Follow the Beat?
Understanding Conducting Gestures from Video

Andrea Salgian[1], Micheal Pfirrmann[1], and Teresa M. Nakra[2]

[1] Department of Computer Science
[2] Department of Music
The College of New Jersey
Ewing, NJ 08628
salgian@tcnj.edu, micheal.pfirrmann@gmail.com, nakra@tcnj.edu

Abstract. In this paper we present a vision system that analyzes the gestures of a noted conductor conducting a real orchestra, a different approach from previous work that allowed users to conduct virtual orchestras with prerecorded scores. We use a low-resolution video sequence of a live performance of the Boston Symphony Orchestra, and we track the conductor's right hand. The tracker output is lined up with the output of an audio beat tracker run on the same sequence. The resulting analysis has numerous implications for the understanding of musical expression and gesture.

1 Introduction

In recent years, numerous artistic and expressive applications for computer vision have been explored and published. Many of these have been for dance, whereby moving dancers trigger various visual and audio effects to accompany their movements [1, 2]. However, there is a small but growing area in which purely musical applications are being researched. In this area, musical conductors are frequently featured, perhaps because conductors are the only musicians who freely move their hands to create sound and whose gestures are not constrained by a rigid instrument.

Several computer-based conducting recognition systems have also relied on on tracking batons equipped with sensors and/or emitters. Most notably, the *Digital Baton* system implemented by Marrin and Paradiso [3], had an input device that contained pressure and acceleration sensors, and the tip of the baton held an infrared LED which was tracked by a camera with a position-sensitive photodiode.

Examples of prior "pure vision" applications featuring musical conducting include the work by Wilson and Bobick [4]. Their system allowed the general public to "conduct" by waving their hands in the air and controlling the playback speed of a MIDI-based orchestral score.

In another project, Bobick and Ivanov [5] took that concept further by identifying a typical musical scenario in which an orchestra musician would need to visually interpret the gestures of a conductor and respond appropriately.

More recently, Murphy *et al.* [6] developed a computer vision system for conducting audio files. They created computer vision techniques to track a conductor's baton, and analyzed the relationship between the gestures and sound. They also processed the audio file to track the beats over time, and adjusted the playback speed so that all the gesture beat-points aligned with the audio beat points.

Until recently, these vision-based techniques aimed at systems that would allow real conductors (and sometimes the general public) to conduct virtual orchestras, by adjusting effects of a prerecorded score. In contrast, the *Conductor's Jacket* created by Nakra [7] was designed to capture and analyze the gestures of a real conductor conducting a real orchestra. However, the input of this system was not visual. Instead, it contained a variety of physiological sensors including muscle tension, respiration and heart rate monitors.

In this paper we take the first steps towards devising a computer vision system that analyzes a conductor conducting a real orchestra. This is an important distinction from earlier work, because a professional conductor reacts differently when conducting a real (versus a virtual) orchestra. His/her motivation to perform the authentic gestures in front of the human orchestra can be assumed to be high, since the human orchestra will be continuously evaluating his/her skill and authenticity. There is scientific value in understanding how good conductors convey emotion and meaning through pure gesture; analysis of real conducting data can reveal truth about how humans convey information non-verbally.

We analyze the low-resolution video footage available from an experiment with an updated version of the *Conductor's Jacket*. We track the right hand of the conductor and plot its height as the music progresses. The vertical component of the conductor's hand movements, together with the beat extracted from the score, enables us to make several interesting observations about musical issues related to conducting technique.

The rest of the paper is organized as follows. In Section 2 we describe the background of our work. In Section 3 we present the methodology for tracking the conductor's hand. In Section 4 we discuss our results. Finally, we conclude and describe future work and possible applications in Section 5.

2 Background

Motivated by prior work, we undertook a joint research project to investigate the gestures of a noted conductor. Our goal was to use computer vision techniques to extract the position of the conductor's hands. Our source video footage featured the Boston Symphony Orchestra and conductor Keith Lockhart. This footage was obtained during a 2006 collaborative research project involving the Boston Symphony Orchestra, McGill University, Immersion Music, and The College of New Jersey. The Boston Symphony Orchestra and McGill University have given us the rights to use their video and audio for research purposes.

Figure 1 shows conductor Keith Lockhart wearing the measuring instruments for this experiment.

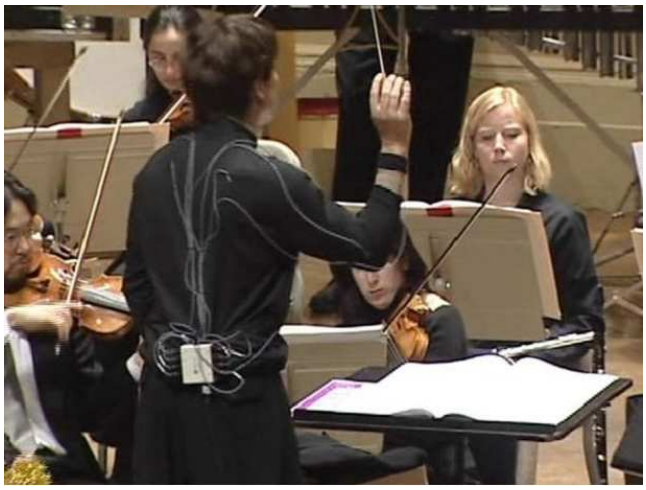

Fig. 1. Conductor Keith Lockhart wearing the measuring instruments (Photo credit: KSL Salt Lake City Television News, April 21, 2006).

The video sequence contains a live performance of the Boston Symphony Orchestra, recorded on April 7, 2006. The piece is the Overture to "The Marriage of Figaro" by W.A. Mozart, and our video has been edited to begin at the second statement of the opening theme. (The reason for the edit is that the beginning of the video features a zoom-in by the camera operator, and the first several seconds of the footage were therefore unusable. This segment begins at the moment when the zoom stopped and the image stabilized.) Figure 2 shows a frame from the video sequence that we used.

Given that image processing was not planned at the time of the data collection, the footage documenting the experiment is the only available video sequence. Hence, we were forced to work with a very low resolution image of the conductor that we cropped from the original sequence (see Figure 3).

Given the quality of the video, the only information that could be extracted was the position of the right hand. It is known that the tempo gestures are always performed by either the conductor's right hand or both hands, and therefore right hand following is sufficient to extract time-beating information at all times [8]. What makes tracking difficult is the occasional contribution of the right hand to expressive conducting gestures, which in our case lead to occlusion.

Our next task was to look at the height of the conductor's right hand - the one that conducts the beats - with the final goal of determining whether it correlated with musical expression markings and structural elements in the score. We have found that indeed, it does. The height of Keith Lockhart's right hand increases and decreases with the ebb and flow of the music.

Fig. 2. A frame from the input video sequence.

Fig. 3. The frame cropped around the conductor.

3 Methodology

As described in the previous section, the frames of the original video were cropped to contain only the conductor. The crop coordinates were chosen manually in the first frame and used throughout the sequence. The frames are then converted to grayscale images. Their size is 71x86 pixels.

The average background of the video sequence is computed by averaging and thresholding (with a manually chosen threshold) the frames of the entire sequence. This image (see Figure 4) contains the silhouettes of light (skin colored) objects that are stationary throughout the sequence, such as heads of members of the orchestra and pieces of furniture.

Fig. 4. The average video background.

For each grayscale image, the dark background is estimated through morphological opening using a circle of a radius 5 pixels as the structural element. This background is then subtracted from the frame, and the contrast is adjusted through linear mapping. Finally, a threshold is computed and the image is binarized. The left side of Figure 5 shows a thresholded frame. This image contains a relatively high number of blobs corresponding to all the lightly colored objects in the scene. Then the average video background is subtracted, and the number of blobs is considerably reduced. We are left with only the moving objects. An example can be seen on the right hand side of figure 5.

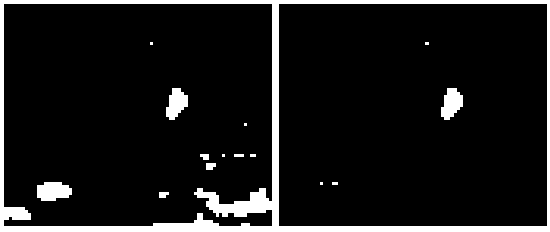

Fig. 5. Thresholded frame on the left, same frame with the average video background removed on the right.

While in some cases background subtraction alone is enough to isolate the conductor's right hand, in other cases, other blobs coming from the conductor's left hand or members of the orchestra can confuse the result. Figure 6 shows such an example.

Fig. 6. Another frame, after thresholding, and background subtraction.

In the first frame, the correct object is picked by the user. In subsequent frames the algorithm tracks the hand using the position detected in the previous frame. More specifically, the coordinates of the object that is closest to the previous position of the hand are reported as the new position. If no object is found within a specified radius, it is assumed that the hand is occluded and the algorithm returns the previous position of the hand. Figure 7 shows a frame with the position of the hand marked.

Fig. 7. Tracked hand.

We then plot the vertical component of the position of the conductor's hand. Based on the conductors' gestures, the local minima and maxima should correspond the tempo of the music being played. To verify this, we extracted the beats from the score using an algorithm developed by Dan Ellis and Graham Poliner [9] that uses dynamic programming. We marked the beat positions on the same plot and generated an output video containing the cropped frames and the a portion of the tracking plot showing two seconds before and after the current frame. Figure 8 shows a frame from the output video sequence.

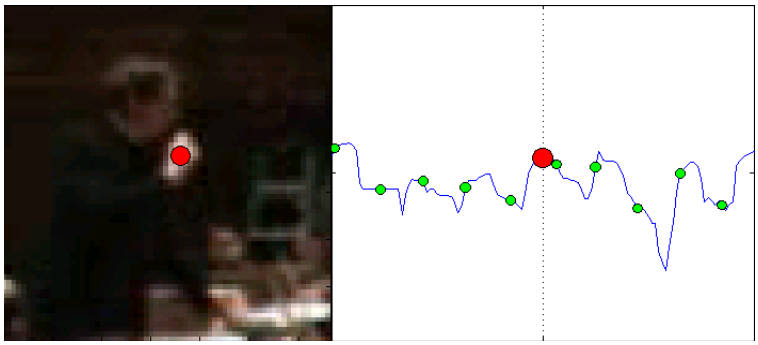

Fig. 8. A frame from the output video sequence.

The left side of the image contains the cropped frame and the detected position of the hand. The right side consists of three overlaid items:

1. a vertical dotted line with a red dot, indicating the intersection of the current moment with the vertical height of Keith Lockhart's right hand.
2. a continuous dark line indicating the output of the hand tracker, giving the vertical component of Keith Lockhart's right hand
3. a series of green dots, indicating the places in time when the audio beat tracker determined that a beat had occurred

4 Results

To analyze the performance of our tracker, we manually tracked the conductor's right hand in the first 500 frames of the video and compared the vertical component with the one extracted by the algorithm. 421 of 500 frames (over 84%) had a detection error of less than 2 pixels. In 24 out of remaining 79 frames the tracking could not be performed manually due to occlusion.

Figure 9 shows the ground truth and the detected y coordinate in the first 500 frames. Ground truth coordinates that are lower than 10 pixels correspond to frames where the hand could not be detected manually. Horizontal segments in the detected coordinates correspond to frames where no hand was detected. In these situations the tracker returns the position from the previous frame.

In the relatively few situations where the tracker loses the hand, it has no difficulty reacquiring it automatically.

Figure 10 shows the vertical component of the right hand position in blue, and the beats detected in the audio score in red. It may seem surprising that there is a delay between the local extrema of the conductor's hand and the audio beats. This is mostly due to the fact that there is a short delay between the time the conductor conducts a beat and the orchestra plays the notes in that beat. (This well-known latency between conductor and orchestra has been quantified in [10] to be 152 +/- 17 milliseconds, corresponding to one quarter of a beat at

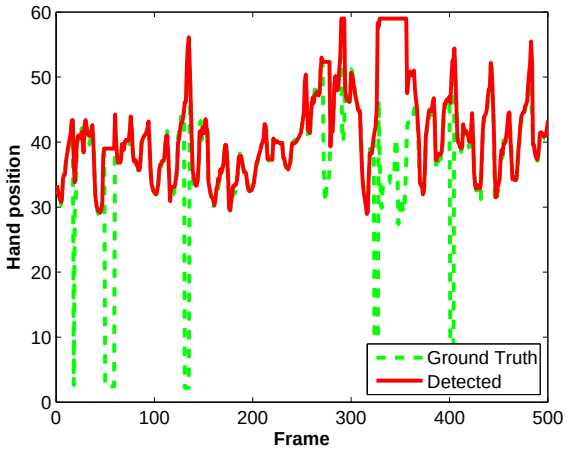

Fig. 9. Tracking performance on the first 500 frames.

100 beats per minute. This study is based upon conductors listening and reacting to a recording, which may have biased the data.) It should also be noted that in the current study, there are places where the conductor's beats are not in phase with the orchestra. It may be assumed that in such places, the conductor is not needed for routine "traffic cop"-type time beating, but rather motivating the orchestra to increase (or decrease) its rate of tempo change.

Using all the visual information provided by the various streams in the video, a musician can make several interesting observations about musical issues related to conducting technique. While these observations strictly refer to the technique of Keith Lockhart, nonetheless it can be assumed that some of these features may also be used by other conductors, perhaps in different ways. Some of the conducting features revealed by our method are as follows:

1. Tiered gesture "platforms" - Lockhart seems to use local "platforms" (horizontal planes) of different heights at different times in the music. The choice of what height to use seems to be related to the orchestration and volume indicated in the music.
2. Height "delta" - at certain times, the height difference between upper and lower inflection points changes. This seems also to be related to expressive elements in the score - particularly volume and density.
3. Smooth versus jagged beat-shapes - sometimes the beats appear almost sinusoidal in their regularity, whereas other times the shape of the peak becomes very jagged and abrupt with no rounding as the hand changes direction. This feature also appears to be controlled by the conductor, depending upon elements in the music.
4. Rate of pattern change - sometimes a particular feature stays uniform over a passage of music, sometimes it gradually changes, and sometimes there are

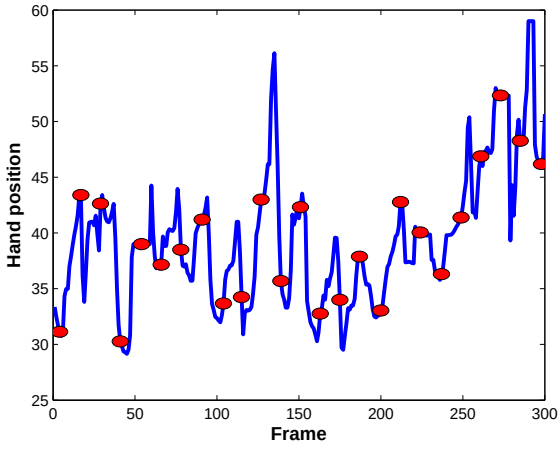

Fig. 10. Hand position and beat in the first 300 frames.

abrupt changes. The quality of the change over time seems also to be related to signaling the musicians about the nature of upcoming events.

5 Conclusions and Future Work

We presented a system that analyzes the gestures of a conductor conducting a real orchestra. Although the quality of the footage was poor, with very low resolution and frequent self-occlusions, we were able to track the conductor's right hand and extract its vertical motion. The tracker was able to reacquire the hand after losing it, and we obtained a recognition rate of 84% on the first 500 frames of the sequence. We annotated these results with the beats extracted from the audio score of the sequence. The data we obtained proved to be very useful from a musical point of view and we were able to make several interesting observations about issues related to conducting technique.

There is much more work to be done in this area. Very little is known about professional conductors' gestures, and it is hoped that with more research some interesting findings will be made with regard to musical expression and emotion. Our next task will be to compare our results with those of the other (physiological) measurements taken during the experiment.

Additional data collections with higher quality video sequences will allow us to devise better algorithms that could track the conductor's hand(s) more accurately and extract a wider range of gestures.

Results of future work in this area are targeted both for academic purposes and beyond. For example, conductor-following systems can be built to interpret conducting gestures in real-time and cause the conductor to control various media streams in synchrony with a live orchestral performance. (Lockhart himself

has agreed that it would be fun to be able to control the fireworks or cannons on the 4th of July celebrations in Boston while conducting the Boston Pops Orchestra.) Human computer interfaces could also benefit from understanding the ways in which expert conductors use gestures to convey information.

Acknowledgments

The authors would like to thank the Boston Symphony Orchestra and conductor Keith Lockhart for generously donating their audio and video recording for this research. In particular, we would like to thank Myran Parker-Brass, the Director of Education and Community Programs at the BSO, for assisting us with the logistics necessary to obtain the image and sound. We would also like to acknowledge the support of our research collaborators at McGill University: Dr. Daniel Levitin (Associate Professor and Director of the Laboratory for Music Perception, Cognition, and Expertise), and Dr. Stephen McAdams (Professor, Department of Music Theory, Schulich School of Music).

References

1. Paradiso, J., Sparacino, F.: Optical tracking for music and dance performance. In: Fourth Conference on Optical 3-D Measurement Techniques, Zurich, Switzerland (1997)
2. Sparacino, F.: (Some) computer vision based interfaces for interactive art and entertainment installations. INTER-FACE Body Boundaries **55** (2001)
3. Marrin, T., Paradiso, J.: The digital baton: A versatile performance instrument. In: International Computer Music Conference, Thessaloniki, Greece (1997) 313–316
4. Wilson, A., Bobick, A.: Realtime online adaptive gesture recognition. In: International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece (1999)
5. Bobick, A., Ivanov, Y.: Action recognition using probabilistic parsing. In: Computer Vision and Pattern Recognition, Santa Barbara, CA (1998) 196–202
6. Murphy, D., Andersen, T.H., Jensen, K.: Conducting audio files via computer vision. In: 5th International Gesture Workshop, LNAI, Genoa, Italy (2003) 529–540
7. Nakra, T.M.: Inside the Conductor's Jacket: Analysis, Interpretation and Musical Synthesis of Expressive Gesture. PhD thesis, Media Laboratory, MIT (2000)
8. Kolesnik, P.: Conducting gesture recognition, analysis and performance system. Master's thesis, McGill University, Montreal, Canada (2004)
9. Ellis, D., Poliner, G.: Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In: Proc. Int. Conf. on Acous., Speech, and Sig. Proc. ICASSP-07, Hawaii (April 2007) 1429–1432
10. Lee, E., Wolf, M., Borchers, J.: Improving orchestral conducting systems in public spaces: examining the temporal characteristics and conceptual models of conducting gestures. In: Proceedings of the CHI 2005 conference on Human factors in computing systems, Portland, Oregon (2005) 731–740