Visual Analysis of Conducting Gestures

Andrea Salgian
The College of New Jersey
Ewing, NJ
salgian@tcnj.edu

Laylah Burke
The College of New Jersey
Ewing, NJ
burkel4@tcnj.edu

Jeffrey Ernest
The College of New Jersey
Ewing, NJ
ernestj1@tcnj.edu

ABSTRACT

The analysis of a conductor's gestures throughout a musical piece has multiple applications, ranging from educational, where the system can provide feedback to a beginner practicing alone, to research, where it can provide statistical information about conductors whose video recordings are publicly available, and performance, where data about conducting can be used as input to devices that augment the musical performance. Previous conductor tracking systems were not able to rely on simple video input, and have used various motion capture devices that were sometimes intrusive. Other systems rely on extensive training data.

In this paper, we introduce an approach that relies on body pose landmark detection performed by the newly introduced Google MediaPipe API. Using simple video recordings, we detect undesired movements, such as swaying and mirroring, and extract tempo and time signature without the need for any training data. Our system works on any video, including previous recordings found on the internet.

1. INTRODUCTION

The act of musical conducting has always fascinated music lovers and researchers alike. Conductors communicate silently with orchestra musicians through hand gestures and facial expressions, conveying a variety of information that includes performance parameters and feedback. The orchestra becomes the conductor's instrument, and the conductor is therefore the only musician (maybe with the exception of the theremin player) who can play an instrument while moving their hands freely.

Conducting is easily understood by humans: its gestures are almost self-explanatory. Computers, however, have a significantly harder time, as visual gesture recognition is a hard problem in general. Previous work that has analyzed conducting relies on devices that can make conducting awkward, such as digital batons [1], Nintendo Wii remotes [2] and other inertial measurement units [3], or motion capture cameras such as the Microsoft Kinect [4]. Such approaches make data collection cumbersome, and prohibit analysis of videos that were prerecorded using simple cameras.

With recent advances in machine learning, gestures can easily be tracked in any video using software packages such as Google's MediaPipe [5]. MediaPipe tracks humans in videos and extracts skeletons similar to the ones extracted by the now obsolete Microsoft Kinect, and can do so on any video, whether it was just recorded with a smartphone, or downloaded from the internet.

In this paper, we present a system that can visually analyze conducting movements in any video, as long as the conductor is visible and facing the camera. We describe methods for detecting swaying, a common beginner conductor mistake that can be distracting for musicians, and mirroring, which may or may not be recommended depending on circumstances. We also describe a method for beat detection, which can then be used to determine conducting tempo, and finally a method for detecting the time signature of the piece based on the conducting pattern. Unlike previous approaches, our methods do not rely on training data, and our beat detection method does not use the time signature.

Our methods were designed with an educational approach in mind. Conductors are the only musicians without an individual instrument, and as such they cannot receive immediate feedback during individual practice. Beginner conductors often practice in front of a mirror, and have to judge their own performance, which can be difficult and subjective. Our system can provide the desired feedback, and can be further customized with thresholds. In addition, our approaches can be used to perform analysis on old conducting videos.

2. BACKGROUND AND RELATED WORK

Conducting students often practice alone, using either a recording or a metronome. However, conducting instructors acknowledge that practicing with a recorded sound source alone is not as effective. Studies have shown that students who practice with computer-based systems offering real-time feedback outperform those who rely only on audio recordings for self-practice [6].

Various music technology applications were developed to capture conducting gestures in order to conduct virtual orchestras [7, 8], or to augment musical performances [3]. Others were focused on music education, specifically on teaching conducting [9, 10].

Some of these systems utilized computer vision [7, 8], but faced challenges in tracking the conducting baton or hand. Others relied on batons equipped with sensors or emitters, such as the Digital Baton system developed by Marrin and Paradiso [1], or a Wii remote [2]. While these systems improved the tracking of conducting gestures, they required

conductors to use devices significantly different from the traditional, lightweight conducting baton. Baba et al. [11] developed a "Virtual Philharmony" incorporating various sensing devices, including a glove with an accelerometer, a baton with an infrared sensor, and a capacitance sensor. Their work, like several previous systems, focused on accurately extracting and reproducing the conducted tempo. Lim and Yeo [12] used a smartphone as a conducting baton and, despite acknowledging ergonomic issues with this approach, achieved promising results in extracting tempo and certain cueing gestures.

More recent approaches avoided the use of intrusive tracking devices by using the Microsoft Kinect [4, 9, 13]. Although the Kinect is now obsolete, it could be replaced by other depth sensing cameras. However, such cameras are not always easily available, and approaches relying on them cannot be used on existing video recordings, unless of course they were recorded with the same depth camera. Detailed motion capture protocols using a variety of devices are being used to generate high quality datasets for the study of conducting [14], but these systems are not accessible to the average music student.

Our approach uses Google's MediaPipe Pose Landmarker [5], which can detect body pose landmarks, that is, the location of the main body joints, and provide their three-dimensional coordinates. Figure 1 shows these joints.

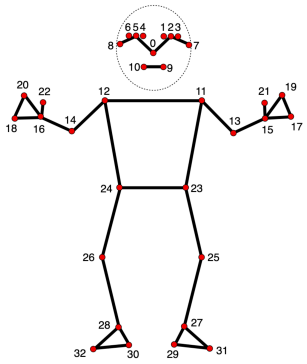

Figure 1. Body landmarks tracked by the Google MediaPipe pose landmarker [5].

We then use the trajectory of some of these joints in the video to extract various characteristics of the conducting performance.

3. METHODOLOGY

Using the coordinates of the hands and torso, we detect swaying and mirroring, and we extract tempo and time signature. Figure 2 shows a frame of our output video, overlaying the body skeleton provided by MediaPipe on top of the image, displaying the frame count, the tempo in beats per minute, and, in this case, signaling that a beat has been detected and mirroring is happening.

3.1 Swaying Detection

Swaying from side to side is a common mistake made by beginner conductors who may be nervous and shift their body weight from one leg to the other. Depending on the intensity, this can distract musicians, as it can interfere with the conducting gestures.

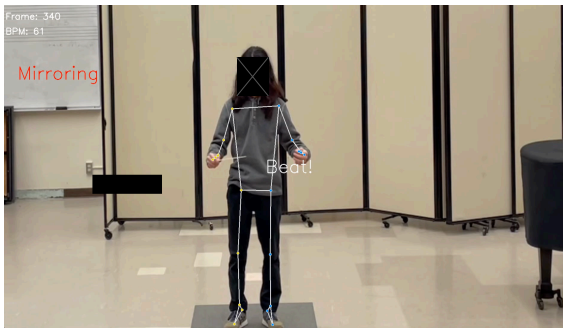

Figure 2. A frame output by our system.

To detect swaying, we track the x coordinate of the midpoint between the two shoulders, which should be relatively stable, and signal whenever this point moves too far from its initial location. Figure 3 shows the graph of the x coordinate of the point we are tracking over time in three different cases. The top graph shows a video where the person is not swaying. The location of the point is shown in blue, while the horizontal red lines show the thresholds between what we consider swaying and not swaying. Note that these thresholds may appear far away, but in fact the graph is zoomed in compared to the other two graphs. The middle graph shows a person swaying to the right, and then back to the left. Finally, the bottom graph shows a special case where the person shifted in the beginning, but then remained stable. This is not usually a problem, and an adaptive threshold that updates after a while could be used to address this situation.

3.2 Mirroring Detection

Mirroring is conducting the same pattern with both hands, using both hands to indicate the beat. While some mirroring might be a personal choice, acceptable for emphasis, formal education discourages it. Therefore we opted to signal every frame where mirroring is detected, and at the end we provide the percentage of time mirroring happened. Users would have the option to fine tune this feedback.

Mirroring happens when the two hands move symmetrically: they are at the same height (i.e., they have the same y coordinate), and are at the same horizontal distance from the center of the body. Figure 4 shows two different scenarios. The top graph shows the x coordinates of mirroring hands. We can see that the trajectories of the right and left hands, shown in blue and in red, are symmetrical, and the midpoint between the two closely follows the center of the chest. In the bottom graph, the left hand is mostly static, and therefore the midpoint between the hands follows the movement of the right hand, moving in and out from the center of the chest. This person is not mirroring.

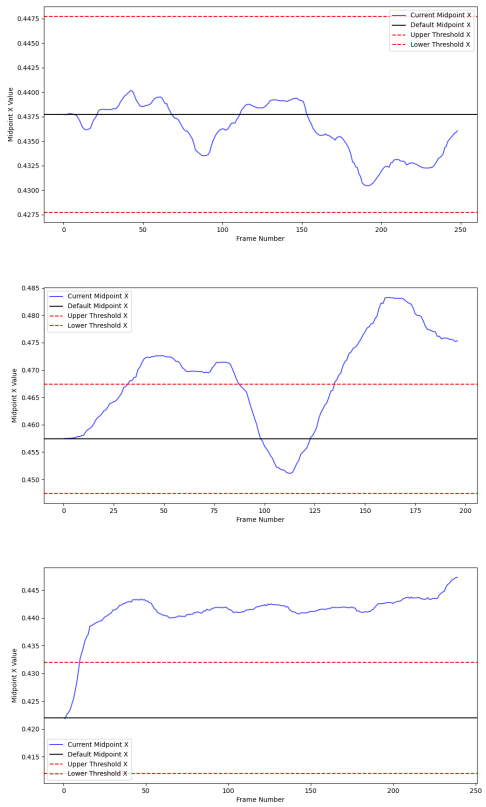

Figure 3. Detecting swaying. Top: not swaying. Middle: swaying right to left. Bottom: moving away from initial position.

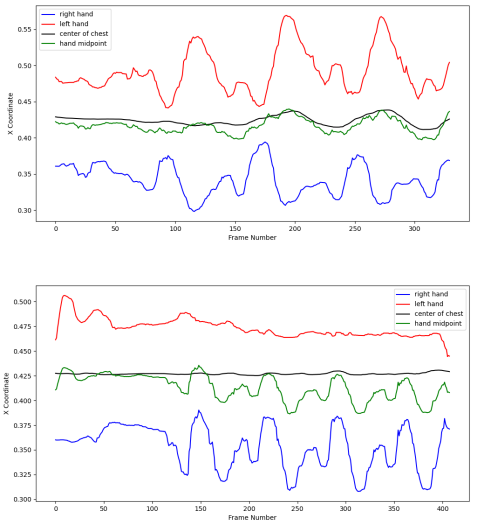

Figure 4. Mirroring detection. Top: Mirroring. Bottom: not mirroring.

3.3 Beat Detection and Tempo Calculation

To detect the beats from the conducting, we need to take a look at the conducting motion. For this work, we limited ourselves to the three most common time signatures: 2/4, 3/4, 4/4. Figure 5 shows their right hand patterns. We can see that, regardless of the time signature, the beat is signaled with an ictus where the right hand reaches a low point and starts to move upward. Therefore the beats should correspond to local minima in the y coordinate of the right hand.

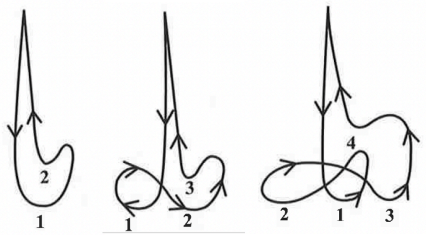

Figure 5. Right hand movement patterns for the most common time signatures: from left to right, 2/4, 3/4, and 4/4 [15].

Looking at the variation of the right hand y coordinate over time, we can see that the ictus is indeed always easy to detect, regardless of time signature. Figure 6 shows the x and y coordinates of the right hand in blue and green respectively, over time. The top graph is for a 2/4 time signature, the middle for 3/4, and the bottom for 4/4. In each case the local minima were easy to detect using a peak detection function, and we marked them on the graph with vertical purple lines. These lines correspond indeed to the instances when beats were conducted.

3.4 Time Signature Detection

Our time signature method relies on the periodicity and general shape of the conducting gestures. Looking at Figure 5, we can see that for each time signature, while the moment of each beat is marked by a local minimum in the y coordinate of the hand, the gesture corresponding to each beat is delimited by two local maxima of the y coordinate. If the conductor follows the textbook pattern, then the hand should always return to the coordinates from the previous measure, at least for the points delimiting the beats. In Figure 6 we can see that while we concentrated on local

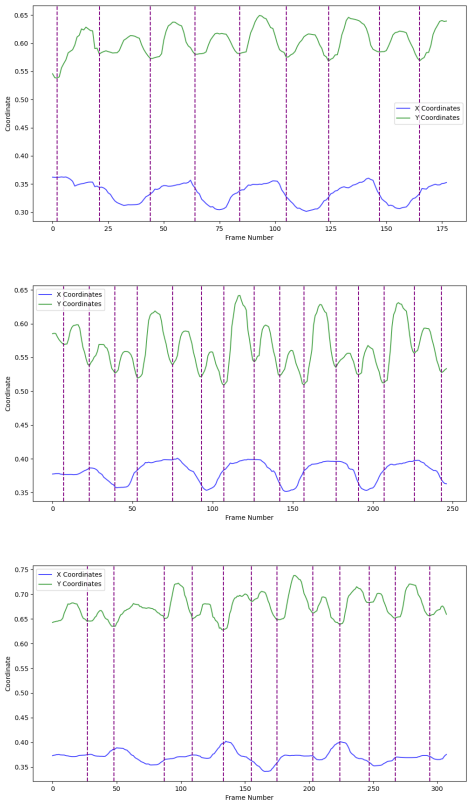

Figure 6. Right hand coordinates over time for the three time signatures: 2/4 (top), 3/4 (middle), 4/4 (bottom).

minima for the beats, local maxima are also easy to extract. Figure 7 shows the locations of these points in the x/y plane for a video with a 3/4 time signature. We can see that the points are easily grouped into three clusters that delineate a triangle that the conductor was following while conducting the piece. The number of clusters gives us the number of beats in the pattern, and therefore the time signature, 3/4.

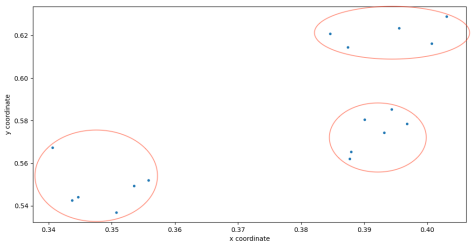

Figure 7. Hand coordinates in a 3/4 time signature.Clusters can be seen easily.

But not all conductors follow the textbook pattern. These could either be beginners who are having trouble, or advanced conductors who usually move away from the textbook and adopt a more personal style. Regardless of the reason, the hand coordinates can't be clustered that clearly anymore. Figure 8 shows such a case.

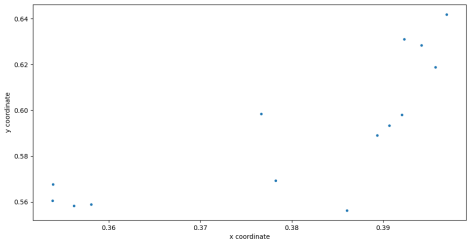

Figure 8. Hand coordinates in a 3/4 time signature. No clear clusters can be extracted.

However, even in this case, conductors will want to mark the beat with an ictus, and clearly delimit one beat from the next, so peaks and valleys in the y coordinate should still be visible. Indeed, Figure 9 shows the x and y coordinates over time for the video from Figure 8.

The difference between how the downbeat and the upbeat are conducted makes the beginning of each measure easy to detect. Each measure will have a number of peaks equal to the number of beats in the measure, and the first peak (between the downbeat and the upbeat) will always be the highest. Therefore, if we find the highest peaks and their

periodicity, we should be able to determine the number of beats in a measure and therefore the time signature. In Figure 9 we marked the highest peaks with a red dot. We can see that the red dot appears at every third beat, and indeed, this graph was generated from a video with a 3/4 time signature. The x coordinate also shows a periodic pattern, but the number of beats can only be extracted using the y coordinate.

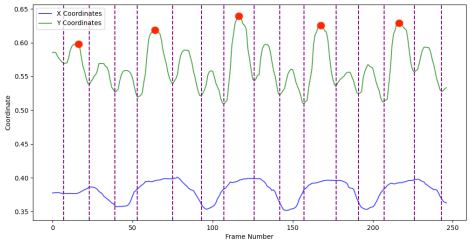

Figure 9. x and y coordinates over time, with the beginning of the first beat marked in each measure.

Depending on the purpose of the application, this approach can be used to signal that the time signature is correct, but the gestures need to be improved.

4. RESULTS

We tested our methods on a variety of videos. Some were obtained by recording conducting students at the beginning of their second college level conducting course, such as the one shown in Figure 2, while others were generic conducting videos scraped from YouTube (Figure 10. MediaPipe was able to track the skeleton of each conductor even if the legs were not shown in the frame.

Figure 10. Output from a video scraped from YouTube.

Our system was able to successfully detect swaying and mirroring, extract the beats, and calculate the tempo. Time signature extraction worked for all those videos where the time signature did not change during conducting, and the gestures were fairly clear. Changes in time signature will have to be indicated manually by the user.

As mentioned in the previous section, our method can also be used to indicate that the gestures used do not follow the textbook.

5. CONCLUSION AND FUTURE WORK

Using Google Mediapipe's pose landmark detection task, we were able to devise a variety of algorithms that can analyze conducting gestures in videos without requiring any additional motion tracking devices. Specifically, our methods can detect swaying and mirroring, can extract the beats and calculate the tempo, and can recognize the time signature if the conducting gestures are consistent.

Our approach works on a variety of videos, ranging from smartphone recordings to older videos found online. Although the main purpose of our project is educational, to help beginners practice conducting, our methods can be used for other purposes as well, such as conducting motion analysis for research, or augmenting the musical performance with other effects.

Future work includes analyzing other conducting patterns for other time signatures and changes in articulation, as well as looking at the left hand and cueing gestures.

Acknowledgments

The authors thank Dr. David Vickerman for providing video recordings and valuable professional conducting advice.

6. REFERENCES

[1] J. P. T. Marrin, "The digital baton: A versatile performance instrument," in *Proceedings of the International Computer Music Conference*, Thessaloniki, Greece, 1997, pp. 313–316.

[2] T. M. Nakra, Y. Ivanov, and C. Ault, "The UBS Virtual Maestro: an Interactive Conducting System," in *Proceedings of the New Interfaces for Musical Expression (NIME) Conference*, Pittsburgh, PA, 2009.

[3] S. Lemouton, R. Borghesi, S. Haapamäki, F. Bevilacqua, and E. Fléty, "Following Orchestra Conductors: the IDEA Open Movement Dataset," in *Proceedings of the 6th International Conference on Movement and Computing*, Tempe, AZ, USA, 2019.

[4] A. Sarasua, "Context-aware gesture recognition in classical music conducting," in *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, 2013, p. 1059–1062.

[5] "Google AI Edge - Gemini API, MediaPipe Solutions Guide," https://ai.google.dev/edge/mediapipe/solutions/guide, 2024, accessed: (January 10, 2025).

[6] D. Hollinger and J. M. Sullivan, "The Effects of Technology-Based Conducting Practice on Skill Achievement in Novice Conductors." *Research and Issues in Music Education*, vol. 5, no. 1, 2007.

[7] R. Behringer, “Conducting Digitally Stored Music by Computer Vision Tracking,” in *Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS’05)*, Florence, Italy, 2005.

[8] D. Murphy, T. H. Andersen, and K. Jensen, “Conducting audiofiles via computer vision,” in *Proceedings of the 5th International Gesture Workshop, LNAI*, Genoa, Italy, 2003, pp. 529–540.

[9] A. Salgian, D. Vickerman, and D. Vassallo, “A Smart Mirror for Conducting Exercises,” in *Proceedings of the Thematic Workshops of ACM Multimedia 2017*, Mountain View, CA, 2017.

[10] A. Rosinski, “Digital Technologies in Teaching Conducting,” *PUPIL: International Journal of Teaching, Education and Learning*, vol. 6, no. 3, pp. 57–67, 2023.

[11] T. Baba, M. Hashida, and H. Katayose, “Virtual Philharmony,” in *Proceedings of the New Interfaces for Musical Expression (NIME) Conference*, Sydney, Australia, 2010.

[12] Y. K. Lim and W. S. Yeo, “Smartphone-based Music Conducting,” in *Proceedings of the New Interfaces for Musical Expression (NIME) Conference*, London, UK, 2014.

[13] G.-F. Chen, “Recognition of Beat-Motion Gestures of Orchestra Conductor using DTW and Nearest Neighbor Method,” in *Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms*, Beijing, China, 2023, p. 59–65.

[14] R. Polfreman, B. Oliver, D. Halford, and C. Metcalf, “Force & Motion: Conducting to the Click,” in *Proceedings of the 6th International Conference on Movement and Computing*, ser. MOCO ’19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3347122.3347139

[15] J. Nowak, *Conducting the Music, Not the Musicians*. Carl Fischer, New York, 2002.