paper.

3.2 Audio-only Pitch Estimation

As pitch estimation in violin music is largely monophonic pitch estimation and is considered a solved problem, a violin specific audio-only pitch estimator developed in [18] and evaluated in [25] is employed in our system implementation. The overall accuracy is 95% in estimating pitch values of our database, described in Section 6.1.

4. VIDEO PROCESSING

In violin playing, note onsets are highly correlated with visual cues of the player, such as the reversals of bowing and finger press/release of a string [3, 25]. Therefore, it is natural to include bowing features (bow reversal moments) and fingering features (finger press and release moments) to improve audio-only onset detection, thus improving automatic violin transcription as a whole.

We employ two ordinary webcams (Microsoft VX3000) to capture bowing and fingering videos along with the audio recording, in 30 frames per second (fps), with the resolution 640×480. One camera is placed in front of the player on a tripod to capture a side view of the bowing. The other camera is placed above the violin body on a tripod (or fixed on the ceiling) to capture a birds eye view of the fingering from the violin neck to the bridge. It should be noted that the camera placement is not critical as long as the bowing/fingering is captured in the video. The movements of the player within certain limits do not degrade the system performance. In short, the two webcams can be easily set up at home, which maximizes the practicality of the system.

4.1 Bowing Analysis for Onset Detection

The right hand of the violin player holds the bow during playing, thus the hand motion reflects the bowing motion in terms of moving direction and speed. Therefore, a hand tracking algorithm is devised to obtain the sequence of the bow moving direction. Sudden changes (around 180 degrees) of the moving direction reflect bow reversal moments.

Hand tracking is achieved using Kalman filter framework [6] with measurements obtained by optical flow [23] and a skin-color Gaussian model [24]. Based on a prior database, the skin-color Gaussian model is pre-calculated in RGB color space as $\mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$, Eq. (3). The distance $dis(\mathbf{c}_{x,y})$ of a pixel at (x, y) with RGB color $\mathbf{c}_{x,y} = [r \quad g \quad b]$ to the skin-color Gaussian model is measured as Mahalanobis distance, Eq. (4). If $dis(\mathbf{c}_{x,y}) < \alpha$, the pixel is considered of skin color. The optimal value of α is found to be 20 in the experiments.

$$\begin{cases} \mu_c = [172 \ 104 \ 45] \\ \mathbf{\Sigma}_c = [764 \ 508 \ 180; \ 508 \ 359 \ 138; \ 180 \ 138 \ 91] \end{cases} \tag{3}$$

$$dis(\mathbf{c}_{x,y}) = (\mathbf{c}_{x,y} - \mu_c)\mathbf{\Sigma}_c^{-1}(\mathbf{c}_{x,y} - \mu_c)^T \tag{4}$$

Before hand tracking, global motion compensation is conducted by referring to the first frame to compensate the body translation of the player, which lessens the influence of the body movement on bowing features for onset detection.

A hand state at time t is defined as $\mathbf{h}_{t,t} = [x_{t,t} \ y_{t,t} \ d_{t,t} \ s_{t,t}]$, where $x_{t,t}$ and $y_{t,t}$ are the pixel coordinate values of the hand in the frame, $d_{t,t}$ is the hand moving direction in radian units and $s_{t,t}$ is the speed of the hand moving in pixel units. The predicted state at time $t+1$ is computed based on the prediction equation:

$$\begin{cases} \mathbf{h}_{t+1,t} = \mathbf{A}_h(\mathbf{h}_{t,t}) + \mathbf{b}_h \cdot w \\ \mathbf{A}_h(\mathbf{h}_{t,t}) = [\ x_{t,t} + s_{t,t} \cdot \cos(d_{t,t}) \\ \qquad\qquad y_{t,t} + s_{t,t} \cdot \sin(d_{t,t}) \quad d_{t,t} \quad s_{t,t}\ ] \end{cases} \tag{5}$$

where $\mathbf{A}_h$ is the non-linear state switching function, w is the system noise defined as a one-dimensional random variable with unit normal distribution, $\mathcal{N}(0, 1)$, and $\mathbf{b}_h$ is a four-dimensional row scale factor, set as unit in the system implementation.

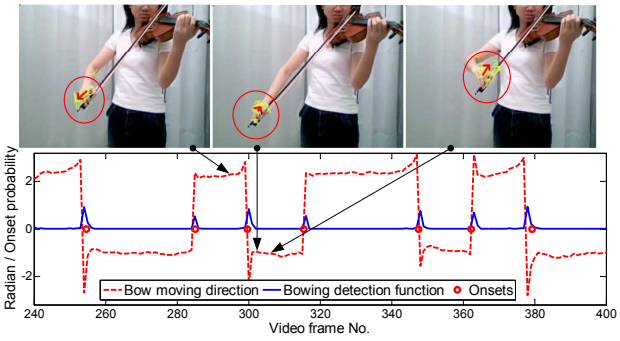


Figure 3: Illustration of bowing analysis for onset detection. Onsets are human annotated as circles.

At next time instance $t+1$, a measurement of the hand state $\mathbf{h}'_{t+1,t+1}$ is obtained by Algorithm 1. Then the predicted hand state $\mathbf{h}_{t+1,t}$ is updated as $\mathbf{h}_{t+1,t+1}$ by taking the current measurement into consideration:

$$\mathbf{h}_{t+1,t+1} = \mathbf{h}_{t+1,t} + \mathbf{G}_h(\mathbf{h}'_{t+1,t+1} - \mathbf{h}_{t+1,t}) \tag{6}$$

where $\mathbf{h}_{t+1,t}$ is the predicted hand state from Eq. (5), the second term is the difference between the measured state and the predicted. The Kalman gain, defined as $\mathbf{G}_h(\mathbf{h}_{t,t}) = \beta_h \cdot \mathbf{h}_{t,t}$, controls how much the filter is influenced by the current measurement in relation to the predicted state. The optimal value of β_h is found to be 0.5 in our system.

Algorithm 1: Hand measurement.

Input: Frame at time t
Output: Hand measurement $\mathbf{h}'_{t,t}$
1 Get an optical flow map with sparse motion points [23];
2 Only retain the motion points of skin-color;
3 Vote in bow moving direction to find the direction category with the most motion points;
4 Compute the average position, direction, and motion speed of the motion points in the selected direction category as the hand measurement $\mathbf{h}'_{t,t}$ and return $\mathbf{h}'_{t,t}$.

After hand tracking, bow moving direction sequence is obtained as $d(t)$ (Fig. 3). To model the bow reversal moments, we derive the detection function of bowing direction, $D_b(t)$, as the absolute value of the first order derivative of $d(t)$:

$$D_b(t) = abs(d'(t)). \tag{7}$$

In $D_b(t)$, peaks indicate the bow reversal moments, i.e., onsets reflected by visual bowing features. As shown in Fig. 3, high correlation between underlying onset times and bow reversal moments can be observed.

The hand tracking algorithm is robust against body movement of the violin player, background clutter and disturbing moving objects in a capturing field with non-skin color background. Implemented in C++ with OpenCV library [1], the algorithm tracks the bowing hand automatically in real-time. Hand tracking results are illustrated in Fig. 3 with motion features and tracked hand state shown as arrows in light (yellow) and dark (red) colors, respectively (see videos of the tracking results online[1]).

[1] `http://www.comp.nus.edu.sg/~bingjun/avamt.html`

4.2 Fingering Analysis for Onset Detection

In order to detect the finger press and release moments in a fingering video, we propose a two-stage fingering analysis algorithm using Kalman filter framework [6]. First the four violin strings are detected, i.e., the starting point (at the violin neck) and the ending point (at the violin bridge) of each string. Then each string is searched to pinpoint the finger positions by using the pre-calculated skin-color Gaussian model, Eq. (3).

The state of string i, $i \in \{1,2,3,4\}$, at time t is defined by a starting point and an ending point, $\mathbf{s}^i_{t,t} = [\mathbf{p}^{i,1}_{t,t}\ \mathbf{p}^{i,2}_{t,t}]$, where a point $\mathbf{p}_{t,t} = [x_{t,t}\ y_{t,t}\ d_{t,t}\ s_{t,t}]$. $x_{t,t}$ and $y_{t,t}$ are the pixel coordinate values of $\mathbf{p}_{t,t}$ in a video frame, $d_{t,t}$ is the moving direction in radian units and $s_{t,t}$ is the moving speed in pixel units of $\mathbf{p}_{t,t}$. The predicted state for string i at time $t+1$ is computed based on the prediction equation:

$$\begin{aligned}\mathbf{s}^i_{t+1,t} &= \left[\ \mathbf{p}^{i,1}_{t+1,t}\quad \mathbf{p}^{i,2}_{t+1,t}\ \right] \\ &= \left[\ \mathbf{A}_f(\mathbf{p}^{i,1}_{t,t}) + \mathbf{b}_f \cdot w \quad \mathbf{A}_f(\mathbf{p}^{i,2}_{t,t}) + \mathbf{b}_f \cdot w\ \right]\end{aligned} \tag{8}$$

where $\mathbf{A}_f$ and $\mathbf{b}_f$ are defined the same as $\mathbf{A}_h$ and $\mathbf{b}_h$ in the previous section (see Eq. (5)).

At time $t+1$, the string state $\mathbf{s}'^i_{t+1,t+1}$ is measured for each string by Algorithm 2. Then the predicted string state $\mathbf{s}^i_{t+1,t}$ is updated as $\mathbf{s}^i_{t+1,t+1}$ by taking $\mathbf{s}'^i_{t+1,t+1}$ into consideration (Eq. 6).

Algorithm 2: String measurement.

Input: Frame at time t
Output: String measurement $\mathbf{s}'^i_{t,t}$

1 Obtain the binary edge image by Canny edge detector;
2 Thin edges into one-pixel width in the edge image;
3 Apply Hough line transform to detect lines [13];
4 Among all detected lines, vote in line direction to find the strings with the dominant line direction;
5 In the string direction category, extract lines not farther than γ pixels from each other as the detected strings;
6 Search along each string i, in the edge image, to find the turning points, $(x^{i,1}_{t,t}, y^{i,1}_{t,t})$ and $(x^{i,2}_{t,t}, y^{i,2}_{t,t})$, at the violin neck and the violin bridge, respectively;
7 Compute $(d^{i,1}_{t,t}, s^{i,1}_{t,t})$ and $(d^{i,2}_{t,t}, s^{i,2}_{t,t})$ with respect to $\mathbf{s}^i_{t-1,t-1}$ to form $\mathbf{s}^i_{t,t}$ and return $\mathbf{s}^i_{t,t}$.

After the extraction of $\mathbf{s}^i_t$ in each frame, we further apply Algorithm 3 to detect only one active finger position f^i_t for each $\mathbf{s}^i_t$ in every frame.

Algorithm 3: Detection of active finger positions.

Input: Frame at time t, and $\mathbf{s}^i_t$, $i \in \{1,2,3,4\}$
Output: Active finger positions f^i_t

1 Starting from the bottom string $\mathbf{s}^1_t$;
2 For each $\mathbf{s}^i_t$, search along the string from $\mathbf{p}^{i,2}_t$ to $\mathbf{p}^{i,1}_t$ to pinpoint a finger position (x^i, y^i) with skin-color;
3 If there is no f^{i-1}_t on $\mathbf{s}^{i-1}_t$ with distance smaller than δ to (x^i, y^i), then set the distance of (x^i, y^i) to $\mathbf{p}^{i,2}_t$ as f^i_t and go to step 5; otherwise search further on $\mathbf{s}^i_t$;
4 If $\mathbf{p}^{i,1}_t$ is reached, set f^i_t as zero and go to step 5;
5 If $i < 4$, start searching $\mathbf{s}^{i+1}_t$ with principles described in step 2, 3 and 4; otherwise, return f^i_t, $i \in \{1,2,3,4\}$.

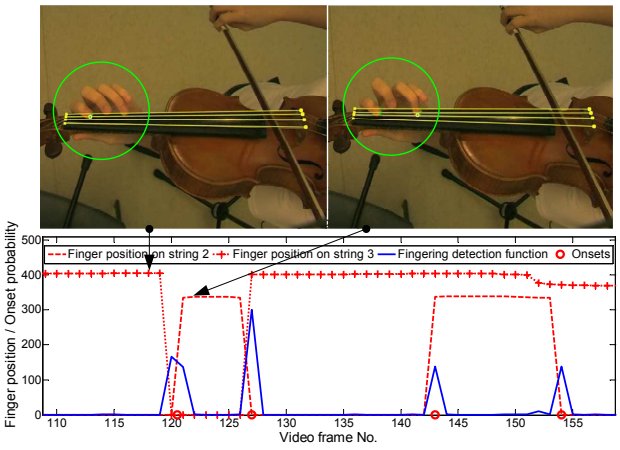


Figure 4: Illustration of fingering analysis for onset detection. Onsets are human annotated as circles. String numbers are in a bottom-up order.

From the fingering video, we extract four active finger position sequences $f^i(t)$, $i \in \{1,2,3,4\}$, each for one string (shown in Fig. 4). A sudden change of the active position (finger pressing position) on any string indicates a finger press or release moment, i.e., an onset timing. To model this, we define the detection function of violin fingering as the summation of the first order derivative of $f^i(t)$:

$$D_f(t) = \sum_{i=1}^{4} abs(f'^i(t)) \tag{9}$$

where peaks in $D_f(t)$ indicates the onset times.

However, because of the limitations of a 2D video, false positives may happen when a finger tip is above but not touching the string. Based on violin domain knowledge, most of the violin notes are monophonic and very few are double-stops or triple-stops. Therefore, peaks in $D_f(t)$ with fewer active finger positions are more likely to be true onset times, and vice versa. To lessen the influence of false active finger positions, we divide $D_f(t)$ by the number of active finger positions at each time instance:

$$D_f(t) = \frac{\sum_{i=1}^{4} abs(f'^i(t))}{\sum_{i=1}^{4} I(f^i(t)) + 1} \tag{10}$$

where $I(x)$ is an indicator function with value 1 when $x > 0$, and 0 otherwise. As shown in Fig. 4, high correlation can be observed between the fingering detection function and underlying note onsets.

Implemented in C++ with OpenCV library [1], the proposed algorithm successfully tracks strings and detects active finger positions in real-time. $\gamma = 10$ and $\delta = 15$ are found optimal for Algorithm 2 and 3 in the experiments. The global motion of the violin body or the player does not influence the fingering analysis algorithm, because we only detect the relative positions of the fingers on the strings. Analysis results are shown in Fig. 4. Tracked strings are marked as bright (yellow) lines with bright (yellow) dots as starting and ending points. All candidates of finger positions are marked by small dark (green) dots, and active finger positions are marked by large bright (yellow) dots (see videos of the analysis results online, footnote 1 in Section 4.1).

5. AUDIO-VISUAL FUSION

In this section, we discuss how to fuse audio-visual data of the violin playing to enhance onset detection. We evaluate state-of-the-art multimodal fusion techniques of both feature level (early) fusion and decision level (late) fusion [12]. The fused audio-visual detection function is expected to be more indicative of onsets than the one from a single modality. During the fusion process, audio-visual features are assumed to be synchronized, as audio and visual streams are captured simultaneously during the violin playing and the incoming audio and visual samples are time stamped in the software level. Visual features (detection functions with 30 Hz sampling rate) are linearly interpolated in time domain to be of the same sampling rate as the audio features (MFCCs or detection functions with 64.7 Hz).

5.1 Feature Level Fusion

In feature level fusion, we evaluate the feature concatenation (FC) technique [12]. For each time index t, we concatenate the audio feature $\mathbf{f}_t$, bowing feature $D_b(t)$, and fingering feature $D_f(t)$ to form a new audio-visual feature $\mathbf{f}_{av,t}$ in a higher dimensional space (47 dimensions, with 45 from MFCCs and 2 from bowing and fingering detection functions). Before concatenation, each attribute of the audio and visual features is normalized into [0, 1].

To derive the audio-visual detection function, the same approach as audio-only case is applied. Firstly, two GMMs with parameters Θ_{av}^{lo} and Θ_{av}^{ln} are trained by the EM algorithm to model the distribution of onset and non-onset audio-visual feature sets F_{av}^{lo} and F_{av}^{ln}, respectively. F_{av}^{lo} and F_{av}^{ln} are assembled in the same way as the audio-only case. Further, with the incoming audio piece and video streams, the audio-visual detection function is calculated as follows based on a time series of audio-visual features with unknown labels, $F_{av} = \{\mathbf{f}_{av,t} | 0 \leq t \leq T\}$:

$$D_{av}^{fc}(t) = Rect(P_{av}^{lo}(\mathbf{f}_{av,t}|\Theta_{av}^{lo}) - P_{av}^{ln}(\mathbf{f}_{av,t}|\Theta_{av}^{ln})) \qquad (11)$$

where the superscript fc of $D_{av}^{fc}(t)$ means feature concatenation fusion.

In data fusion literature [16], cross-modal correlation techniques, such as Principle Component Analysis (PCA), Latent Semantic Indexing (LSI), Canonical Correlation Analysis (CCA), etc., have been proposed to derive the correlation among modalities or reduce dimensionality in the fused feature space. However, based on the violin domain knowledge in our application, the correlation between audio and visual modalities has already been derived by audio processing and video tracking algorithms, because the audio-visual features extracted are both indicative of onset times. In addition, the dimensionality of the audio and visual feature spaces has also been reduced (45 dimensions for MFCCs, 1 dimension for bowing features and 1 dimension for fingering features). Therefore, for feature level fusion of our system, feature concatenation is directly applied without bothering any cross-modal correlation technique.

Feature level fusion allows early correlation between audio and visual modalities. In addition, it requires only one training phase to derive the overall audio-visual detection function. However, the caveat is that if the feature separability between onset and non-onset classes for a particular modality is poor, feature level fusion may corrupt the higher dimensional space after fusion and make its feature separability worse than the best single modality before fusion.

5.2 Decision Level Fusion

In decision level fusion, we evaluate both a rule-based fusion technique, i.e., linear weighted sum (LW) fusion, and a classification based fusion technique, i.e., Support Vector Machine (SVM) based fusion [10]. For each fusion technique, onset detection functions or decision scores from the three data streams, $D_a(t)$, $D_b(t)$ and $D_f(t)$, are firstly normalized into [0,1], and then fused to derive an overall decision score series $D_{av}(t)$ as the audio-visual detection function.

In comparison with feature level fusion, decision level fusion has the same representation for each modality, i.e., onset decision scores in our system, which makes the fusion easier to conduct. In addition, decision level fusion is scalable in terms of modalities, which is not easily achievable for feature level fusion. However, decision level fusion fails to utilize the feature level correlations between modalities. It normally requires an additional training phase, which is less efficient than feature level fusion.

5.2.1 Linear Weighted Sum Fusion

For each violin piece, given the normalized decision scores from the three data streams, $D_a(t)$, $D_b(t)$ and $D_f(t)$, the linear weighted sum fusion calculates the overall decision score series as:

$$D_{av}^{lw}(t) = w_a^{lw} \cdot D_a(t) + w_b^{lw} \cdot D_b(t) + w_f^{lw} \cdot D_f(t) \qquad (12)$$

where w_a^{lw}, w_b^{lw}, and w_f^{lw} are the weights for the corresponding data streams; and the superscript lw of $D_{av}^{lw}(t)$ means linear weighted sum fusion. After fusion, D_{av}^{lw} is normalized into [0, 1] as the final audio-visual detection function.

The rationale behind using the linear weighted sum fusion is that each peak indicative of an onset in any single data stream will be revealed in the fused detection function with the peak value multiplied by the corresponding weight. Therefore, if the detection function of each data stream is noiseless, the fusion will complement among the data streams and produce an audio-visual detection function which could reveal onsets missed by an individual data stream. However, if there are false peaks in the detection function from any single data stream, the false peaks will also propagate to the fused one, thus making the overall detection function more noisy.

Linear weighted sum fusion is computationally inexpensive, whereas its fusion performance is sensitive to the combination weights, of which the optimal values can only be found based on extensive experiments.

5.2.2 SVM based Fusion

Multimodal fusion can also be considered as a pattern classification problem [10]. In our application, scores from individual data streams can be viewed as input patterns to be recognized as onsets or non-onsets. Among classification based fusion methods, Support Vector Machines (SVM) [7] have been found effective in multimodal fusion literature [10], where SVM outperformed other evaluated classifiers, including Multilayer Perceptrons, k-Nearest Neighbours, etc. In our system, SVM is further evaluated to show its effectiveness in multimodal fusion for onset detection and violin transcription.

To evaluate SVM based fusion in our system, decision scores from three data streams, $D_a(t)$, $D_b(t)$ and $D_f(t)$, are concatenated at each time t to form a decision vector series $\mathbf{D}(t) = [D_a(t)\ D_b(t)\ D_f(t)]^T$. Then the training data of

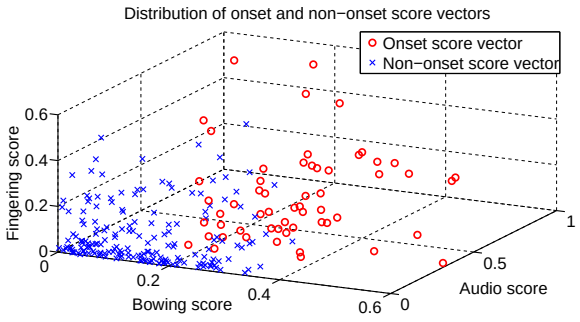


Figure 5: Score vector distribution of onset and non-onset frames.

SVM for onset and non-onset classes are selected from $\mathbf{D}(t)$ in the same way as training GMM.

With the training data, the SVM firstly maps features into a higher dimensional space by a mapping function, $\Phi : \Re^n \rightarrow \Re^{n'}$, $n < n'$. SVM then finds the optimal linear separating hyperplane in $\Re^{n'}$, $H(\mathbf{w}, \mathbf{b}) = \{\mathbf{x}|\ \mathbf{w}^T \cdot \mathbf{x} + \mathbf{b} = 0\}$, with the largest distance to the marginal data (support vectors) of either class, where $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{b}$ are vectors in $\Re^{n'}$.

During the recognition, given an input decision vector $\mathbf{D}_t^l$ with unknown label l, the decision value is obtained as:

$$D(\mathbf{D}_t^l) = \Phi(\mathbf{D}_t^l)^T \cdot \mathbf{w} + \mathbf{b} \tag{13}$$

In our implementation, a radial basis function is used as the kernel function:

$$\begin{aligned} K(\mathbf{D}_t^l, \mathbf{D}_j^{l_j}) &= \Phi(\mathbf{D}_t^l)^T \cdot \Phi(\mathbf{D}_j^{l_j}) \\ &= \exp(-\tfrac{1}{3}\|\mathbf{D}_t^l - \mathbf{D}_j^{l_j}\|^2) \end{aligned} \tag{14}$$

Since $D(\mathbf{D}_t^l)$ is an uncalibrated value, and not a probability, it cannot be used as the fused decision score. Therefore, we further calculate a calibrated value from it as the audio-visual decision score by the method proposed in [19]:

$$D_{av}^{svm}(t) = D_{av}^{svm}(\mathbf{D}_t^l) = \frac{1}{1 + \exp(A \cdot D(\mathbf{D}_t^l) + B)} \tag{15}$$

where A and B are estimated scalars by minimizing the negative log-likelihood function using training data and their decision values; and the superscript svm of $D_{av}^{svm}(t)$ means SVM based fusion. After normalization, $D_{av}^{svm}(t)$ is used as the audio-visual detection function obtained by SVM based fusion.

In pattern classification literature, SVM based fusion is considered superior to linear weighted sum fusion because of its capability in finding a non-linear yet optimal separating hyperplane based on the training data. As shown in Fig. 5, the onset and non-onset score vectors are not linearly separable. Therefore, an optimal non-linear separating hyperplane of SVM is potentially superior to a linear separating plane produced by the linear weighted sum fusion. The non-linearity of SVM is desirable when the noise pattern in a particular modality is consistent enough for SVM to generalize well, in which case SVM may fuse useful information of different modalities effectively while mostly discarding noisy patterns. As observed in our experimental results, SVM based fusion performs the best among all evaluated fusion techniques.

5.3 Audio-Visual Violin Transcription

After multimodal fusion, onset times are detected by a peak picking algorithm. This approach finds the local maxima from a detection function $D(t)$ subtracted by a series of thresholds $\tilde{\delta}(t)$. This process is illustrated in Fig. 1 as the onset time peaking. The thresholds are dynamically derived from the detection function based on median filtering [8]:

$$\tilde{\delta}(t) = \delta + \lambda \cdot \text{median}(D(t - W), ..., D(t + W)) \tag{16}$$

where δ is the base threshold, whose best value is different for various onset detection functions; λ is the ratio threshold tuned as 0.7 for all methods during experimentation. W is the half window length for the median filter, and is tuned to be 4 experimentally.

After onset detection, a whole violin piece is segmented into individual notes. Then an energy based activity detector is used to find the most active portion in energy of the audio signals within each note. The audio-only pitch estimator described in Section 3.2 is applied to that portion to calculate the pitch values. If the active portion is less than a ratio (experimentally found to be 30%) of the whole note duration, that note is considered silent, and no pitch estimation is conducted for it.

After pitch estimation, a MIDI (piano-roll) notation of a violin piece is obtained as the learning feedback to violin students. Each note of the piece is parameterized by its starting time (its onset), ending time (onset of the next note) and pitch values.

6. EVALUATION

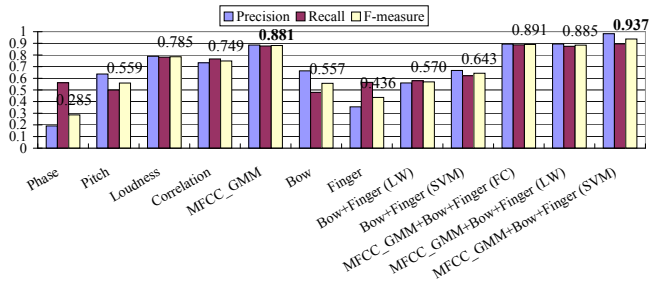
In this section, different onset detection approaches with several fusion techniques are evaluated. To compare with state-of-the-art methods for onset detection, implementations of the phase change based method (Phase) [4], pitch change based method (Pitch) [9], equal loudness change based method (Loudness) [14], and spectrum correlation change based method (Correlation) [18] are evaluated on the same audio-visual violin database. In addition, by combining different onset detection approaches with the same audio-only pitch estimation method, described in Section 3.2, the overall transcription performance is evaluated.

6.1 Audio-Visual Violin Database

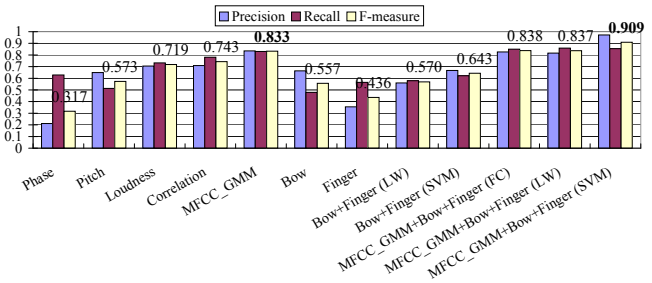
The evaluation of different approaches for onset detection and music transcription is conducted based on an audio-visual database with 36 violin pieces of 6340 notes. All pieces are recorded in an ordinary room (SNR=24dB) with the same audio and visual settings described in the previous sections. The selected pieces cover a wide range of violin playing styles, e.g., vibrato, legato, staccato, double stop, etc. The tempo of the pieces ranges from slow to fast. Each piece is recorded twice, once without vibrato and once with vibrato playing style. Human annotation is carried out by trained musicians as the evaluation ground truth. To further evaluate the usefulness of the visual modality in noisy conditions for violin transcription, SNR of the original database is reduced to 15, 0, and -5dB with additional white noise.

6.2 Evaluation Metric

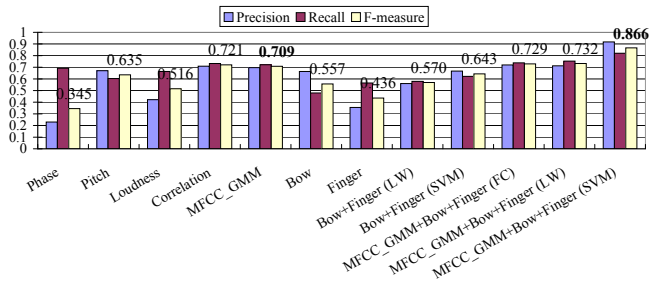
For onset evaluation, a detected onset within 50 milliseconds to the left or right of a human annotated onset is considered correctly detected. Three metrics are used to eval-



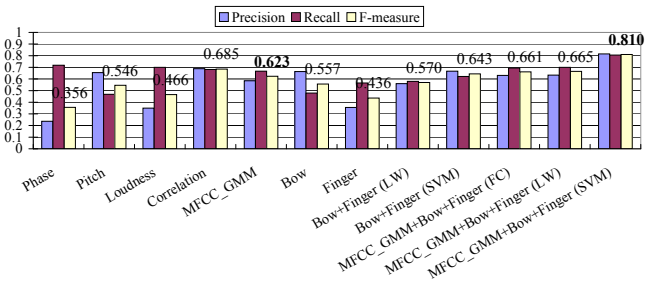
(a) Results on original database (SNR=24dB).



(b) Results on noisy database (SNR=15dB).



(c) Results on noisy database (SNR=0dB).



(d) Results on noisy database (SNR=-5dB).

Figure 6: Performance comparison of different onset detection approaches.

uate performance of onset detection. They are **precision**, **recall** and the balanced **F-measure**, where precision is the percentage of correctly detected onsets over all detected onsets, and recall is the percentage of correctly detected onsets over all annotated onsets.

In violin transcription evaluation, a note is correctly transcribed by the system, if its starting time, ending time and pitch values are all correctly detected. **Accuracy** is used to evaluate transcription approaches. It is calculated as the percentage of correctly transcribed notes over all annotated notes.

6.3 Experimental Results

During evaluation, the grid search method is applied on the parameter combination space of each onset detection and transcription method to find the best F-measure or accuracy. Best parameters found for each part of the system are described in the corresponding sections. On the original database (SNR=24dB), two-fold cross validation is conducted for GMM and SVM related approaches. When evaluated on more noisy databases, GMM and SVM are trained based on the original database. After evaluation, the best F-measure with the corresponding precision and recall for each onset detection method is illustrated in Fig. 6. And the best accuracy for audio-only and audio-visual transcription approaches is shown in Fig. 7.

6.3.1 *Performance Comparison of Onset Detection*

In the audio-only case, as shown in Fig. 6, the proposed onset detection approach (MFCC_GMM) achieves 88% and 83% F-measure on databases with SNR=24, 15dB, respectively, which outperforms state-of-the-art methods (Phase, Pitch, Loudness, and Correlation) by 10% and 9% F-measure. On databases with SNR=0, -5dB, MFCC_GMM still performs much better than state-of-the-art methods except the spectrum correlation change based one (Correlation). Phase change based method (Phase) performs the worst, because of the vibrato playing style in the database. Therefore, the MFCC_GMM onset detection approach is superior to state-of-the-art methods in less noisy environment (e.g., SNR=24, 15dB), and generally good in more noisy environment (e.g., SNR=0,-5dB).

In the video-only case, bowing (Bow) and fingering (Finger) generate 55% and 43% F-measure, respectively. With linear weighted sum fusion (LW) and SVM based fusion (SVM), the visual modality (Bow+Finger) generates 57% and 64% F-measure, respectively. The optimal weights for bowing and fingering data streams are 0.6 and 0.4, respectively. As the visual modality is not affected by acoustic noise, their performance is stable for onset detection in different noisy conditions.

In audio-visual fusion (MFCC_GMM+Bow+Finger), three fusion techniques (feature concatenation based fusion in feature level, FC; linear weighted sum fusion and SVM based fusion in decision level, LW and SVM) generally improve the onset detection performance. In linear weighted sum fusion, the optimal weights for audio, bowing and fingering data streams are found by extensive experiments: 0.7, 0.2, 0.1 for the database with SNR=24dB; 0.7, 0.2, 0.1 for the database with SNR=15dB; 0.6, 0.3, 0.1 for the database with SNR=0dB; and 0.5, 0.3, 0.2 for the database with SNR=-5dB. SVM based decision level fusion is found to be the most effective fusion approach, as it improves over the best single modality by 5%, 7%, 16%, and 18% F-measure in respective databases. The more noisy the audio modality is, the more improvement is accomplished by fusing the visual modality (shown in Fig. 7). The best F-measure achieved by fusing audio and visual modalities using SVM are 93%, 90%, 86%, and 81% on databases with SNR=24, 15, 0, and -5dB, respectively.

As revealed in the experimental results, feature concatenation fusion in feature level and linear weighted sum fusion

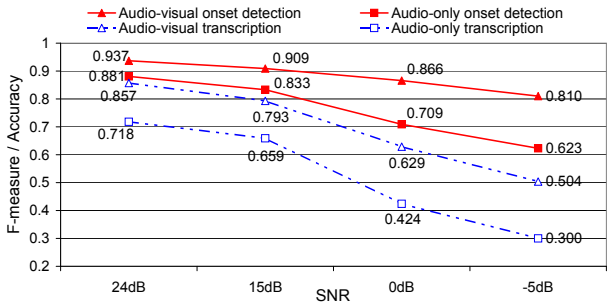


Figure 7: Performance improvement by the visual modality with SVM based decision level fusion in different noisy conditions.

in decision level are inferior to SVM based decision level fusion for onset detection. This is mainly because bowing and fingering streams (with 55% and 43% F-measure) are more noisy than the audio modality in general (with 62% to 88% F-measure on different noisy conditions).

For the feature concatenation fusion in feature level, the noise of the visual modality remains in the fused feature space, thus the feature separability for onset and non-onset classes does not improve as much as in the SVM based fusion. In addition, the audio feature dimensionality (MFCCs with 45 dimensions) and visual feature dimensionality (detection functions with 2 dimensions) are out of balance. Most of the modeling efforts will fall into the audio features rather than the visual features in the concatenated feature space. This further affects the improvement brought by the visual modality in feature concatenation fusion. One way to balance audio and visual dimensions is to apply additional dimensionality reduction techniques, such as PCA, on the audio feature space. However, according to our extra experiments, after reducing the 45 dimensional MFCCs into 3 to 5 dimensions by PCA, the classification performance by the audio-only modality suffers due to the loss of information. After fusing the more balanced audio and visual modalities, the overall onset detection performance is not better than the the performance by the unbalanced feature concatenation fusion. With this dilemma, the feature concatenation fusion in feature level is not suitable for our application.

For the linear weighted sum fusion in decision level, the noise of the visual modality propagates to the fused onset detection function more severely than in the SVM based fusion, which results in less improvement by the linear weighted sum fusion. Compared with other fusion methods, the SVM based fusion in decision level improves the onset detection performance by the most amount of F-measure. This not only reveals the advantage of decision level fusion, in which the audio and visual modalities have the same representation (detection function for onsets) and balanced dimensions (1 dimension for each data stream), but also verifies the effectiveness of SVM's non-linearity and optimal separating hyperplane in fusing audio and visual modalities of a violin playing for onset detection.

6.3.2 *Performance Comparison of Transcription*

The best transcription accuracy for audio-only and audio-visual transcription approaches is shown in Fig. 7. In the audio-only case, with MFCC_GMM onset detection, the overall transcription accuracy is 71%, 65%, 42%, and 30% on the four databases, respectively. In fusing audio and visual modalities, the best transcription performance are 85%, 79%, 62%, and 50% on the four databases, which improves over audio-only approaches by 14% to 20% accuracy (shown in Fig. 7).

As shown in the experimental results, visual modality is helpful in improving onset detection performance and transcription accuracy. Especially for violin practice at home, where the acoustic conditions are far from ideal, introducing visual modality is beneficial to high performance music transcription system.

7. RELATED WORKS

Few works have been published on music transcription by fusing multimodal features. Drum transcription in [11] is the first system we found dealing with percussive sounds using both audio and visual modalities. Tempo analysis of sitar performance based on multimodal sensor fusion are found in [5]. Our previous work in [25] is the first attempt for violin transcription with audio-visual inputs. However, the previous system used markers to aid bowing and fingering analysis, which is less practical compared with the system in this paper. One attempt to automatic fingering analysis without markers has also been conducted by us in [28]. Nevertheless, the finger tracking algorithm in [28] is more computationally expensive and less suitable for practical applications compared with the work in this paper. The correlation between violin music and the visual modality, bowing and fingering, has been shown in cognitive brain research [2] and other violin literature [3]. Inspired by those works, we introduced the visual modality to utilize the complementary information between the audio and visual modalities in violin transcription. Audio-visual fusion based approach significantly improves violin transcription performance based on our experimental results. Superior performance has also been observed by using multiple modalities in audio-visual speech recognition [20], audio-visual biometric [10], concept detection in multimedia data [26], etc.

8. CONCLUSIONS

In this paper, we have built an audio-visual fusion based music transcription system for violin practice in home environment. To address the difficulties in onset detection of PNP sounds, such as from the violin, we have proposed an audio-only onset detection approach based on supervised learning. Two GMMs are used to classify onset and non-onset audio frames based on MFCC features. MFCC feature models the spectrum envelop effectively, which forms the basis of superior classification performance. In addition, due to the efficient modeling approach by GMM and the low dimensionality of MFCCs, the proposed audio-only onset detection method is computationally efficient, thus suitable for practical applications.

To further enhance audio-only onset detection, the visual modality of violin playing, including bowing and fingering, is introduced into our system. Two webcams of the system can be easily placed to capture bowing and fingering videos in home environment. Fully automatic and real-time algorithms have been devised to conduct bowing and fingering analysis, which maximizes the practicality of the system.

State-of-the-art multimodal fusion techniques have been evaluated to fuse the audio and visual modalities for en-

hanced performance of onset detection and overall transcription. SVM based decision level fusion is verified to be superior to feature concatenation fusion in feature level and linear weighted sum fusion in decision level. With the help of the visual modality and SVM based decision level fusion, both onset detection and transcription performance are improved significantly. Especially in home environment, where the acoustic conditions are far from ideal, the performance improvement by the visual modality is more substantial.

Based on the above contributions and extensive evaluations, the violin transcription system has achieved good performance even in acoustically inferior conditions. This transcription system is able to provide more accurate transcribed results as feedback to students when they practice violin at home. With efficient and automatic audio-visual analysis algorithms, the system can be easily set up once and for all in a home environment.

9. ACKNOWLEDGMENTS

We would like to thank Mohan S. Kankanhalli for providing us with references in multimodal fusion.

10. REFERENCES

[1] `http://opencvlibrary.sourceforge.net`.

[2] A. P. Baader, O. Kazennikov, and M. Wiesendanger. Coordination of bowing and fingering in violin playing. *Cognitive Brain Research*, 23:436–443, 2005.

[3] A. Bachmann. *An encyclopedia of the violin*. Da Capo Press, 1975.

[4] J. Bello and M. Sandler. Phase-based note onset detection for music signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 49–52, 2003.

[5] M. S. Benning, A. Kapur, B. C. Till, and G. Tzanetakis. Multimodal sensor analysis of sitar performance: Where is the beat? In *IEEE Workshop on Multimedia Signal Processing*, pages 74–77, 2007.

[6] E. Brookner. *Tracking and Kalman Filtering Made Easy*. John Wiley & Sons, 1998.

[7] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[8] N. Collins. A comparison of sound onset detection algorithms with emphasis on psycho-acoustically motivated detection functions. In *Proceedings of AES118 Convention*, 2005.

[9] N. Collins. Using a pitch detector for onset detection. In *Proceedings of International Conference on Music Information Retrieval*, 2005.

[10] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez. A comparative evaluation of fusion strategies for multimodal biometric verification. *Audio- and Video-based Biometric Person Authentication*, pages 1056–1056, 2003.

[11] O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 205–208, 2005.

[12] D. L. Hall. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Inc., 2004.

[13] N. Kiryati, Y. Eldar, and A. M. Bruckstein. A probabilistic hough transform. *Pattern Recognition*, 24(4):303–316, 1991.

[14] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3089–3092, 1999.

[15] A. Lacoste and D. Eck. A supervised classification algorithm for note onset detection. *EURASIP Journal of Applied Signal Processing*, 2007(1):153–153, 2007.

[16] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of ACM Conference on Multimedia*, pages 604–611, 2003.

[17] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

[18] A. Loscos, Y. Wang, and W. Boo. Low level descriptors for automatic violin transcription. In *Proceedings of International Conference on Music Information Retrieval*, 2006.

[19] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 2000.

[20] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[21] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

[22] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[23] J. Shi and C. Tomasi. Good features to track. In *Proceedings of IEEE Computer Society Conference Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[24] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proceedings of Graphicon*, pages 85–92, 2003.

[25] Y. Wang, B. Zhang, and O. Schleusing. Educational violin transcription by fusing multimedia streams. In *Workshop of Educational Multimedia and Multimedia Education*, 2007.

[26] Y. Wu, E. Y. Chang, K. C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of ACM Conference on Multimedia*, pages 572–579, 2004.

[27] J. Yin, Y. Wang, and D. Hsu. Digital violin tutor: An integrated system for beginning violin learners. In *Proceedings of ACM Conference on Multimedia*, pages 976–985, 2005.

[28] B. Zhang, J. Zhu, Y. Wang, and W. Leow. Visual analysis of fingering for pedagogical violin transcription. In *Proceedings of ACM Conference on Multimedia*, 2007.