AN AUDIO-VISUAL FUSION PIANO TRANSCRIPTION APPROACH BASED ON STRATEGY

Xianke Wang *, *Wei Xu* †, *Juanting Liu*, *Weiming Yang*, *Wenqing Cheng*

Smart Internet Technology Lab
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan 430074, China
{M202072113, xuwei, juanting, M202072117, chengwq}@hust.edu.cn

ABSTRACT

Piano transcription is a fundamental problem in the field of music information retrieval. At present, a large number of transcriptional studies are mainly based on audio or video, yet there is a small number of discussion based on audio-visual fusion. In this paper, a piano transcription model based on strategy fusion is proposed, in which the transcription results of the video model are used to assist audio transcription. Due to the lack of datasets currently used for audio-visual fusion, the OMAPS data set is proposed in this paper. Meanwhile, our strategy fusion model achieves a 92.07% F1 score on OMAPS dataset. The transcription model based on feature fusion is also compared with the one based on strategy fusion. The experiment results show that the transcription model based on strategy fusion achieves better results than the one based on feature fusion.

1. INTRODUCTION

Piano transcription is a fundamental problem in the field of music signal processing and music information retrieval. It has a wide range of applications in music education, music creation and information retrieval. The complete piano transcription infers information about onset, offset, pitch, velocity and pedal from the audio signal and then obtains score-level representation. Due to transcription complexity, most of the current transcription models can only get accurate onset and pitch [1].

Audio-based transcription, video-based transcription and audio-visual fusion transcription are three methods of piano transcription. The current mainstream transcription models are based on audio, among which the Onsets and Frames model [2] is the most advanced model. Besides, video-based transcription models and audio-visual fusion transcription models have been gradually developed in recent years. In general, the keyboard registration is carried out in the video-based model [3] to obtain the segmented single-key images.Then the single-key images are fed into the classifier to get the transcription results. Compared with video-based and audio-based transcription, there are a few studies on audio-visual fusion transcription. Wan [4] proposed a novel piano-specific transcription system, using both audio and visual features for the first time. A new onset detection method was proposed using a specific spectrum envelope matched filter on multiple frequency bands. And a computer-vision method was proposed to enhance audio-only piano music transcription by tracking the pianist's hands on the piano keyboard. Lee [5] introduced a novel two-stream convolutional neural network that took video and audio inputs together for detecting pressed notes and fingerings. However, due to the lack of implementation details and open-source datasets, the actual performance of the models [4, 5] remains to be verified.

At present, researches based on video transcription and audio transcription are relatively sufficient, but the single-mode transcription models cannot achieve satisfactory results in actual performance scenario. Therefore, a piano transcription model based on strategy fusion is proposed, as shown in Figure 1, and the model's real performance is discussed in this paper. At the same time, the feature fusion idea of Lee [5] is also reconstructed. Finally, we compare the transcription models' performance based on feature fusion and strategy fusion.

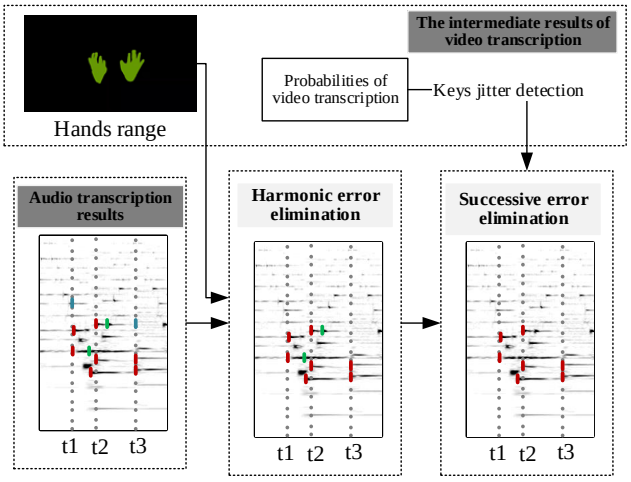


Figure 1: *Piano transcription based on strategy fusion.*

As shown in Figure 1, the audio transcription model only relies on the audio signal to predict notes, but the audio is a mixture of fundamental and harmonics. When a piano key is pressed, both the fundamental and harmonics have strong energy. As a result, audio models often detect harmonics as notes with a higher fundamental frequency, resulting in extra note detection errors. We refer this type of errors as harmonic errors. At the same time, the

* This work is supported by The National Natural Science Foundation of China (No. 61877060)

† Corresponding author

actual playing can be fast or slow. For successive rapid notes, the audio models are also prone to extra note detection errors of the same pitch We refer this type of errors as successive errors. In this paper, the transcription model based on strategy fusion takes the audio transcription results as the preliminary results and then uses the range of hands in the video to limit the pitch range of playing notes to eliminate harmonic errors. Besides, video-based transcription can detect key jitters in continuous playing, helping the audio model eliminate successive errors.

In order to study the effect of video transcription on audio transcription, the ablation studies of hand range and jitter detection are carried out. The experiment results show that the strategy fusion model with hand range and jitter detection achieves a 92.07% F1-score on OMAPS dataset, which is better than the single-mode transcription model.

The feature fusion transcription is also studied and compared with the transcription based on strategy fusion. Since the feature fusion model [5] directly takes the whole image as the video module's input, many interfering pixels will be brought in, which may have a negative impact. Therefore, the influence of full image, only keyboard region, only hand region and differential image as input on the feature fusion model is also studied. On OMAPS dataset, the F1 score of the strategy fusion model is 5.18% higher than that of the best model based on feature fusion, which proves the superiority of the strategy-based fusion model.

In general, the main contributions of this paper are as follows:

- A piano dataset, OMAPS dataset [1], is proposed for audio-visual fusion transcription research, which contains complete video, audio and aligned MIDI annotations.
- A piano transcription algorithm based on strategy fusion is proposed, and for the first time the hand range and jitter detection mechanisms is used for audio-visual fusion transcription.
- The input characteristics of the transcription model based on feature fusion is also studied. And compared with the best model based on feature fusion, the strategy fusion algorithm proposed in this paper is better.

2. RELATED WORK

2.1. Audio-based Transcription Research

Before deep learning became popular, NMF was often used for piano transcription. The simplest NMF [6, 7] first established the spectrum template through the single note signal, then used the spectrum template to decompose the piano signal to get the activation matrix, and finally obtained the transcription results through the post-processing of the activation matrix. However, simple NMF transcription methods do not consider the differences of signals at different stages, and the models' performance is often unsatisfying. Cheng [8] proposed an NMF transcription method based on attack-decay, using different spectral templates to decompose the attack and decay stages of the piano signal, and achieved an 81.80% F1 scores on MAPS dataset. However, NMF is a linear model and its modelling ability is limited, so the development of NMF in transcription soon reaches a bottleneck.

With the emergence of deep learning, piano transcription technology has been further developed. The most common input of the piano transcription models [9, 10, 11] based on deep learning is the two-dimensional spectrogram obtained by time-frequency transformation. Then the spectrogram representation is sent to the convolutional neural network for feature extraction. Finally, the results are obtained through the classifier. The transcription models based on CNN make up for the lack of modelling capability, but they don't consider the time dependence between sound signals. To solve this problem, Hawthorne [2] and Kong [12] proposed the piano transcription models based on CRNN. These models used CNN to extract the spectral features and then used Bi-LSTM to learn the time dependence. Compared with the model only using CNN, the CRNN models can capture the correlation of sound signals and achieve better transcription performance. In addition, there are transcription studies based on generative adversarial networks (GAN) [13, 14], language models [15, 16, 17], etc.

2.2. Video-based Transcription Research

At present, there are two kinds of video-based transcription methods. In the first kind of methods, a single key image is obtained by keyboard location and keyboard segmentation. These images are then fed into a binary classifier to get transcription results. However, the accuracy of keyboard segmentation is critical in this non-end-to-end transcription methods. Before the emergence of deep learning, traditional image processing methods such as Sobel operator, Hough transform, morphological expansion and corrosion [18, 19, 20] were generally used to complete keyboard location and keyboard segmentation. However, such methods are often sensitive to camera distortion and illumination changes, leading to the accuracy decline of the keyboard segmentation and ultimately affecting the entire transcription system's performance. Akbari [21, 22] added an illumination correction step in their pipeline, but the limitations for drastic light changes or vibrations of the camera or piano were reported. The methods based on deep learning make keyboard segmentation more robust. Li [3] proposed keyboard location and segmentation models based on semantic segmentation, making the model more tolerant to illumination changes and camera distortion. The second kind of methods is to directly send the whole image into the models for end-to-end learning. Since the transcription of a single image often ignores the association between successive frames, Koepke [23] has used 3D convolution for feature extraction to learn the temporal correlation between continuous video images. Besides, Rho [24] used a depth camera to capture the depth change and speed of the keypress, which provides a new idea for video-based piano transcription.

Single-key models reduce interference in images and focus on keypress detection. But there are also shortcomings, such as the lack of hand position perception and key position correlation learning. Full keyboard input retains the original hand position information. However, full image as input will bring a lot of irrelevant information to the models, which may interfere with the model's decision making.

2.3. Audio-visual Fusion Transcription Research

Piano transcription methods based on audio-visual fusion have been gradually developed in recent years. There are two methods to complete audio-visual fusion transcription. The first method is to use the transcription results of one mode to assist the transcription of another mode. For example, the audio-visual fusion transcription method proposed by Wan [4] adopted an energy envelope matching filter for the audio transcription part. Then hand

[1] https://github.com/itec-hust/OMAPS

information transcribed in the video was used to enhance the audio transcription results. Another fusion method is to fuse audio and video modes' feature matrices to get the final transcription results. For example, Lee [5] proposed a two-stream convolutional neural network that took video and audio together for detecting pressed notes and fingerings. Lee finally achieved an accuracy of 75.37% on its own piano dataset, which proved the feasibility of audio-visual fusion transcription.

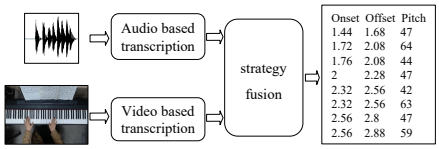


Figure 2: *The overall structure of our system.*

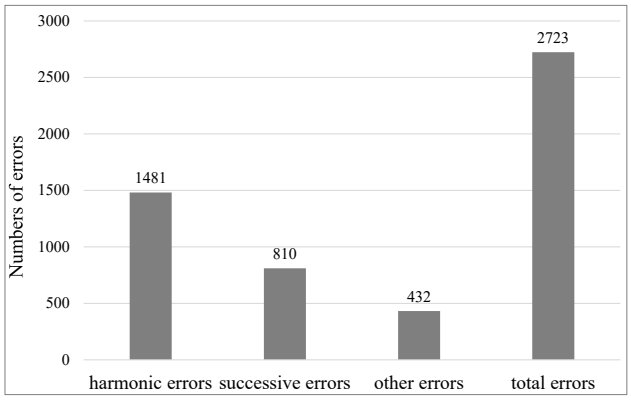


Figure 3: *Extra note detection errors' distribution.*

3. METHOD

Because the single-mode piano transcription models in the real scenario are not excellent, a fusion model based on strategy is proposed in this paper. The fusion model based on strategy utilizes the current optimal audio transcription model and video transcription model. The system's overall structure is shown in Figure 2, which is divided into three modules: video transcription, audio transcription, and strategy fusion. Each module will be introduced in detail below.

3.1. Audio transcription model

3.1.1. Implementation of audio transcription model

We use the Transition-aware model [25] as our audio-based transcription model. This model includes two branches: frame estimation and onset detection. The overall structure is shown in Figure 4. Frame estimation is used to predict the existence of 88 notes; Onset detection is used to predict the onset probabilities of 88 notes. The structure of onset detection and frame estimation branches is the same. Firstly, a multi-layer convolutional network is used to extract features from the input spectrogram. Then, Bi-LSTM is used to conduct time-dependent modeling. The onset detection branch features are fused into the frame estimation branch to improve the frame estimation branch's performance.

The structure of the Transition-aware model is similar to that of the Onsets and Frames model. The following improvements have been made to the Transition-aware model to achieve a better transcription performance:

1. Use CQT spectrogram instead of mel spectrogram as input. Compared with mel spectrogram, the frequency points of the CQT spectrogram are distributed exponentially and have different frequency resolutions for high and low frequencies, so it is more suitable for music signal processing.
2. The continuous five frames near the onset are annotated so that the model could better consider the spectral changes at the edge of the onset moment and improve the transcription accuracy.
3. Use peak selection to obtain onset transcription results. Compared with the taking threshold method, peak selection is more robust to noise interference and can reduce extra note detection errors.

3.1.2. Problems in audio transcription

We have tested the Transition-aware model on OMAPS dataset and achieved an F1 score of 87.51%. The number of extra note detection errors and missing note detection errors on OMAPS dataset of the Transition-aware model is 2723 and 2235, respectively, which indicates that extra note detection errors are the principal contradiction of the audio transcription model. We have analyzed the extra note detection errors' distribution on OMAPS dataset, as shown in Figure 3. It can be seen that there are mainly harmonic errors generated by harmonics and successive errors generated by continuous playing, which is consistent with the discussion in the introduction part. The audio-based transcription model performs well, but the remaining extra note detection errors are challenging to be solved by audio-based methods, which prompts us to propose a fusion model based on strategy.

3.2. Video transcription model

3.2.1. Implementation of video transcription model

We adopt the best video transcription model [3], which includes four components: keyboard location, keyboard segmentation, hands location and classifier. The overall structure is shown in Figure 5.

As can be seen from Figure 5, a complete input image contains the keyboard, hands and other interfering pixels. Putting the whole image directly into the neural network will bring in many interfering factors, which will affect the model's performance. The keyboard location is generally carried out first, and only the keyboard's filed is retained to improve the performance of the video transcription model. Then the coordinates of each key can be obtained according to the geometric relationship. On the other hand, the keyboard area covered by each playing action is limited to the hands' range. Therefore, the hand detection module is used to locate hands, and the approximate range of playing keys is obtained. We intercept the keyboard within the rectangular field of the hands as our detection range. At the same time, the coordinates of all keys are used for single key segmentation. Finally, a single key's image is sent into the binary classifier to get the detection results. The video transcription model uses differential images as input of the binary classifiers to obtain better transcription results. Considered the influence of optical factors, two independent classifiers are used for white keys and black keys, respectively, and the com-

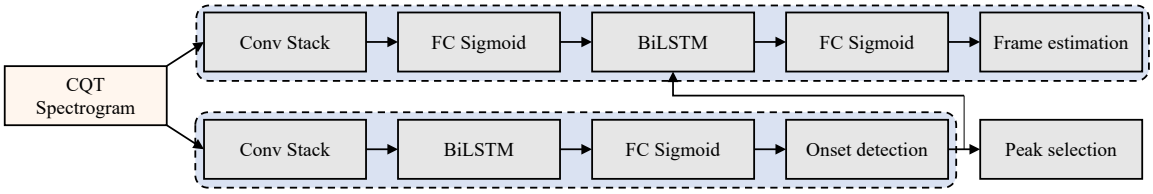


Figure 4: *Audio-based transcription model.*

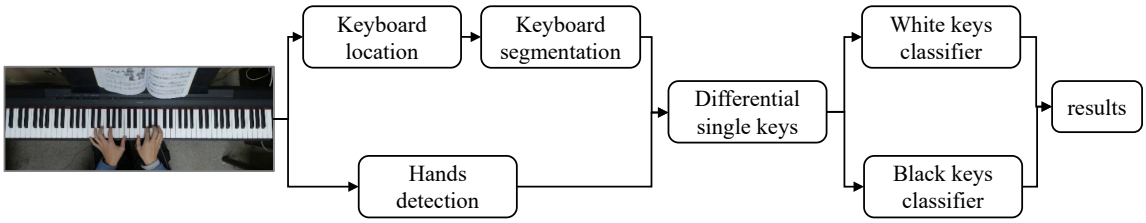


Figure 5: *Video-based transcription model.*

plete transcription results are finally obtained by combining them.

Keyboard location and hand detection are implemented using the semantic segmentation model, PSPNet [26], proposed by SenseTime company. To reduce the model's computational load, MobileNet-v2 [27] is used to accomplish feature extraction in PSPNet. After locating the keyboard, each key's position can be obtained based on the geometric relationship to achieve keyboard segmentation. Finally, the classifiers of black keys and white keys are implemented by the five-layer convolutional neural network, and the specific parameters are shown in Table 1. The convolution layer parameters, H*W@C, refer to the height of the convolution kernel as H, width as W and the number of channels as C. The max-pooling layer parameters, PH * PW/PSH * PSW, indicate that the height of the pooling area is PH, the width is PW, the step size along the height direction is PSH, and the step size along the width direction is PSW.

Table 1: *Parameters of the classifiers.*

Input	Layer & Parameter	Output
$112 \times 32 \times 1$	Convolution:3×3@8	$112 \times 32 \times 8$
$112 \times 32 \times 8$	Max-pool:$2 \times 2/2 \times 2$	$56 \times 16 \times 8$
$56 \times 16 \times 8$	Convolution:3×3@8	$56 \times 16 \times 8$
$56 \times 16 \times 8$	Max-pool:$2 \times 2/2 \times 2$	$28 \times 8 \times 8$
$28 \times 8 \times 8$	Convolution:3×3@16	$28 \times 8 \times 16$
$28 \times 8 \times 16$	Max-pool:$2 \times 2/2 \times 2$	$14 \times 4 \times 16$
$14 \times 4 \times 16$	Convolution:3×3@32	$14 \times 4 \times 32$
$14 \times 4 \times 32$	Max-pool:$2 \times 2/2 \times 2$	$7 \times 2 \times 32$
$7 \times 2 \times 32$	Reshape+Drop:0.5+Fc:256	256
256	Fc:2	2

3.2.2. *Problems with video transcription*

We have found that the transcription models based on video can detect the white keys well, but there are many missing note detection errors for the black keys. There are two reasons for this phenomenon. First, the hands will cover part of the keys during the playing process, while the color of the human skin is closer to the color of the black keys, which has a greater impact on the black keys. Second, a black gap will appear when a piano key is pressed. For white keys, the model can recognize the gap better when the piano keys are pressed. However, for black keys, the color of the gap is comparable to that of the black keys, so the model cannot identify the gap's characteristics well. These two factors lead to the detection accuracy of the black keys being much lower than that of the white keys, which ultimately leads to the video models' performance is inferior to that of audio models.

Although the model's overall performance based on video is not as good as that based on audio, some information in the video can assist the audio model in completing the transcription task better. For example, the video model's hand range judgment can help the audio transcription model remove harmonic errors. Meanwhile, the video transcription model can perform jitter detection for white key, which helps eliminate successive errors of the audio model. Both mechanisms are described in detail below.

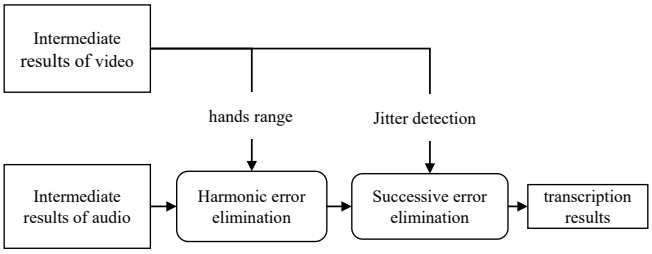


Figure 6: *Concrete implementation of our strategy.*

3.3. Strategy fusion

Our fusion model's basic idea based on strategy is to use the intermediate results of the video-based model to assist the audio-based model. As shown in Figure 6, the video-based model's hand range is used to eliminate harmonic errors, and the jitter detection results are used to eliminate successive errors.

The mechanism of using the hand range to eliminate harmonic errors is straightforward. We take the notes within the hand range detected by the video transcription model as the candidate notes. If

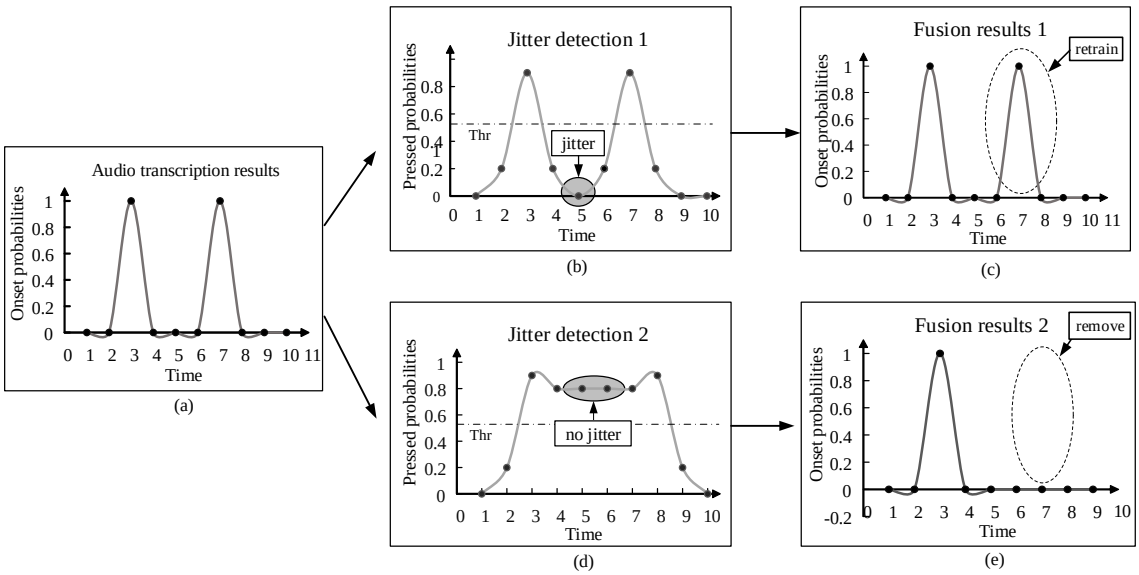


Figure 7: *Jitter is used to eliminate successive errors.*

the note onsets of the audio transcription results are not within the candidate notes, we consider them as harmonic errors and remove them from the final results. Using the jitter detection mechanism to eliminate successive errors is shown in Figure 7a. There may be successive detection of the same pitch in the audio transcription results. Some of these results represent continuous playing, while the other part represents successive errors. To distinguish, we use the detection probabilities of the video model to perform jitter analysis. As shown in Figure 7b, the video transcription model's output probabilities show a jitter between two probability peaks, indicating that the video model has detected continuous key playing. Therefore, the continuous detections in the audio model results are retained, and the final transcription result is shown in Figure 7c. As shown in Figure 7d, the video transcription model's output probabilities have no jitter between the two peaks, indicating that this is the playing action of one onset. Hence, the audio part's continuous detection results represent successive errors and will be removed later. The final result is shown in Figure 7e.

For the video transcription model, optical factors lead to a high error rate for black keys. Therefore, we carry out different fusion strategies for white keys and black keys:

1. For the black keys, we directly use the hand range to eliminate harmonic errors in the audio. Due to the low detection accuracy for the black keys, the jitter detection of the black keys cannot be well detected. Therefore, we do not conduct successive errors correction for the black keys.

2. For white keys, the video model has high accuracy, so we adopt a more refined processing strategy. Firstly, all the probabilities of white keys within the range of hand are obtained, and then white keys above the threshold are selected as candidate keys to eliminating harmonic errors. Then we use a jitter detection mechanism for white keys to eliminating successive errors.

4. EXPERIMENTS

4.1. OMAPS dataset

At present, the datasets used in piano transcription research include MAPS dataset [28], MAESTRO dataset [2] for audio transcription, and MTA dataset [20] for video transcription. However, the dataset used for audio-visual fusion transcription has not been proposed yet. To evaluate the performance of different audio-video fusion models, we have established the OMAPS dataset.

The OMAPS (Ordinary MIDI Aligned Piano Sounds) dataset was recorded from Yamaha electric piano P115 by a piano player. The Logitech C922 Pro HD stream webcam was used to record video and audio simultaneously. The Logitech camera is available in both 1080p/30fps and 720p/60fps video configurations. To ensure the resolution of the video, we used the 1080p/30fps configuration. The Logitech camera audio module's sampling rate is 44100Hz. Since the recorded videos and piano MIDI files were out of sync, we manually aligned the exported MIDI files as annotations. The OMAPS dataset contains 106 different pieces for a total of 216 minutes, with an average of two minutes per piece. The amount of notes played per second is used to measure the playing speed. According to the playing speed, the OMAPS dataset is divided into a training set and a test set. The training set and the test set have the same playing speed distribution. The training set contains 80 videos, and the test set contains 26 videos, as shown in Table 2.

Table 2: *Statistics of the OMAPS dataset.*

Split	Performance	Duration, minutes	Size, GB	Notes
Train	80	123	3.18	60,589
Test	26	53	1.03	19,135
Total	106	176	4.22	79,724

4.2. Evaluation metrics

Precision, recall and F1 score are used to evaluate the performance of the piano transcription models. Precision represents extra note detection errors, recall represents missing note detection errors, and F1 score represents the model's comprehensive performance. The calculation formula of precision, recall and F1 score is as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

Where TP is the number of correct detected notes, FP is the number of extra detected notes, and FN is the number of missed detected notes. At present, the evaluation algorithm implemented in mir_eval library [29] is commonly used to evaluate transcriptional models, and the time tolerance of onset is set to ±50ms.

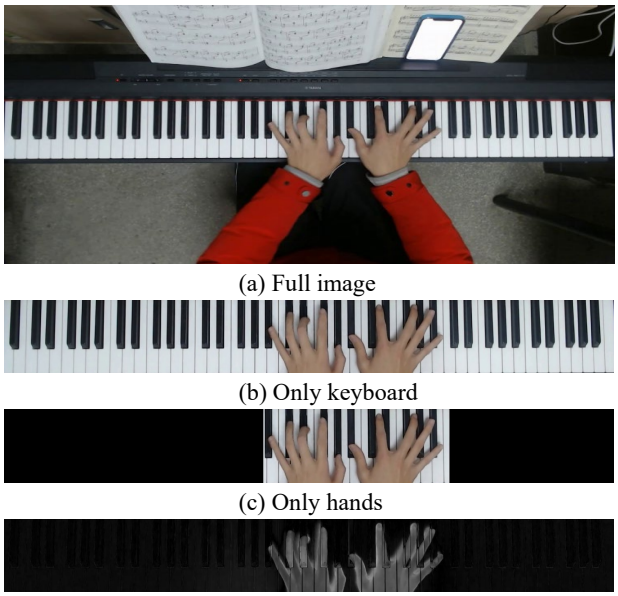
4.3. Study on ablation of strategy fusion

The fusion model based on strategy in this paper adopts a non-end-to-end method to conduct post-fusion of audio and video transcription results. The audio part adopts the Transition-aware model and the video part adopts Li's model. Due to many parameters in the Transition-aware model and to prevent overfitting of the model, we pre-trained the Transition-aware model using the MAESTRO dataset and then fine-tuned the model on OMAPS training set. The OMAPS training set contains more than 200,000 images, which is enough to train Li's video transcription model. So we directly use the OMAPS dataset to train the video transcription model.

To investigate hand range and jitter detection, we studied the effects of three model configurations: Transition-aware, Transition-aware combined with hand range, and Transition-aware combined with hand range and jitter detection. As shown in Table 3, the hand range eliminates many harmonic errors and increases precision by 8.6%, which is the main contribution improving of the model's performance. After combining jitter detection, the precision is only increased by 1.01%. On the one hand, there are a few successive errors, and the improvement of reducing successive errors is limited. On the other hand, the video transcription model has a poor jitter detection performance for black keys, so it is only implemented for white keys. At the same time, it can also be found that the recall of using the hand range and jitter detection mechanism is unchanged, which indicates that the proposed strategy rarely brings in new missing detection errors into the audio transcription results.

Table 3: *Performance of the three model configurations on OMAPS test set.*

Model	P	R	F1
Transition-aware	85.98	89.22	87.51
Transition-aware+Hands	94.58	88.56	91.40
Transition-aware+Hands+Jitter	95.59	88.94	**92.07**



(a) Full image

(b) Only keyboard

(c) Only hands

(d) Differential image

Figure 8: *Four types of input images.*

4.4. Input research of feature fusion

We replicated the idea of the feature fusion model [5] and conducted training and testing on OMAPS dataset. A complete image was directly fed into the feature fusion model, including many interfering pixels, so the transcription performance might not be perfect. To explore the potential of the feature-fusion based model, we investigated the effects of full-image input (Figure 8a), only-keyboard input (Figure 8b), only-hands input (Figure 8c) and differential input (Figure 8d), respectively. The full-image input (Figure 8a) is the input configuration in Lee's paper. The performance of each model on OMAPS test set is shown in Table 4.

Table 4: *The feature fusion model's performance on different input images.*

Configuration	P	R	F1
Full image	90.62	83.96	86.67
Only keyboard	92.83	84.26	87.74
Only hands	95.74	84.77	89.69
Differential	**96.15**	86.32	**90.22**

As shown in Table 4, only an 86.67% F1 score is obtained on OMAPS dataset when the entire image is directly used as the model input. The F1 score of the only-keyboard input reaches 87.74%, which indicates that reducing interference makes the model achieve better performance. The F1 score and precision of the only-hands input reaches 89.69% and 95.74% respectively, and the recall isn't decreased, indicating that only-hands input features can make the transcription model locate the playing pitch range more accurately. For differential input, the transcription model achieves an F1 score of 90.22% on OMAPS dataset and achieves the best performance among the feature fusion models. The differential image allows the model to directly focus on each finger, which is more accurate than the only-hands images and further re-

duces extra note detection errors. Besides, the differential image is not dependent on the hand detection module. This also avoids missing note detection errors caused by inaccurate hand detection, thus improving the recall rate.

4.5. Comparison with other methods

We tested Onsets and Frames, Li's model and the our strategy fusion model on OMAPS test set, and the experiment results are shown in Table 5.

Table 5: *Results of several transcription models on OMAPS test set.*

Model	P	R	F1
Onsets and Frames[2]	81.03	90.22	85.17
Li[3]	93.69	77.73	83.15
Feature fusion [5]	96.15	86.32	90.22
Strategy fusion	**95.59**	88.94	**92.07**

As can be seen from Table 5, the F1 score of Onsets and Frames is 2.02% higher than Li's video-based transcription model. The transcription models based on feature fusion and strategy fusion achieve 90.22% and 92.07% F1 scores on OMAPS dataset, respectively, which are higher than the single-mode transcription model based on video or audio. The fusion models' performance indicates that it is effective to combine video and audio information.

The strategy fusion model achieves a 92.07% F1 score on OMAPS dataset, which is the best among the above models. This model directly post-processes the audio transcription results to reduce the audio model's extra note detection errors. In this way, no missing detection errors are brought in, thus maintaining the advantage of the audio-based transcription model's high recall rate. In contrast, due to the hand occlusion problem in the video part and the same weight given to the image feature and the audio feature, new missing note detection errors are brought in for the feature fusion model. Its final recall is lower than that of our strategy fusion model. Compared with the best feature fusion model, our strategy fusion model's precision decreased slightly, recall rate increased by 2.62%, and F1 score increased by 1.85%, indicating that the current strategy fusion model's performance is better than that of the feature fusion model.

5. CONCLUSION

The OMAPS dataset is proposed for audio-video fusion transcription studies in this paper, consisting of 106 videos. Besides, a transcription model based on strategy fusion is also presented. The experiment results show that our strategy fusion model has achieved a 92.07% F1 score on OMAPS dataset, which is better than the feature fusion model [5]. Besides, the ablation studies of the hand range and the jitter detection mechanisms are conducted, proving the effectiveness of our proposed strategy fusion.

Piano transcription based on audio-visual fusion is a new research field, and there are a few relevant pieces of research at present. However, studies based on multimode is essential. We need machines to perceive the external world from a variety of perception simultaneously, just like humans. In the future, we will continue to study the audio-visual piano transcription and further explore the piano transcription models based on feature fusion.

6. ACKNOWLEDGMENTS

This work is supported by the national natural science foundation of China (no. 61877060). We also thank the player, Tianyue Yu, for recording the OMAPS dataset.

7. REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[2] C. Hawthorne, A. Stasyuk, and A. Roberts, "Enabling factorized piano music modeling and generation with the maestro dataset," in *Proceedings of International Conference on Learning Representations*, 2019.

[3] J. Li, W. Xu, Y. Cao, W. Liu, and W. Cheng, "Robust piano music transcription based on computer vision," in *Proceedings of the High Performance Computing and Cluster Technologies Conference & the International Conference on Big Data and Artificial Intelligence*, 2020, pp. 92–97.

[4] Y. Wan, X. Wang, R. Zhou, and Y. Yan, "Automatic piano music transcription using audio-visual features," *Chinese Journal of Electronics*, vol. 24, no. 3, pp. 596–603, 2015.

[5] J. Lee, B. Doosti, Y. Gu, D. Cartledge, D. Crandall, and C. Raphael, "Observing pianist accuracy and form with computer vision," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1505–1513.

[6] L. Gao, L. Su, and Y. H. Yang, "Polyphonic piano note transcription with non-negative matrix factorization of differential spectrogram," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 291–295.

[7] A. Cogliati, Z. Duan, and B. Wohlberg, "Piano transcription with convolutional sparse lateral inhibition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 392–396, 2017.

[8] T. Cheng, M. Mauch, and E. Benetos, "An attack/decay model for piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 584–590.

[9] Q. Wang, R. Zhou, and Y. Yan, "A two-stage approach to note-level transcription of a specific piano," *Applied Sciences*, vol. 7, no. 9, pp. 901–901, 2017.

[10] S. Liu, L. Guo, and G. A. Wiggins, "A parallel fusion approach to piano music transcription based on convolutional neural network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 391–395.

[11] R. Kelz, S. Böck, and G. Widmer, "Multitask learning for polyphonic piano transcription, a case study," in *Proceedings of IEEE International Workshop on Multilayer Music Representation and Processing*, 2019, pp. 85–91.

[12] Q. Kong, B. Li, , and X. Song, "High-resolution piano transcription with pedals by regressing onsets and offsets times," in *arXiv preprint arXiv:2010.01815*, 2020.

[13] A. Ycart, D. Stoller, and E. Benetos, "A comparative study of neural models for polyphonic music sequence transduction," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 470–477.

[14] J. W. Kim and J. P. Bello, "Adversarial learning for improved onsets and frames music transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 670–677.

[15] S. Sigtia, E. Benetos, and N. Boulanger-Lewandowski, "A hybrid recurrent neural network for music transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2061–2065.

[16] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[17] A. Ycart, A. McLeod, and E. Benetos, "Blending acoustic and language model predictions for automatic music transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 454–461.

[18] A. Goodwin and R. Green, "Key detection for a virtual piano teacher," in *Proceedings of the International Conference on Image and Vision Computing*, 2013, pp. 282–287.

[19] P. Suteparuk, "Detection of piano keys pressed in video," Tech. Rep., Dept. of Comput. Sci., Stanford Univ, 2014.

[20] M. Akbari, J. Liang, and H. Cheng, "A real-time system for online learning-based visual transcription of piano music," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 25513–25535, 2018.

[21] M. Akbari, *claVision: visual automatic piano music transcription*, Ph.D. thesis, University of Lethbridge, 2014.

[22] M. Akbari and H. Cheng, "Real-time piano music transcription based on computer vision," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2113–2121, 2015.

[23] A. S. Koepke, O. Wiles, Y. Moses, and A. Zisserman, "Sight to sound: An end-to-end approach for visual piano transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 1838–1842.

[24] S. Rho, J. I. Hwang, and J. Kim, "Automatic piano tutoring system using consumer-level depth camera," in *Proceedings of the IEEE International Conference on Consumer Electronics*, 2014, pp. 3–4.

[25] X. Wang, W. Xu, J. Liu, W. Yang, and W. Cheng, "Transition-aware: A more robust approach for piano transcription," in *Proceedings of the 23th International Conference on Digital Audio Effects (DAFx2020)*, 2021.

[26] H. Zhao, J. Shi, and X. Qi, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[28] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2009.

[29] C. Raffel, B. McFee, and E. J. Humphrey, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2014, pp. 367–372.

.