Supplementary Document

Anonymous WACV 2023 Algorithms Track submission

Paper ID 1616

1. Supplementary

1.1. Recap of Motivation

To decide on the piano scenario, we interviewed a number of pianists and teachers prior to this research to understand their needs, which was unfortunately omitted for the page limitations. Surprisingly, most of them are struggling with understanding the difference between themselves and recorded videos (or remote lessons, which are popular under the current pandemic), since the fixed angle and 2D visualization limit the training and teaching. They are athirst for a technology that can precisely estimate 3D poses from these videos to make an all-around comparison. However, models trained with a "generalized dataset" are far from satisfactory (as evident from the qualitative study) and cannot correctly estimate the 3D hand poses in piano playing. Therefore, a specialized dataset focusing on dexterous finger motions is required, which needs a heavy workload for annotation if using conventional methods. Therefore, we decide to study the possibility of translating Mocap-captured images into common barehand images. That is also the reason why all data in our dataset are taken from a top camera angle which is commonly used in piano instructions.

1.2. Detailed Explanation on Piano Task

In the present data collection described in the section *Piano*, we prepared a variety of musical excerpts that are played with either the right hand, left hand, or both hands. The basic criteria of the task selection were first to cover various hand

Figure 1. Examples of the collected pianists' hand postures.

postures that can involve transient occlusion of some finger joints from the cameras while playing, and second to cover a wide range of spatial locations of the keys to be played. Here, the tasks were categorized into two groups; the fundamental sets of tasks (L&R) that are considered to be basic modules of various musical pieces (e.g. scale, arpeggio, chord, trill), and musical excerpts of various musical pieces (B1-B11) that have been played by many pianists. Below is the name and the composer of the specific excerpts:

B1: Wolfgang Mozart: Piano Sonata K. 545 1st mov.
B2: Frederic Chopin: Etude Op.10 No.1
B3: Frederic Chopin: Etude Op.10 No.2
B4: Frederic Chopin: Etude Op.10 No.8
B5: Frederic Chopin: Etude Op.10 No.5
B6: Frederic Chopin: Etude Op.10 No.4
B7: Frederic Chopin: Etude Op.10 No.10
B8: Frederic Chopin: Etude Op.25 No.12
B9: Wolfgang Mozart: Piano Sonata K. 331 3rd mov.
B10: Wolfgang Mozart: 12 Variations on "Ah vous dirai-je, Maman", K. 265/300e
B11: Alexander Scriabin: 24 Preludes Op. 11 No.1

For both the scale and arpeggio tasks, two different excerpts for each task involve different fingering and different geometrical locations of the keys to be played (i.e. only white keys or a mixture of the black and white keys) over a wide space (i.e. 2 octaves). Also, these two tasks require different hand postures (i.e. closed or opened) with the thumb-under maneuver, where the thumb moves under the palm to dynamically change the hand location on the keyboard. In addition, we choose some chord tasks that are played with different combinations of multiple fingers (e.g. with all fingers, with different sets of two fingers) over different spatial locations (e.g. playing the same keys repetitively, playing different keys over time).

For musical excerpts derived from different music, the pieces are chosen from different musical styles (e.g. classical, romantic, and modern music, which corresponds to Mozart, Chopin, and Scriabin in this study, respectively). However, we mainly use the pieces of the etudes composed by Frederic Chopin (B2-B8), which are partly included in the CMU Panoptic Hand Dataset [1]. Firstly because it is a famous excerpt that is mandatory to play at most of the major piano competitions and entrance examinations of music conservatories, and secondly because the motor skills to play the etudes have been considered as a basis for playing a variety of music composed by different composers. Accordingly, the dataset of the finger motions to play such pieces can be generalizable for various types of music. In addition, different articulations (e.g. staccato, legato) can also require different finger postures (e.g. extended or flexed), and therefore pieces played with different articulations are included (e.g. 3 pieces by Wolfgang Mozart B1, B9, B10). Finally, we also included a task where part of the two hands sometimes overlap each other (B11).

The complete notes for all tasks are also attached in the supplementary documents (TaskNotes.pdf).

1.3. Details of Qualitative Experiment

The naturalness of the estimated hand poses in piano playing cannot be only told from the quantitative metrics. For example, an estimated result which is totally shifted 10 mm has the same MPJPE as another result where only one finger has a 50-mm error. It is difficult to discuss which of them is more correct for piano since the previous one shows the correct posture but plays totally different notes and the latter one vice versa. Therefore, we believe that it is necessary to combine objective opinions from experts by a qualitative study. A screenshot of the program for qualitative study is shown in Fig 2, and the whole procedures of the qualitative study are as follow:

1. A video clip obtain from online piano instruction is viewed by the subjects, together with the corresponding notes.
2. The video stops at a specific frame and the estimated 3D postures of the 4 models appear at the bottom in a counterbalanced order using Latin-Square.
3. The subjects can change the view point (angle of camera) freely by dragging a scrollbar.
4. The subjects repeat the above procedure until they decide their score for naturalness of the 4 estimated hand poses in 7-point-Likert-Scale.
5. In total, Step 1-4 are performed for 8 different videos (V01-V08) for each subject.

Overall, among the results of the eight clips of videos, the mean score of six videos suggests the proposed MR-Net-based method performs the best while is ranked the second best in the other two videos. The detailed scores of each video are shown in Table 1.

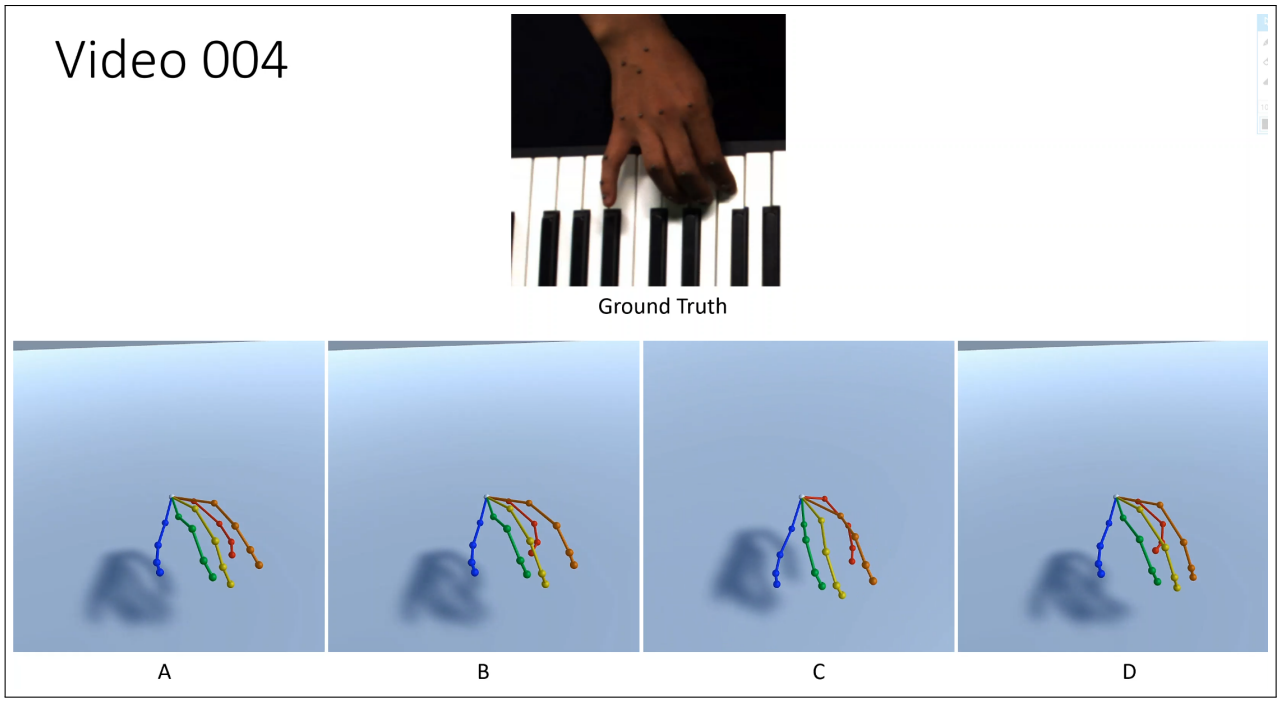

Figure 2. A screenshot of the program for the qualitative user study.

Method	V01	V02	V03	V04	V05	V06	V07	V08	Avg.
Zhou et al.	4.50	5.38	5.0	2.88	4.88	4.13	4.25	2.88	4.23
InterHand2.6M	3.75	3.63	**5.25**	4.00	3.50	2.88	4.13	4.25	3.93
Ours (Synthetic)	**5.13**	4.13	3.75	4.50	4.50	5.00	5.13	4.63	4.60
Ours (MR)	4.63	**5.63**	5.13	**5.25**	**5.68**	**5.38**	**5.25**	**5.50**	**5.31**

Table 1. Qualitative Study of 10 professional pianists' impression on estimated hand poses. (7-point Likert scale)

1.4. Example Dataset

Since the whole dataset cannot be uploaded due to its size and anonymity, we attached a tiny part of example data to show the contents as well as the data format. These data are captured from one of the subject playing *Etude Op.10 No.2*, which contain the following files:

1. A 9.5-second play of raw 1080p video data at 120-fps (1141 frames).
2. The corresponding 2D hand postures label file (every two columns are one 2D coordinates of a specific key point).
3. The corresponding 3D hand postures label file (every three columns are one 3D coordinates of a specific key point).
4. The corresponding MIDI generated from the pose data.
5. The bounding-box information of each frame (showing the upper left (x1,y1) and lower right (x2,y2) coordinates of both hands).
6. A video rendering the 2D results and bounding-box on the raw video.

The final released dataset will also include the processed training data which is down-sampled to 30-fps.

References

[1] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2