COMBINING HAND POSE ESTIMATION WITH AUDIO FOR NOISE-ROBUST PIANO SCORE FOLLOWING

Kaitlin Pet[*]

Indiana University Bloomington
Luddy School
Bloomington, Indiana, USA

Akira Maezawa

Yamaha Corporation
Music Informatics Group, R&D Division
Hamamatsu, Shizuoka, Japan

ABSTRACT

Recently, deep learning-based pose estimation tools have allowed for accurate, real-time hand position tracking. In this paper, we show how such hand tracking tools can be incorporated into multi-modal piano score following. We created an Hidden Markov Model (HMM)-based score follower that incorporates both piano audio and hand position features yielded from top-down video of a pianist's hands on the keyboard. This multi-modal system has the ability maintain score follower performance in highly noisy audio conditions.

Index Terms— Score following, multi-modal score alignment, music information retrieval

1. INTRODUCTION

Online score following [1, 2], allows a given input music performance to be continually temporally matched to its corresponding symbolic score. Traditionally, score followers use performance audio alone as input. In these cases, the score follower architecture is often a variant on a Hidden Markov Model (HMM) where states, emission probability and transition probabilities are derived from the pitches and rhythms specified in the musical score [3, 4, 5].

Audio-based score followers perform best in situations with a high signal-to-noise ratio; that is, the audio being followed is easy to hear. Many score follower becomes unreliable when the performance audio becomes unintelligible or inaudible. In such situations, understanding the player's posture can be a complementary visual input, like how a conductor will look at the pianist's body language for cues during loud *tutti* sections of a piano concerto.

Multi-modal score followers have the potential to boost score follower performance in low SNR situations. Some systems incorporate additional inputs at a few carefully selected locations [3, 6]. Others use visual data as a complementary time series to directly infer score position. For example, the Max/MSP-based IMuSE automatic accompaniment system [7] tracks performances with score followers operating on different types of data, trusting the score follower with the smoothest series of inferred note onsets. One video-based source is the pianist's hand position across time, from which the center of mass is extracted by computer vision techniques. IMuSE uses IRCAM's `gf` [8] object to learn the hand's trajectory during rehearsal, then infers the most likely position in the score at any given time based on the new center-of-mass trajectory during performance.

The advent of deep learning-based pose estimation creates the potential to more easily and accurately incorporate gestural information into score following. For example, MediaPipe's Hand Landmarker [9] infers the most likely position of 21 "landmarks" on each hand from static images or video feed. Previous papers have used pose estimation in piano transcription [10], but to our knowledge, this is the first attempt to use hand landmark motion patterns in score following.

Inspired by multi-modal score following systems [7], we fuse video and audio sources for score following using both hand positions and audio. In addition, we incorporate a feature weighing mechanism to allow the importance of different modalities. Our contributions are as follows: (1) we present a first score follower that incorporates deep learning-based pose estimation, (2) we present a method to fuse audio and video feeds and weigh their importance, (3) we evaluate the proposed multi-modal score follower on a varied dataset of clapping and background music

2. METHODS

2.1. Data Collection

We obtained rehearsal and performance takes of multi-modal piano performance data, capturing performance MIDI and top-down video. The camera was suspended above the piano to give a top-down view of the performance (see Figure 1). All performances were recorded with the same camera angle and lighting. Audio was synthesized from performance MIDI using the Steinway Piano virtual instrument in Logic Pro X.

Performance data (rendered audio and video recordings) were score-matched – i.e. for each etude, all notes in its sym-

[*] Research performed when working at Yamaha.

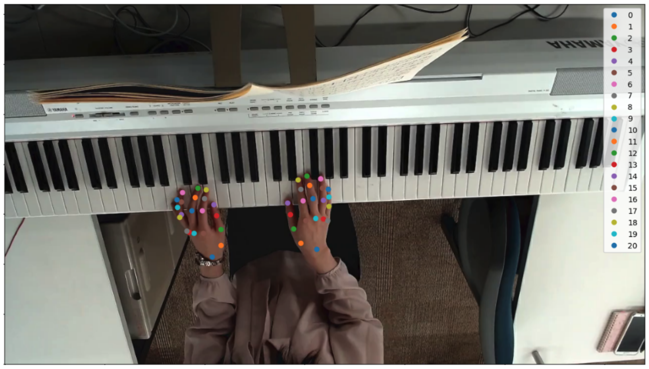

Fig. 1. Frame of top-down performance video with MediaPipe landmarks colored.

bolic score were matched to a unique timestamp in the performance. This offline score match was achieved with a combination of custom code, the Orchestra program [11], and hand corrections. There was not a one-to-one relationship between note onsets specified in the score and actual note onsets in the performance. Performances contained mistakes, missing notes, and extra notes. For example, when the pianist did not play a phrase to her satisfaction, she would sometimes repeat the phrase one or more times before continuing on with the performance. In cases of missing notes, a sensible location was chosen based on the score's rhythm and the player's tempo. In cases where notes were added or short sections were repeated, the nearest note onset was placed either before or after the added material.

2.2. Hand Landmarking

MediaPipe's Hand Landmarker [9] was used to identify 21 "landmark locations" on each hand on all frames of the performance video (See Figure 1 for labeled landmarks). If a landmark is not observed or a hand is missing from the video frame, the missing landmark locations are set to their last observed positions. We found the landmarks' x-coordinate most relevant for score position, so discarded the y and z coordinates for each landmark provided by MediaPipe. We then further processed landmark features by extracting each hand's center of mass and discarding all landmarks but the fingertip locations. We surmised that since fingertips were directly involved in piano sound production, their location would be the most relevant to a score follower. This yielded a total of 12 hand position features across two hands.

2.3. Multi-modal HMM for Score Following

Our method matches a reference ("rehearsal") piano performance to another performance of the same piece in real-time. To track the performance in an online manner, our method models the performance using an HMM and decodes its state sequence by computing the forward probabilities at each time frame.

We represented the score by a sequence of n hidden states, $\{x_1 \dots x_n\}$. These hidden states are matched to m frames of observed features $\{y_1 \dots y_m\}$. The features of y could consist of **(1) audio features** only, **(2) hand landmark features** only, or **(3) a combination** of the two. Emission probability for each state x_n was the multivariate normal distribution $P(y|x_n) \sim N(\mu_n, \Sigma_n)$, where μ_n is the mean of a combination of audio and/or visual features associated with score state x_n, and Σ_n is the covariance matrix associated with score state x_n.

μ and Σ are trained based on the piece's music score (i.e. its *intended* pitch content during each state), the first take's MIDI data (i.e. its *actual* pitch content during each state), and the first take's hand landmark positions. To compute audio-derived means, reference audio was first generated by rendering the symbolic score with the Steinway Piano virtual instrument in Logic Pro X. The audio was transformed into a constant-Q chromagram, then normalized such that each frame summed to 1. Visual means were computed from the reduced hand landmark features described in Section 2.2.

When computing the covariance Σ of each state, we assumed audio features, right/left-hand fingertips, and right/left-hand center of mass were independent of each other. Within each of these data sources, there were sometimes insufficient samples per state for covariance computation. We thus hand-chose a minimum covariance for each data source. This minimum covariance was set as the emission covariance should the state be too short for covariance computation; otherwise, this minimum covariance was added to the computed covariance values.

When fusing audio and visual features, it is preferable to adjust the relative importance of the audio features with respect to visual features. Thus we introduce an **audio variance penalty** parameter γ such that for state n, given audio feature covariance matrix $\Sigma_n^{(a)}$ and visual feature covariance $\Sigma_n^{(v)}$, Σ_n is given as diag($[\gamma\Sigma_n^{(a)}, \Sigma_n^{(v)}]$).

3. EXPERIMENTS

We evaluate the effect of incorporating audio and visual feeds under a noisy environment, depending on the nature of the noise and the audio variance penalty γ.

3.1. Dataset

Performance MIDI and video of Burgmüller's 25 etudes for piano were recorded by a late-intermediate to early-advanced pianist on a digital piano (Yamaha P125). Each piece was performed twice. In the performances, certain takes were mistake-free, while others contained missing/additional notes or repeated phrases. Of the 25 etudes, twelve were retained

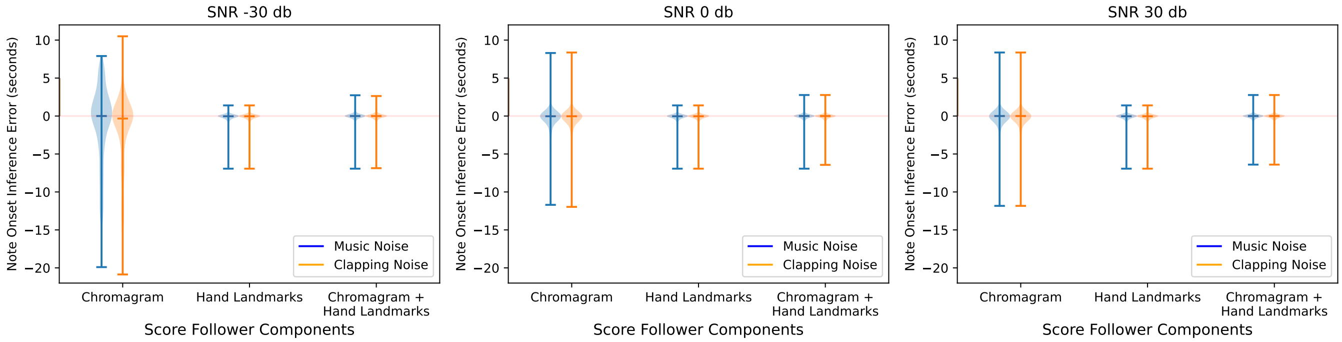

Fig. 2. Highlighting behavior at the chromagram penalty level of 30. Comparing raw note onset error of all pieces in the test set with differing SNR and noise types. Positive values indicate the inferred onset time is late, negative values indicate the inferred onset time is early. Violin plots show the full range of onset errors, as well as highlighting the median.

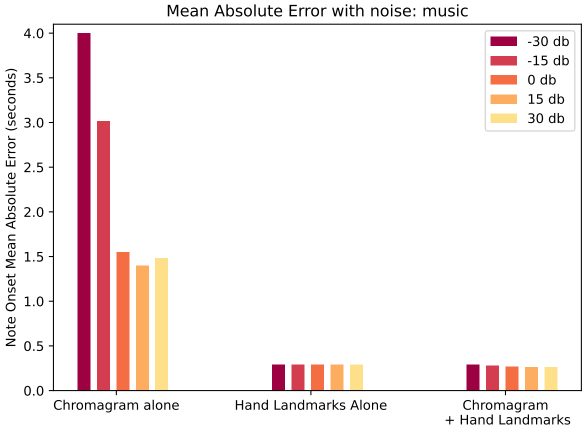

Fig. 3. Highlighting behavior at the chromagram penalty level of 30. The mean absolute error of different score follower ablation conditions across a range of SNRs. Performances exposed to clapping as opposed to music showed similar error trends.

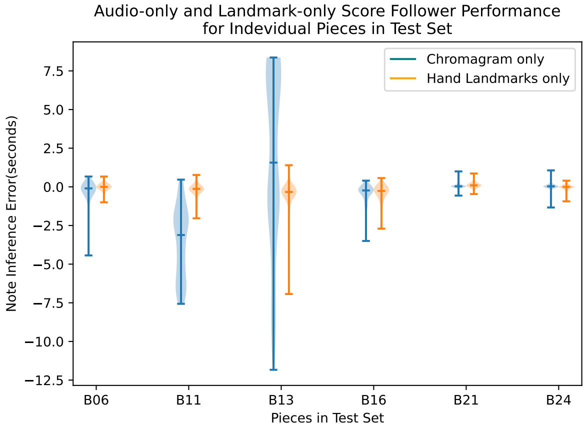

Fig. 4. Violin plots of audio-only and landmark-only score followers across individual pieces in the test set with 30dB SNR.

for analysis[1].

3.2. Experimental Conditions

Six of the twelve etudes were used as a validation dataset to hand-tune the penalty associated with the hand landmarks. We report the results from the remaining 6, which were used as test data.

For each piece, the HMM was trained using the first take of the piece and the score. States of the HMM were created by partitioning by score by the union of note onsets and running sextuplets. Transition probabilities between states adjacent in the score depended on the ratio of their score-based lengths.

Score followers were tested with three types of input data: (1) audio alone, (2) hand landmarks alone, and (3) combined audio and hand features.

The evaluation was performed online by feeding audio and video of the performance take to the score follower in 30-ms frames. After receiving each new frame, the score follower HMM was used to compute the forward posterior maximum state.

To evaluate the effectiveness of each score follower, we compared the observed note onset times in the performance take to the inferred note onset times from the HMM. However, given our online method of state determination, it is possible that certain states corresponding with note onsets will be skipped, repeated multiple times, or never reached. We thus computed the time of each note onset state x_o, pos(x_o), as

[1] Two etudes contained too many mistakes, two contained pickup notes which were incompatible with the pipeline, and nine could not be effectively pre-processed for other reasons.

follows:

$$\text{pos}(x_o) = \max(l(x < x_o), l(x \geq x_o)), \tag{1}$$

where $l(x)$ is the frame in the test sequence where state x is observed. This ensures an onset time exists for every note.

3.3. Experiment 1. Effect of Noise on Score Following

We computed the error between ground truth note onset times and inferred note onset times from the score follower as different noises were added and as the signal-to-noise ratio (SNR) was varied. Specifically, we add (1) another music audio [12] and (2) clapping audio to the piano performance audio, at SNR of -30 dB to 30 dB in 15 dB increments. Clapping audio is meant to simulate clapping during a live show, while music audio is intended to simulate a "backstage" or "practice room" environment where alternate music sources cannot be isolated from the input audio.

A violin plot of absolute error across all test pieces with γ set at 30 can be seen in Figure 2, and the mean absolute onset error in shown in Figure 3. As can be seen, the landmark-only score follower outperformed the audio-only score follower by a large margin, even in high SNR conditions which were anticipated to favor audio. The error decreases slightly as video and audio data are fused, suggesting that the two media complement each other. Furthermore, the performance of chromagram-only score follower drops as SNR decreases regardless of the noise type, showing the weakness of audio-based score follower under generally noisy environments.

The audio-only score follower performed poorly compared to the landmark-only score follower, even at high SNR where chromagram features should not have been adversarially affected by noise. In fact as shown in Figure 4, the audio-only score follower and landmark-only score follower performed comparably in certain test set pieces. We hypothesize the poorer performance of chromagram-only score follower in only certain pieces may be attributed to a higher sensitivity to performance mistakes. That is, pianists make audible mistakes while executing more-or-less same performance movements, like hitting a neighboring key or striking a key on an unintended finger. This means that mistakes produce significant differences in the chromagram but little difference in the hand landmarks.

3.4. Experiment 2. Effect of Audio Variance Penalty

We evaluated the alignment error as the audio variance penalty γ was varied while keeping the SNR fixed at 0 dB. Results are shown in Figure 5. We can see that at low γ values, the combined visual-audio score follower more closely tracks the audio-only score follower's behavior, while at larger γ values, the combined score follower's behavior is more similar to the visual-only one. For a certain set of penalties, the combined score follower outperforms both audio-only and visual-only score followers. This suggests that a well-tuned combined score follower can consolidate observations in a way that enhances the ability to track a pianist's behavior.

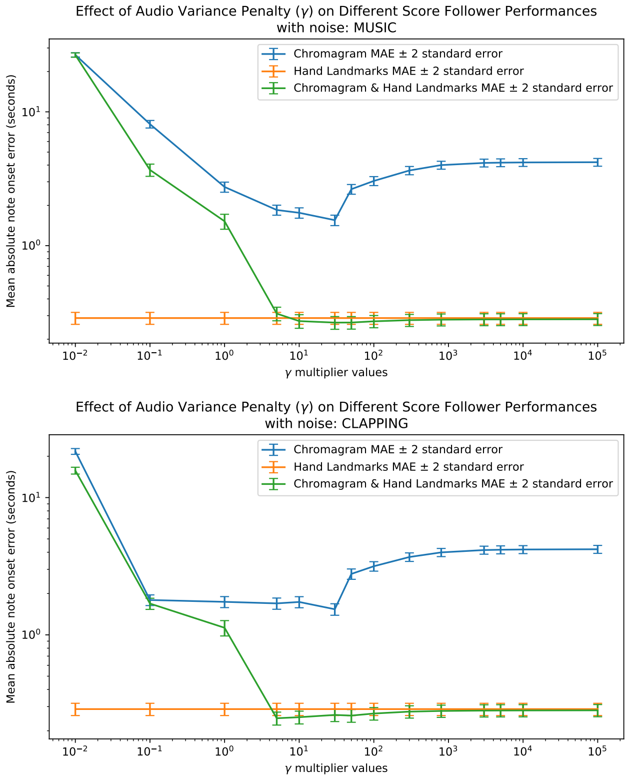

Fig. 5. Highlighting behavior of varying γ (SNR is fixed at 0 dB). Error bars are +- 2 standard errors of the mean. Combined score follower (green) exhibits more chromagram-like behavior at low penalty values and more landmark-like behavior at high penalty values.

4. CONCLUSION

This paper presented a multi-modal score following system that fuses deep learning-based pose features with audio. We demonstrated that audio and visual data sources can complement each other when an appropriate feature weighting scheme is used. Furthermore, we found visual features were robust when performance contained many audible mistakes. We believe our work expands the applicability of piano score following in noisy environments, such as practice rooms or music genres for which cheering or clapping during a performance is the norm. In the future, we hope to continue inquiry into how this system performs in less ideal lighting conditions that more closely simulate a live concert scenario, as well as better partitioning of visual features into states.

5. REFERENCES

[1] Roger Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proceedings of the International Conference on Computer Music (ICMC)*, 1984, pp. 193–198.

[2] Barry Vercoe, "The synthetic performer in the context of live performance," in *Proceedings of the International Conference on Computer Music (ICMC)*, 1984, pp. 199–200.

[3] Akira Maezawa and Kazuhiko Yamamoto, "MuEns: A Multimodal Human-Machine Music Ensemble for Live Concert Performance," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, May 2017, pp. 4290–4301, ACM.

[4] Arshia Cont, "ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music.," in *International Computer Music Conference (ICMC)*, Belfast, Ireland, Aug. 2008, pp. 33–40.

[5] Christopher Raphael, "Music Plus One and Machine Learning," in *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, 2010, pp. 21–28.

[6] Takeshi Mizumoto, Angelica Lim, Takuma Otsuka, Kazuhiro Nakadai, Toru Takahashi, Tetsuya Ogata, and Hiroshi Okuno, "Integration of flutist gesture recognition and beat tracking for human-robot ensemble," *Proc of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, 11 2010.

[7] Martin Ritter, Keith Hamel, and Bob Pritchard, "Integrated multimodal score-following environment," in *International Computer Music Conference*, 2013.

[8] Frédéric Bevilacqua, Norbert Schnell, Nicolas Rasamimanana, Bruno Zamborlin, and Fabrice Guédy, "Online Gesture Analysis and Control of Audio Processing," in *Musical Robots and Interactive Multimodal Systems*, Jorge Solis and Kia Ng, Eds., pp. 127–142. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[9] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[10] Jangwon Lee, Bardia Doosti, Yupeng Gu, David Cartledge, David Crandall, and Christopher Raphael, "Observing pianist accuracy and form with computer vision," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1505–1513.

[11] Christopher Raphael, "A hybrid graphical model for aligning polyphonic audio with musical scores," in *International Society for Music Information Retrieval Conference*, 2004.

[12] Ludwig van Beethoven, "Symphony No. 5 in C Minor," IMSLP, 2002, Fulda Symphonic Orchestra.