Motion Analysis of Music Ensembles with the Kinect

Aristotelis Hadjakos
Zentrum für Musik- und
Filminformatik
HfM Detmold / HS OWL
Hornsche Straße 44
32756 Detmold, Germany
hadjakos@hfm-detmold.de

Tobias Großhauser
Electronics Laboratory
ETH Zürich
Gloriastrasse 35
8092 Zürich
tobias@grosshauser.de

Werner Goebl
Institute of
Music Acoustics (IWK)
University of Music and
Performing Arts Vienna
Austrian Research Institute for
Artificial Intelligence
Vienna, Austria
goebl@mdw.ac.at

ABSTRACT

Music ensembles have to synchronize their performances with highest precision in order to achieve the desired musical results. For that purpose the musicians do not only rely on their auditory perception but also perceive and interpret the movements and gestures of their ensemble colleges. In this paper we present a method for motion analysis of musical ensembles based on head tracking with a Kinect camera. We discuss first experimental results with a violin duo performance and present ways of analyzing and visualizing the recorded head motion data.

Keywords

Kinect, Ensemble, Synchronization, Strings, Functional Data Analysis, Cross-Correlogram

1. INTRODUCTION

Members of music ensembles have to synchronize to one another with highest precision in order to achieve the desired common musical goal. How musical ensembles achieve such a delicate synchronization is a wide and rich topic for research. Many aspects play a role, such as the musical style, the configuration of the ensemble (piano, string instruments, etc., and perhaps also a conductor or dancers), the experience of the musicians, and many others. Synchronizing requires the musicians to not rely on their auditory perception alone but also to perceive and interpret the movements and gestures of their ensemble colleagues. In order to pursue further research in this direction, we developed a Kinect-based method for motion analysis of musical ensembles. Our method concentrates on head movements which are clearly visible and which the musician may use to communicate with the other ensemble members and the audience. Research in ensemble synchronization could provide new pedagogical insights for ensemble musicians. Furthermore, a better understanding of ensemble synchronization could lead to better computer accompaniment since current solutions [6] are not based on an informed model of (human) ensemble synchronization. Head motion has been already previously shown to play an important communicative role in piano duets [2]. However, those studies have used obtrusive sensor technologies such as inertial sensing or marker-based motion capture.

This paper contributes a method for motion analysis of musical ensembles based on head tracking from depth camera images. This provides an unobtrusive and affordable method to examine synchronization by movement analysis in musical ensembles. Furthermore, we present first experimental results with a violin duo.

2. RELATED WORK

The Kinect has been used in many musical projects such as those described in [9, 7, 11, 12]. Originally, the Kinect was intended for capturing human movement unobtrusively. The standard algorithm [8] that is shipped with the Kinect is based on a decision forest that is trained with an extensive training set. This training set is composed of recordings of actors that were filmed with a depth camera while their movements were simultaneously tracked with a makerbased optical motion capture system. Furthermore, artificial training data was constructed by simulating and rendering human movement. This is possible since the depth information is much less variable than RGB information usually varying between users due to different clothing and and different lighting conditions from recording to recording. The method shipped with the Kinect is not suited for capturing instrumentalist movements, since such conditions (having a violin in the hand, sitting at the piano) were not reflected in the training set. It would be possible to adopt the approach and construct a training in order to apply Shotton's et al. method [8]. However, the large effort to construct such a dataset makes such an approach unpractical for musical applications. Therefore, other solutions have to be found for musical applications, such as for capturing pianist movements [3].

In this paper we provide a method for analysis of head movements in music ensembles. In contrast to [3], which provides unobtrusive pianist motion capture of a large range of joints of a single person, we detect the head movements of multiple ensemble members. Furthermore, our method determines not only the head position but also the viewing direction of the performers. We report first experimental results and data analysis with a violin duet performance.

3. IMAGE ANALYSIS

Setup & Recording: A Kinect depth camera is mounted facing downwards so that it records the music ensemble from above (see Fig. 1). The optimal height of the Kinect is empirically determined with the ensemble in place so that the heads of the ensemble members are always visible during the performance of the piece, taking into account head

NIME'13, May 27–30, 2013, KAIST, Daejeon, Korea.

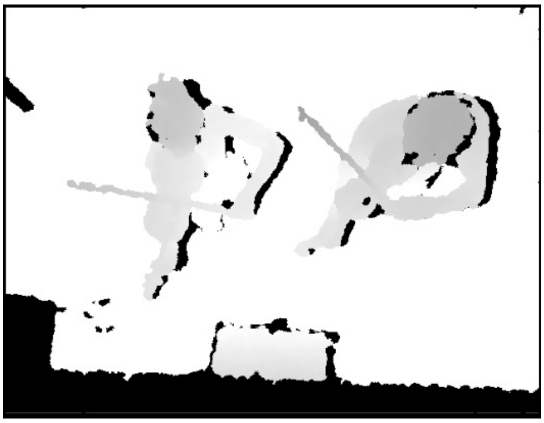

Figure 1: The raw image provided by the Kinect. Darker areas are closer to the camera; lighter areas are farther away. The heads and the bow tips are closest to the camera.

swaying motions that are typical during instrument performance. Our analysis algorithm assumes that the heads of the ensemble members are the highest areas in the depth image (i.e., closest to the camera). Therefore, the recording area has to be free of other high objects. The depth camera images are recorded in a lossless format for later analysis at a frame rate of 30 frames per second.

Algorithm overview: We track the head positions of the ensemble members in order to provide means to analyze gestural ensemble communication and examine movement synchronization of the ensemble members. The head seems to be well suited for expressive performance analyses as shown by previous work [1]. The swaying motion of the head, which is a compound movement by the entire body, is well visible and has usually no specific function in operating the instrument. It is therefore available for communication with the audience and ensemble members. In order to make the most from the depth data, the direction of the head (which is an indicator for the viewing direction) and the position of the head of all ensemble members are tracked. The design of the analysis algorithm takes into account computational efficiency to enable future use in real-time interactive computer music projects. The image analysis consists of 2 steps, which will be discussed in the next sections: head position detection and ellipse matching.

3.1 Head position detection

The Kinect measures depth by projecting an infrared dot pattern into the space. The dot pattern is recorded with an infrared camera. By identifying the dot patterns in the image and evaluating the distance between the dots, the distance from the camera can be determined [10]. The raw Kinect depth image can be seen in Fig. 1. The different shades of grey correspond to different distances from the camera. Darker colors (i.e., lower values) correspond to points that are close to the camera; lighter colors correspond to points that are further away. Due to shadows and reflections, it is not always possible to determine the distance. Areas, in which the distance measurements fail, are marked with zero values, visualized as black areas in the raw data image.

The heads are the highest areas in the image. In order to find the first head, the highest point in the image is identified by iterating through the depth values. It sometimes happens that the bow tip is even higher than the head. In order to filter out such values, the neighborhood of the "candidate head pixel" is examined. The neighborhood is a rectangular area centered around the candidate head pixel (see Fig. 2). If the candidate head pixel is really the highest point on the head then the surrounding pixels will also be head pixels and thus have very similar depth values. On the other hand, if the candidate head pixel is in fact a bow tip pixel then only some of the surrounding pixels will be bow pixels and many other pixels will be floor pixels and have distinctively different depth values. By examining the fraction of pixels in the neighborhood that have similar depth values to the candidate head pixel, these two conditions can be differentiated effectively.

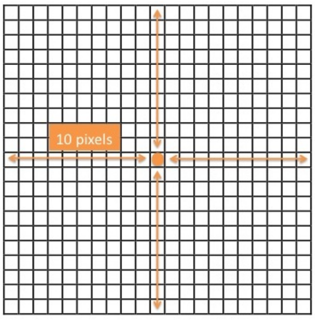

Figure 2: Neighborhood around the candidate head pixel. The rectangle is spanned by 10 pixels in each direction.

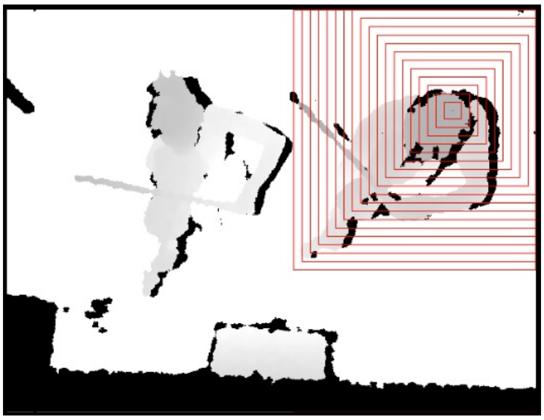

Figure 3: The shaded area centered around the head position of the taller player is excluded in order to detect the head position of the second player.

The head position of the tallest ensemble member is determined through the method described above. In order to detect the head position of the next ensemble member, the same method is repeated. However, a large rectangular area centered around the the previously detected head(s) is excluded from the analysis (see Fig. 3). The overall process is continued until all ensemble members are detected.

3.2 Ellipse matching

In the previous step, the approximate positions of the heads of the ensemble players were detected. In this step the position of the head is refined and the head direction is determined. First, all head pixels of each player are determined. This is done by comparing the depth values of the surrounding pixels in a rectangular area with the depth value of the highest head pixel determined in the previous step. If the

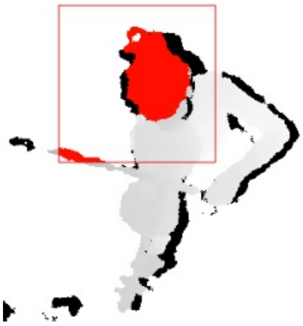

Figure 4: Head pixels are detected in a rectangular area around the highest head point based on the depth difference. Sometimes bow pixels are incorrectly labeled as head pixels.

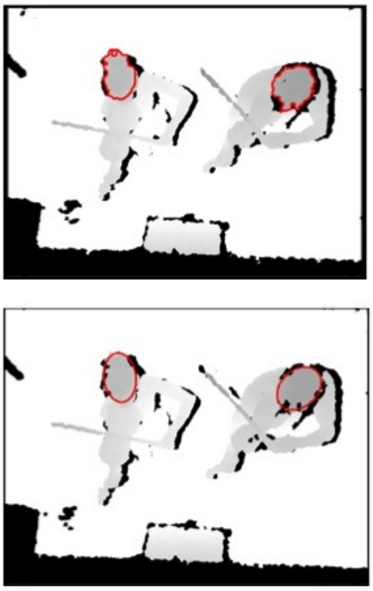

Figure 5: The contours of the head (upper) and the matched ellipses (lower)

depth difference of the pixel amounts to only a few centimeters, it is recognized as a head pixel (see Fig. 4). Sometimes bow pixels are located within that rectangular area and are labeled incorrectly. To avoid this problem a contour detection algorithm is used. This algorithm finds the contours of the regions of connected pixels. The largest contour is then recognized as the head contour. This provides an effective way of differentiation since the contours originating from the bow are rather small. An ellipse is matched onto the contour of the head (see Fig. 5). The center point and the direction of the matched ellipse correspond to the center of the head and the head direction.

4. EXPERIMENTAL RESULTS

We recorded two violinists performing a short piece with a Kinect camera mounted above the musicians. The head position and orientation was extracted with the above algorithm. The resulting head position and head orientation trajectories are plotted in Fig. 6. The forward-backward (y) and the sideways motion (x) of both musicians do not adhere to a strict period as one would expect if there was a one-to-one correspondence to the pulse of the music. Although the trajectories of player B (blue) are not strictly periodical, they show a high regularity, grouping time into small chunks according to the fine-grained musical structure of the piece. Player A's movements (red) on the other hand are freer and less regular. Judging from the motion graphs alone it seems that player B (blue) has the lead in controlling the ensembles tempo, as evident by the busier graphs and more regular motions. We did not detect any systematic variation in the diagram showing the viewing direction.

Acceleration, the second derivative of position, has been shown to contain visual information on timing cues used in ensemble performance, particularly in conducting gestures [4]. Therefore the x and y position data were converted to a functional form using Functional Data Analysis [5] in a further analysis step. Order 6 b-splines were fit to the second derivative (acceleration), with knots placed every 5 data points, and smoothed using a roughness penalty on the fourth derivative ($\lambda = 10^{-5}$), which smoothed the second derivative (acceleration). Head acceleration of x and y was combined by taking the root of the summed squares of x and y acceleration trajectories. The compound head acceleration (indicating head acceleration in any direction) is plotted in Fig. 7 (top panel) for both players.

To elucidate any fine-grained temporal relationships in the two musicians' head movements, we computed multiple cross-correlations between the two compound head acceleration trajectories. The bottom panel of Fig. 7 shows the color-coded coefficients of cross-correlations calculated on windows of 3.33 seconds (or 200 samples at a re-sampling rate of 60 fps) shifted 12.5% sideways, resulting in about 2.5 analyses per second. Red colors reflect regions of high correlation (in-phase movements between the musicians) while blue colors show negative correlations (anti-phase motion). Negative lags (in seconds) mean that A's head movements lead the others' movements, while positive lags point to the opposite (B's movements anticipating A's movements). This "cross-correlogram" reveals longer regions of dark red color: from about 13–24 s player A seems to anticipate the other by about half a second, while the opposite occurs between 36 s and 47 s. This novel way of presenting motion synchronicities over time may represent a powerful analysis tool to unseal otherwise hidden motion relationships between performing musicians.

5. CONCLUSION

The members of a musical ensemble have to synchronize one another with highest precision to achieve the desired musical goal. The musicians do not only rely on acoustic information but also anticipate timing and communicate with each other based on gestures and movements. There has been quite some research on ensemble synchronization (see [2] for a discussion of existing works). However, up to now motion analyses with ensembles have been performed using intrusive technologies, such as inertial sensing or marker-based optical motion capture systems. Particularly the latter are very expensive in both prime costs and data evaluation. In this paper we proposed a head tracking method using a Kinect depth camera which is both very inexpensive in its prime costs and, even more importantly, and unobtrusive in the sense that it does not require markers to be glued on the participants. Furthermore, we have demonstrated the opportunities of our motion tracking method for head motion analysis revealing complex interaction patterns hidden in the complex kinematics of musicians' body motion.

Future work will evaluate this tracking and analysis method

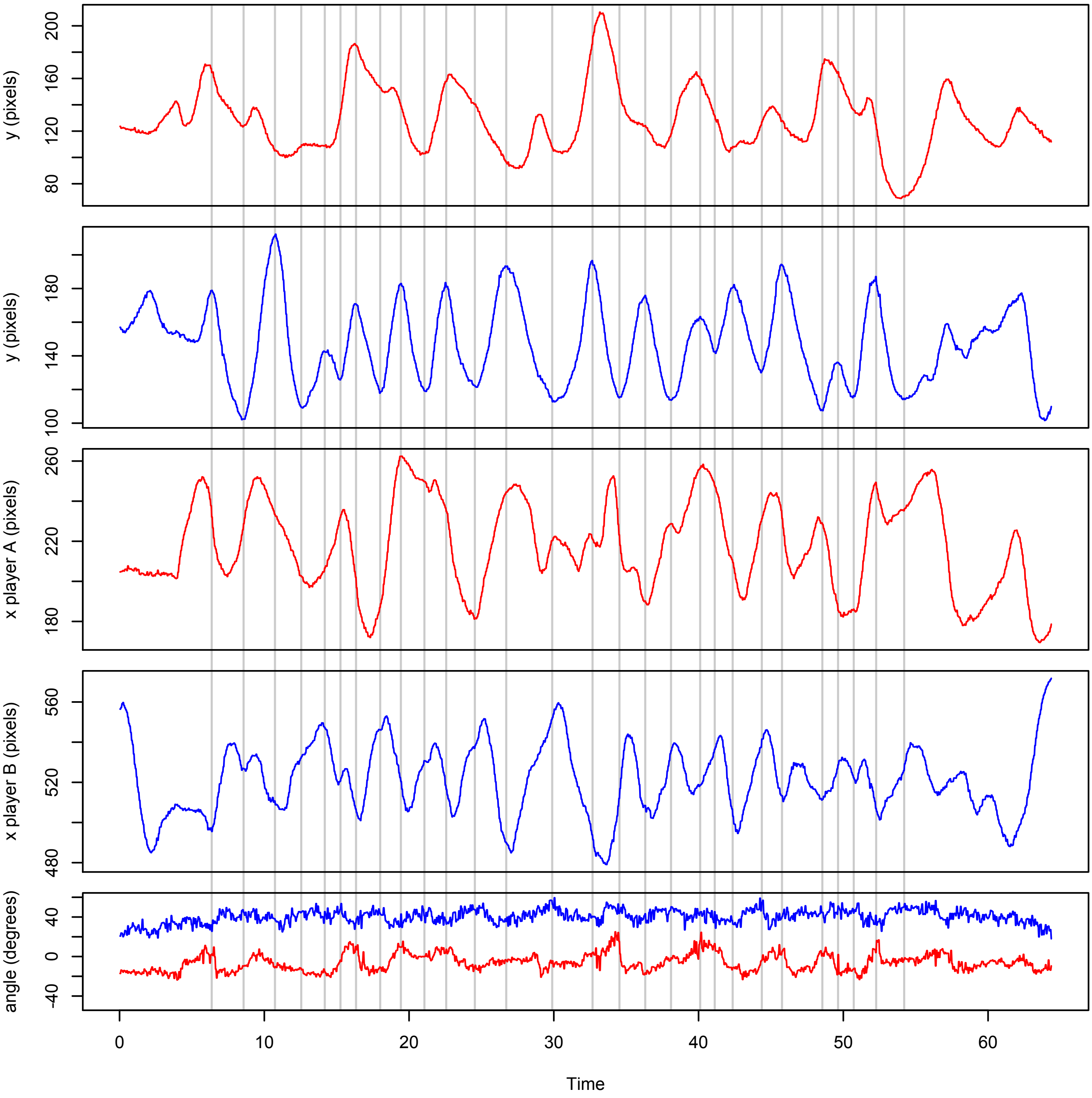

Figure 6: The head position trajectories of player A (red) and player B (blue). The first two diagrams show the forward-backward motion of the musicians (along the image y-axis). The next two diagrams show the sideways motions of the musicians (along the image x-axis). The last diagram shows the musicians head orientation (an indicator for viewing direction). The horizontal gray lines crossing all diagrams are placed at maxima and minima of player B's forward-backward motion (the second diagram).

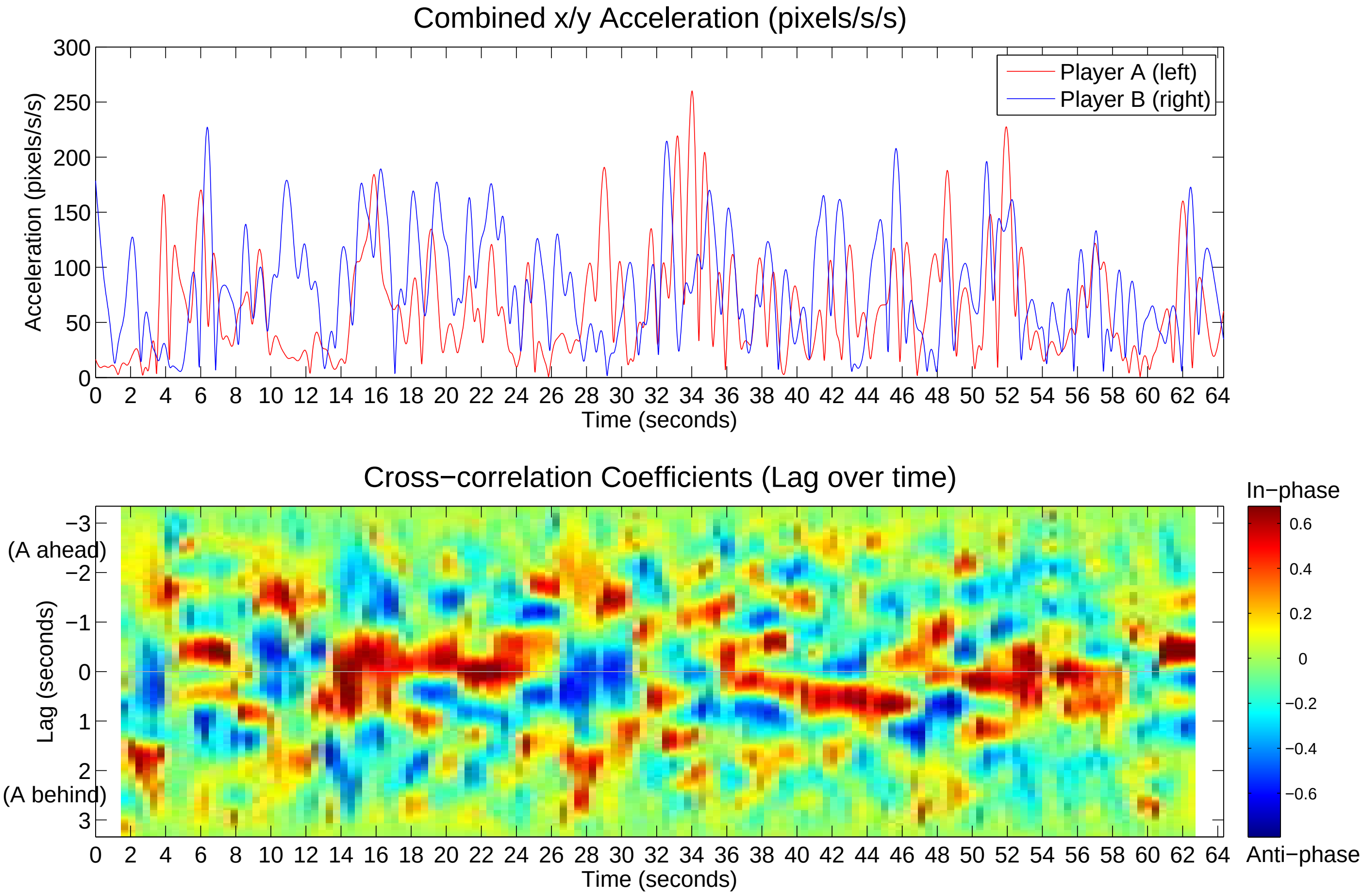

Figure 7: Violin duet performance: Compound head acceleration (in pixels/s^2) against time in seconds (upper panel) and cross-correlation coefficients (color-coded) for lag (in seconds) over time (in seconds). Regions of dark red indicate kinematic in-phase relationships at various lag times.

in controlled real-life experiments. Another path of extension is to enable the algorithm to capture and analyze data from multiple daisy-chained and synchronized Kinect cameras. This would enable us to monitor larger ensembles up to an orchestra and explore the widely unknown kinematic dynamics of music expression and communication evolving during performances of large music ensembles.

6. ACKNOWLEDGMENTS

We thank Anne Weber for performing in the violin duo recording sessions. This work is in part supported by the Austrian Science Fund (FWF project P 24546).

7. REFERENCES

[1] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, pages 103–119, 2008.

[2] W. Goebl and C. Palmer. Synchronization of timing and motion among performing musicians. *Music Perception*, 2009.

[3] A. Hadjakos. Pianist motion capture with the Kinect depth camera. In *SMC 2012*, 2012.

[4] G. Luck and J. A. Sloboda. Spatio-temporal cues for visually mediated synchronization. *Music Percept*, 26(5):465–473, 2009.

[5] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, New York, 2nd edition, 2005.

[6] C. Raphael. Current directions with musical plus one. In *SMC-09*, 2009.

[7] S. Şentürk, S. W. Lee, A. Sastry, A. Daruwalla, and G. Weinberg. Crossole: A gestural interface for composition, improvisation and performance using Kinect. In *NIME-12*, 2012.

[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, volume 2, 2011.

[9] S. Trail, M. Dean, G. Odowichuk, T. F. Tavares, P. Driessen, W. A. Schloss, and G. Tzanetakis. Non-invasive sensing and gesture control for pitched percussion hyper-instruments using the Kinect. In *NIME-12*, 2012.

[10] Wikipedia. Kinect — Wikipedia, the free encyclopedia, 2013. [Online; accessed 25-April-2013].

[11] Q. Yang and G. Essl. Augmented piano performance using a depth camera. In *NIME-12*, 2012.

[12] M.-J. Yoo, J.-W. Beak, and I.-K. Lee. Creating musical expression using Kinect. In *NIME-11*, 2011.