frontiers
in Digital Humanities

Extracting Coarse Body Movements from Video in Music Performance: A Comparison of Automated Computer Vision Techniques with Motion Capture Data

Kelly Jakubowski[1*], Tuomas Eerola[1], Paolo Alborno[2], Gualtiero Volpe[2], Antonio Camurri[2], Martin Clayton[1]

[1]Music, Durham University, United Kingdom, [2]DIBRIS (Department of Informatics, Bioengineering, Robotics, and Systems Engineering), University of Genova, Italy

Submitted to Journal:
Frontiers in Digital Humanities

Specialty Section:
Digital Musicology

ISSN:
2297-2668

Article type:
Original Research Article

Received on:
08 Jan 2017

Accepted on:
21 Mar 2017

Provisional PDF published on:
21 Mar 2017

Frontiers website link:
www.frontiersin.org

Citation:
Jakubowski K, Eerola T, Alborno P, Volpe G, Camurri A and Clayton M(2017) Extracting Coarse Body Movements from Video in Music Performance: A Comparison of Automated Computer Vision Techniques with Motion Capture Data. *Front. Digit. Humanit.* 4:9. doi:10.3389/fdigh.2017.00009

Extracting Coarse Body Movements from Video in Music Performance: A Comparison of Automated Computer Vision Techniques with Motion Capture Data

Kelly Jakubowski[1]*, Tuomas Eerola[1], Paolo Alborno[2], Gualtiero Volpe[2], Antonio Camurri[2], Martin Clayton[1]

[1]Department of Music, Durham University, Durham, UK

[2]Casa Paganini Research Centre, DIBRIS (Department of Informatics, Bioengineering, Robotics, and Systems Engineering), University of Genova, Italy

*** Correspondence:**
Kelly Jakubowski
kelly.jakubowski@durham.ac.uk

Keywords: movement, motion tracking, music performance, musical ensemble coordination, computer vision, video analysis

Abstract

The measurement and tracking of body movement within musical performances can provide valuable sources of data for studying interpersonal interaction and coordination between musicians. The continued development of tools to extract such data from video recordings will offer new opportunities to research musical movement across a diverse range of settings, including field research and other ecological contexts in which the implementation of complex motion capture systems is not feasible or affordable. Such work might also make use of the multitude of video recordings of musical performances that are already available to researchers. The present study made use of such existing data, specifically, three video datasets of ensemble performances from different genres, settings, and instrumentation (a pop piano duo, three jazz duos, and a string quartet). Three different computer vision techniques were applied to these video datasets—frame differencing, optical flow, and kernelized correlation filters (KCF)—with the aim of quantifying and tracking movements of the individual performers. All three computer vision techniques exhibited high correlations with motion capture data collected from the same musical performances, with median correlation (Pearson's r) values of .75 to .94. The techniques that track movement in two dimensions (optical flow and KCF) provided more accurate measures of movement than a technique that provides a single estimate of overall movement change by frame for each performer (frame differencing). Measurements of performer's movements were also more accurate when the computer vision techniques were applied to more narrowly-defined regions of interest (head) than when the same techniques were applied to larger regions (entire upper body, above the chest or waist). Some differences in movement tracking accuracy emerged between the three video datasets, which may have been due to instrument-specific motions that resulted in occlusions of the body part of interest (e.g. a violinist's right hand occluding the head whilst tracking head movement). These results indicate that computer vision techniques can be effective in quantifying body movement from videos of musical performances, while also highlighting constraints that must be dealt with when applying such techniques in ensemble coordination research.

1. **Introduction**

The extraction and quantification of human movement data from musical performances offers a range of potential uses to researchers of musical interaction. Movement data from performers can be instrumental to research on interpersonal synchrony and entrainment between musicians, leader-follower relationships within an ensemble, and musical gestural analysis, to name just a few examples. Extraction of such data from video recordings can be particularly useful in situations where more complex or costly motion capture technologies are not feasible, such as field research and various other ecological performance contexts (e.g., gigs at nightclubs, rehearsals in music practice rooms, ritual ceremonies and religious events, etc.). One area that offers a variety of promising techniques for extracting features of human movement from video is the field of computer vision (Moeslund and Granum, 2001). The work of computer vision scientists is focussed around developing computational methods that perform similar tasks to the human visual system using digital images and videos, including object recognition, event detection, object tracking, and motion estimation (Forsyth and Ponce, 2002).

Researchers have recently begun to test the efficacy of computer vision techniques for capturing and indexing human body movements during social motor coordination tasks (Romero et al., 2016) and dance (Solberg and Jensenius, 2016). The work of Romero et al. (2016) suggests that computer vision methods, as applied to video recordings, can perform similar tracking of body movements to more expensive techniques, such as motion capture (MoCap) systems or Microsoft Kinect, under certain conditions. This is advantageous, as specialised MoCap technologies are not only costly, but can also be invasive in that markers need to be fixed to a person's body (or for some systems a specialised suit needs to be worn), time-consuming in terms of set-up and calibration procedures, and difficult to implement in ecological settings outside of specialised motion capture laboratories. Previous research has revealed that the conditions under which computer vision methods applied to video most closely approximate MoCap tracking in terms of body movement quantification include a fixed video camera angle (e.g., no zooming or panning), stable lighting within the recording setting, no other movements occurring in the background, and the separation of participants in space so as to avoid occlusions or the movements of one participant being included in the analysis space of another (Paxton and Dale, 2012; Romero et al., 2016). However, limitations of the use of computer vision methods for motion tracking include that these methods have previously proved more feasible for tracking large-scale, full-body movements than movements of individual body parts (Paxton and Dale, 2012; Romero et al., 2016) and only measure movements in two dimensions (cf. MoCap and sensors such as accelerometers, which measure movements in three dimensions). Additionally, computer vision techniques are generally applied to data sources with a lower temporal resolution than MoCap technologies; standard video recordings tend to be recorded at a frame rate of around 25 frames per second (fps), whereas MoCap data is often recorded in the range of 100 to 200 fps.

Music performance serves as another highly relevant case for testing the capabilities of computer vision techniques, as group music making employs a variety of movement cues to facilitate the coordination of timing and expressivity between performers. This coordination of timing and expressivity is sometimes referred to as interpersonal entrainment (Clayton et al., 2005). When producing video recordings of musical performances it is also often possible to implement solutions to minimise some of the challenges to the application of computer vision techniques listed above. For instance, the lighting and camera angle may be able to be fixed to a standardised setting throughout a performance and the performers may be situated within the performance space such that they do not occlude one another (at least in small ensembles).

Coordination in musical ensembles is achieved through the use and integration of both auditory (instrumental and vocal sounds) and visual (body movement and eye contact) cues. The accuracy of temporal coordination in the auditory domain is typically in the order of tens of milliseconds in expert ensemble performance (e.g., Keller, 2014; Rasch, 1988; Shaffer, 1984). The movements that produce these sounds, such as finger movements of a pianist or bowing movements of a violinist, often evolve at similarly short timescales. In addition to these instrumental, sound-producing movements that are required in performance, musicians also make use of a variety of communicative and sound-facilitating movements that can serve to coordinate timing and expressive intentions between performers (Jensenius et al., 2010). These ancillary movements (e.g., head nods, body sway) typically evolve over longer timescales than instrumental movements (e.g. in the order of seconds; Davidson, 2009; Wanderley et al., 2005). Importantly, systematic relationships have been observed between coordination at the level of ancillary body movements and musical sounds (Keller and Appel, 2010; Ragert et al., 2013). Thus, the analysis of such movements can provide information about the overall level of interpersonal coordination within an ensemble performance. In contrast to acoustic features and instrumental movements, ancillary body movements tend to generalise across performers regardless of the instrument played and are also prevalent in vocal performance. Additionally, the fact that ancillary movements tend to take place across longer timescales than instrumental movements allows them to be tracked within video recordings despite its lower temporal resolution in comparison to MoCap. Therefore, it is of great interest to music researchers to measure and analyse ancillary movements from video recordings of musical performances.

There are a variety of areas within the field of music performance research that may benefit from the use of computer vision techniques to measure movement data with a view to quantifying interpersonal coordination. For instance, such techniques could be applied to study temporal relationships between performers within commercial video recordings of classical or popular music, or to quantify corporeal interactions between a music or dance therapist and his/her clients. Ethnomusicologists often make video recordings of musical performances in ecological settings in which access to sophisticated technologies such as motion capture is not feasible. Indeed, a large amount of archival material of video recordings of music performances from across the world already exists. For example, the JVC Video Anthology of World Music and Dance (JVC, Victor Company of Japan, 1990) comprises some 30 volumes of field recordings from across the world and the Ethnographic Video for Instruction & Analysis (EVIA) Digital Archive Project (http://www.eviada.org/default.cfm) is a repository of ethnographic videos, including many music performances, which aims to preserve these materials for the long-term in a digital, online format. As such, if video-based analysis methods prove to be fruitful in providing new insights about musical interaction, a large amount of useful research could be done that makes use of such existing video archives (with the appropriate permissions and taking account of ethical considerations), which could thereby minimise the costs that are necessarily incurred when collecting new data. The present study served as a test case in this regard, as it also made use of existing data—in this case, three existing datasets in which both video and motion capture recordings had been collected (as reported in Glowinski et al., 2013, Moran et al., 2015, and one previously unpublished dataset). Our study was therefore able to test whether computer vision techniques could be used to quantify body movements from video recordings that had originally been obtained for other research purposes.

The computer vision field offers a diverse range of possible techniques for tracking moving elements and changes in image sequences that were considered for use in the present study. As the majority of materials in our datasets of musical performances presented a situation in which only the to-be-tracked targets (the performers) were moving, we first considered background subtraction techniques. These techniques aim to distinguish an object(s) (in this case, the performers) in the foreground from a static background and perform further processing (e.g., tracking or motion detection) on the foreground object. The background subtraction-based technique that we applied was frame differencing. Frame differencing is one of the oldest and most widely-used computer vision techniques, which measures the overall change in pixels within the foreground from one frame to the next (Wren et al., 1997; see also Jensenius et al., 2005, for an implementation for studying musical gestures). We then explored two techniques that provide more detailed information on the direction of motion of each performer. Specifically, we employed a technique based on the variation of the motion field, known as optical flow (Farnebäck, 2003), and a technique based on pattern similarity calculation, known as kernelized correlation filters (hereafter referred to as KCF; Henriques et al., 2015). Optical flow is a technique that has been widely applied within the computer vision literature (e.g. Fleet and Weiss, 2006; see also Latif et al., 2014, for an application in studying interpersonal coordination), whereas KCF is a comparatively recently developed technique. Both of these techniques were used to track the direction of movement of the performers by providing both horizontal and vertical position data of each performer within each frame.

To summarise, in the present project we applied three automated computer vision techniques (frame differencing, optical flow, and KCF) to a set of video recordings of musical performances comprising a variety of performers, performance settings, instrumentations, and musical styles. The aims were 1) to test the robustness of the computer vision techniques for capturing body movements across the different performance conditions and 2) to test how closely these techniques were able to capture the actual motion of performers, as indexed by motion capture data from the same performances. Finally, as previous studies comparing motion capture data to computer vision techniques have primarily examined full-body movements (e.g. Romero et al., 2016), we extended this area of research to include analysis of video data within predefined regions of interest (i.e., head, upper body) to test whether the video analysis techniques could also be effective in quantifying movements of specific parts of the body. If it was found that computer vision techniques could be effectively applied to measure movement in specific body parts such as the head, this would suggest that in some cases it may be possible to differentiate sound-producing, instrumental movements from sound-facilitating, ancillary movements of musical performers by isolating a part of the body that does not play a role in both types of movement (e.g., a guitar or cello player does not typically use head movements to produce sounds but rather for communicative purposes).

2. **Methods**

2.1 **Materials**

The project made use of three existing datasets (see Figure 1), in which both video recordings and MoCap data of the same musical performances had been collected for other research purposes.[1] The first dataset (previously unpublished and hereafter referred to as the "Piano Duo") comprised seven songs performed by singer-songwriters Konstantin Wecker and Jo Barnikel. Wecker has been described as one of Germany's most successful singer-songwriters, with a career spanning 40 years at the time of the recording, and Barnikel is a leading film and TV composer who had been accompanying Becker on recordings and concert tours for over 15 years.

The second dataset consisted of three performances by jazz duos, a subset of the Improvising Duos corpus described in Moran et al., 2015. In this subset (hereafter referred to as "Mixed Instrument Duos"), two duos performed free jazz improvisations and one performed a jazz standard (*Autumn Leaves* [J. Kosma, 1945]). Performers in these duos were recruited on the basis of public performance experience of around 10 years in their respective styles. Data from five of the six performers from this dataset were analysed in respect of performers' permissions on data reuse.

The third dataset ("String Quartet") comprised eight recordings by the Quartetto di Cremona string quartet performing the first movement of Schubert's *String Quartet No. 14* ("Death and the Maiden"; Glowinski et al., 2013). Two of these recordings featured only the first violinist performing his part alone. For the other six recordings, two of the four performers were selected for whom the least occlusions were observed (i.e. another player was not moving in front of him/her regularly). In total, the three datasets allowed for the analysis of 33 cases of 10 different performers playing six different instruments (see Table 1).

-INSERT FIGURE 1 ABOUT HERE-

[1] In all instances the primary focus of the original research was on the collection of MoCap data, thus the performance settings were optimised for MoCap data collection and video was collected as a secondary measure for reference purposes only.

For each of the three datasets, the recordings were made in the same room under similar performance conditions (e.g. all string quartet recordings were made with performers situated in a similar position on the same stage using the same video camera and MoCap system). The Piano Duo and Mixed Instrument Duos were both recorded at the Max Planck Institute in Leipzig, Germany, using a Vicon Nexus 1.6.1 optical motion capture system with ten cameras and a sampling rate of 200 Hz. A SONY HDR-HC9 camera was used to make the video recordings. The video files were recorded in AVI format at a frame rate of 25 fps and frame size of 720 x 576 pixels. The String Quartet was recorded at Casa Paganini Research Centre (University of Genova, Italy), using a Qualisys Oqus300 motion capture system with eleven cameras and a sampling rate of 100 Hz. A JVC GY-HD-251 camera was used to capture video of the performances. The video files were recorded in AVI format at a frame rate of 25 fps and frame size of 720 x 576 pixels.

-INSERT TABLE 1 ABOUT HERE-

2.2 **Analysis**

2.2.1 **Motion capture data**

All MoCap data were processed using the MoCap Toolbox (Burger and Toiviainen, 2013) in Matlab. Each dataset was first rotated in order to orient the MoCap data to the same perspective as the camera angle of the video recording. This was done manually by inspecting animations generated from the MoCap data in comparison to the video recording (see Figure 2). Once the optimal rotation was achieved, a subset of markers was selected from each performer, comprising one marker from the head and one from the torso or each shoulder (if a torso marker was not present, as was the case for the String Quartet). If multiple markers were present for a specific body part (e.g. four head markers), the marker for which the least amount of data points were missing was selected. Markers were also selected in consideration of the camera angle of the video. For instance, if only the back of the head of a performer was visible in the video, a marker from the back of the head was selected. The three-dimensional coordinates from each selected marker were saved for further analysis. The horizontal and vertical coordinates of the MoCap data are subsequently referred to as the x- and y-dimensions respectively, which were compared to the two-dimensional data that were derived from the video recordings by the computer vision techniques.

2.2.2 **Video data**

The computer vision techniques (frame differencing, optical flow, and KCF) were implemented in EyesWeb XMI 5.6.2.0 (http://www.infomus.org/eyesweb_ita.php). The first step when applying each technique was to manually define relevant regions of interest (ROIs) on which to apply the technique to each video. A rectangular ROI was selected around each performer whilst ensuring that only that individual performer was serving as the main source of motion in the ROI (see Figure 2). This was generally achieved to a high standard, although there were a few cases in which the hands or bows of another performer occasionally moved into the ROI in the Piano Duo and String Quartet. Two sets of ROIs were defined for each performer in each video—a larger ROI that comprised the upper body (from the mid-chest or the waist up to the top of the head, depending on how much of the performer could be seen in the video[2]) and a smaller ROI around the head only. Frame differencing and optical flow were both applied using the same sets of upper body and head ROIs for each video. A slightly different set of upper body and head ROIs were defined for KCF, due to the way this technique is implemented. In typical implementations of KCF, the entire ROI moves dynamically throughout the process of tracking the performer. Conversely, frame differencing and optical flow were applied on static ROIs that do not move during the analysis process. As such, larger ROIs were needed that could encompass the whole range of movement of a performer for frame differencing and optical flow, whereas KCF is more suited to smaller ROIs since the ROI shifts from frame to frame.

In frame differencing, the foreground, i.e. the moving element(s) of interest (in this case, the performers), is separated from the background and further processing is performed on the foreground. In the present study, frame differencing was implemented using the Pfinder algorithm of Wren et al. (1997). A version of this algorithm has previously been implemented in EyesWeb for studying interpersonal musical coordination in Indian duos (Alborno et al., 2015). The Pfinder algorithm uses adaptive background subtraction, in which the background model that is subtracted from the foreground is constantly updated throughout the analysis process. The speed at which the background model is updated is determined by the alpha constant, which was set in the present study to 0.4, following an optimisation process in which this parameter was manually adjusted to a range of values and tested on a subset of the present videos. The analysis that was performed on the foreground elements measures the overall Quantity of Motion (QoM) in each ROI for each frame, which is computed based on the number of pixels that change in the foreground from one frame to the next. This analysis produces one column of output values for each performer.

Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. In optical flow, characteristics such as edges or angles are identified within each section of the video frame. In the next frame, such characteristics are sought again. A speed is then associated to each pixel in the frame; the movement is determined by the ratio between the distance in pixels of the displacement of the characteristic in question and the time between one frame and another. The version of optical flow that was implemented in the present study is known as dense optical flow[3] and is based on the algorithm of Farnebäck

---

[2] In some cases the waist of a performer could not be seen, as it was behind their instrument (e.g. for some pianists).

[3] Traditional optical flow methods (e.g. as implemented by Lucas and Kanade (1981)) compute optical flow for a sparse feature set, i.e. using only specific parts of the image, such as detected corners. Dense optical flow, as implemented by Farnebäck (2003), performs optical flow computation on all pixels in the image for each frame. The use of dense optical flow can increase the accuracy of the optical flow results, with a tradeoff of slower computation speed.

(2003). This technique has previously been implemented in EyesWeb in work of Alborno et al. (2015) on Indian music duos, as well as to develop a "virtual binocular" installation in which users' movements are tracked and estimated by computation of optical flow on the face (Camurri et al., 2010). A similar optimisation procedure was followed to that used for frame differencing in which the "pyramid layers" parameter was adjusted to a range of values and tested on a subset of the present videos. This parameter allows for the tracking of points at multiple levels of resolution; increasing the number of pyramid layers allows for the measurement of larger displacements of points between frames but also increases the number of necessary computations. The optimal value that was selected for this parameter was 12. The resulting output that was provided by the optical flow analysis was two columns of data per performer, which represent movement of the barycentre of the ROI along the x- (horizontal) and y- (vertical) axes. The barycentre of the ROI is computed based on pixel intensities. The video image is converted to greyscale and the barycentre coordinates are calculated as a weighted mean of the pixel intensities within the ROI; this is done separately for the x- and y-dimensions.

KCF is a relatively recently developed tracking technique (Bolme et al., 2009), based on older correlation filter methods (Hester, 1980), that works using pattern similarity calculations on a frame-by-frame basis. KCF was implemented in EyesWeb[4] in the present study using the OpenCV C++ implementation[5] of the algorithm of Henriques et al. (2015). When the KCF algorithm is initialised, a visual tracker is placed at the centre pixel of the pre-defined ROI for the first frame of the video. In the second frame, similarity and classification computations are performed by searching for the set of pixels with the maximum correlation to the initial tracker position in terms of its multi-channel RGB colour attributes, and so on for each subsequent frame. In effect, this allows the technique to track the movement of the performers across the ROI. Similarly to optical flow, the output of the KCF analysis is two columns of data per performer, which represent movement of the barycentre of the ROI along the x- and y-axes. In this case, since the ROI moves dynamically with the performer, the barycentre that is used is the geometric barycentre at the intersection of the two diagonals of the rectangular ROI.

2.2.3 **Motion capture and video comparison**

As video data collection was not the primary focus of the original studies, the video and MoCap data were not synchronised with an external timecode. As such, these two data sources were aligned in the present study using automated cross-correlational methods. Each video analysis output from EyesWeb was cross-correlated with its corresponding MoCap target (e.g. the x-coordinate of the head from the optical flow analysis within the head ROI was cross-correlated with the x-coordinate of the MoCap head marker). This allowed us to determine the optimal lag time for each trial, which was defined as the lag at which the maximum correlation value between the video and MoCap data was reached. The median optimal lag time from all cross-correlational analyses from the same video (taking account of

[4] The KCF block has recently been released within the Image Processing Library of EyesWeb.
[5] http://docs.opencv.org/trunk/d2/dff/classcv_1_1TrackerKCF.html

analysis of all position data from both performers in each video) was taken as the optimal lag time for that particular video. The median optimal lag time across all video and MoCap pairings in the dataset was 0.05 seconds (range = -0.10 to 0.42 seconds). Before computing any statistical comparisons between the video and MoCap data, the MoCap data were down-sampled to match the lower sampling rate of the videos at 25 fps, and all video and MoCap data outputs were de-trended and normalised. Figure 2 depicts the data preparation and extraction process for video and MoCap for one example performance from the Mixed Instrument Duos.

-INSERT FIGURE 2 ABOUT HERE-

3. **Results**

The main focus of the subsequent data analysis was to compare the efficacy of the three computer vision techniques (frame differencing, optical flow, and KCF) for measuring body movements of musical performers across the three different datasets (Piano Duo, Mixed Instrument Duos, and String Quartet)[6] and two sets of ROIs (upper body and head). For the upper body ROI, we compared the outputs of the computer vision analyses to the coordinates of the torso marker from the MoCap data (or the right shoulder marker, in the case of the String Quartet[7]) for each trial. For the head ROI, we compared the computer vision data to the coordinates of the MoCap head marker.

Since frame differencing provides a single, overall estimate of movement of each performer (rather than two-dimensional tracking), the optical flow, KCF, and corresponding MoCap data were converted from Cartesian (x and y) to polar (radial and angular) coordinates. We then computed the absolute change of the radial coordinate on a frame-by-frame basis for each trial; this absolute change measure was used in subsequent comparisons to the one-dimensional frame differencing results. Both the resultant absolute change data and the QoM data from frame differencing were kernel smoothed in R using the Nadaraya–Watson kernel regression estimate with a bandwidth of 1.[8] The video and MoCap data for each trial were then compared using correlations (Pearson's *r*); a summary of these comparisons is reported, by dataset, in Table 2.[9] These descriptive statistics suggest that the two-dimensional tracking methods (optical flow and KCF) tend to perform more accurately than the more coarse-

[6] Although the primary research question is focused on evaluating and comparing the three computer vision techniques within the two ROIs, "dataset" is also included as an independent variable in subsequent analyses to take account of the fact that the three datasets vary on a number of parameters, including setting, recording session, lighting, camera angle, and instrumentation.

[7] This analysis was also tested with the left shoulder marker and the average of the left and right shoulder markers, however these analyses revealed similar patterns of results and did not increase the overall correlations.

[8] This smoothing procedure was applied because both the video and MoCap data contained small random fluctuations, which were smoothed without tampering with the overall shape of the trajectories. Filtering had a minor positive effect on the overall results (mean increase in video/MoCap correlation values of 0.07).

[9] Median values (rather than means) are reported as descriptive statistics throughout this paper due to some non-normal data distributions and the relative robustness of the median to the presence of statistical outliers.

grained method (frame differencing) and that performance of all three computer vision techniques is improved when concentrated on a smaller ROI (head, as compared to upper body).

-INSERT TABLE 2 ABOUT HERE-

For the data using the upper body ROI, a 3x3 mixed ANOVA was conducted to test the effects of computer vision technique (frame differencing, optical flow, KCF) and dataset (Piano Duo, Mixed Instrument Duos, String Quartet) on accuracy of overall movement measurement (as indexed by the correlation of each video analysis output with the MoCap data; see Table 2). Prior to entering the correlation values as the dependent variable in the ANOVA, these values were subjected to a Fisher z-transformation to normalise the distribution. The ANOVA revealed significant main effects of computer vision technique ($F(2, 60) = 16.51$, $p < .001$, $\eta_p^2 = .355$) and dataset ($F(2, 30) = 15.41$, $p < .001$, $\eta_p^2 = .507$), as well as a significant technique by dataset interaction ($F(4, 60) = 18.82$, $p < .001$, $\eta_p^2 = .557$). Bonferroni-corrected, paired-samples t-tests indicated that optical flow provided a more accurate measure of performers' movements than both frame differencing ($t(32) = 3.67$, $p = .003$) and KCF ($t(32) = 3.38$, $p = .006$); no significant difference was found between the frame differencing and KCF techniques. Tukey HSD tests revealed that overall movement measurements were more accurate for the Piano Duo than both the Mixed Instrument Duos (mean difference = 0.528, SE = 0.152, $p = .004$) and the String Quartet (mean difference = 0.583, SE = 0.110, $p < .001$); no significant difference was found between the Mixed Instrument Duos and the String Quartet. Bonferroni-corrected, independent-samples t-tests indicated that the optical flow technique exhibited more accurate performance for the Piano Duo than the Mixed Instrument Duos ($t(17) = 4.06$, $p = .009$) and the String Quartet ($t(26) = 6.80$, $p < .001$). The KCF technique also achieved more accurate performance for the Piano Duo than the String Quartet ($t(26) = 3.39$, $p = .018$). All other pairwise comparisons of the three datasets by computer vision technique failed to reach statistical significance.

An analogous 3x3 mixed ANOVA was conducted for the data using the head ROIs. A significant effect of computer vision technique was found ($F(2, 60) = 24.23$, $p < .001$, $\eta_p^2 = .447$), with no significant effect of dataset ($F(2, 30) = 3.14$, $p = .058$, $\eta_p^2 = .173$). The technique by dataset interaction term was statistically significant ($F(4, 60) = 5.59$, $p = .001$, $\eta_p^2 = .272$). Bonferroni-corrected, paired-samples t-tests revealed that optical flow and KCF both provided more accurate measures of performers' movements than frame differencing ($t(32) = 3.88$, $p = .001$ and $t(32) = 8.38$, $p < .001$, respectively) and KCF provided a more accurate measure than optical flow ($t(32) = 2.77$, $p = .027$). Bonferroni-corrected, independent-samples t-tests indicated that the optical flow technique achieved more accurate performance for the Piano Duo than the String Quartet ($t(26) = 4.42$, $p = .001$). All other pairwise comparisons of the three datasets by computer vision technique failed to reach statistical significance.

Finally, we compared performance of the computer vision techniques between the upper body ROI versus the head ROI. A paired-samples t-test indicated that movement measurement was more accurate overall when restricted to a smaller ROI (the head) than a larger ROI (upper body), $t(98) = 2.54$, $p = .013$.

We next looked in more detail at tracking in the horizontal versus vertical dimensions for both optical flow and KCF, as compared to the MoCap data. These results are displayed in Table 3, broken down by tracking dimension. Paired-samples t-tests for both the optical flow (upper body ROI: $t(32) = 6.22$, $p < .001$; head ROI: $t(32) = 5.21$, $p < .001$) and KCF data (upper body ROI: $t(32) = 6.82$, $p < .001$; head ROI: $t(32) = 5.77$, $p < .001$) indicated that tracking by the computer vision techniques was significantly more accurate in the horizontal than the vertical dimension. To probe this difference further, we explored whether the overall lower performance in vertical movement tracking might be due to the computer vision techniques also picking up on the missing, third dimension (depth) in which movement can be made, in addition to the vertical dimension. It is plausible that this might especially be the case when a performer is orthogonal to the video camera, and thus movement forward and backward appears in the video as increases or decreases in the size of the performer. We conducted two sets of regression analyses in which 1) the vertical dimension of the MoCap data was used as a predictor of the vertical dimension of the video data and 2) the vertical dimension of the MoCap data *and* the depth dimension of the MoCap data were used as predictors of the vertical dimension of the video data. We then computed the change in adjusted R^2 values between the two regression analyses. For optical flow analysis, the adjusted R^2 values for the Mixed Instrument Duos and String Quartet only increased on average by 0.03 and 0.06 respectively when taking the third MoCap dimension into account. In both of these datasets the performers were viewed from the side or were situated diagonally with respect to the camera (see Figure 1). However, in the Piano Duo, where the performers were seated orthogonally to the camera (see Figure 1), the R^2 values of the regression models increased on average by 0.22 when the depth dimension of the MoCap data was added as a predictor in addition to the vertical dimension. Although all of the increases in adjusted R^2 values were statistically significant (Mixed Instrument Duos: $t(9) = 2.59$, $p = .029$; String Quartet: $t(27) = 3.92$, $p = .001$; Piano Duo: $t(27) = 3.99$, $p < .001$), the raw adjusted R^2 values indicate that the inclusion of the depth dimension made the most substantial contribution to explaining the previously unaccounted variance in the Piano Duo. A similar pattern emerged for the KCF data (adjusted R^2 change values: Piano Duo = 0.13, Mixed Instrument Duos = 0.03, String Quartet = 0.06). This change was statistically significant within the Piano Duo ($t(27) = 3.95$, $p = .001$) and String Quartet datasets ($t(27) = 3.44$, $p = .002$) but not the Mixed Instrument Duos ($t(9) = 2.12$, $p = .063$).

-INSERT TABLE 3 ABOUT HERE-

4. **Discussion**

The results of the present study indicate that the quantification of movement of musical performers from video using computer vision techniques closely approximates measurements from more sophisticated and costly technologies such as motion capture systems under certain conditions. Specifically, frame differencing, optical flow, and KCF techniques all achieved generally high correlations with MoCap data collected from the same musical performances, with median correlation values of .75 to .94, depending on the ROI, dataset, and computer vision technique. These results are in line with the work of Romero et al. (2016), who found specifically that frame differencing methods could provide a close approximation to MoCap data when tracking movement during social coordination tasks involving tapping, pointing, and clapping. It should also be noted that the promising results of the present study were obtained despite the fact that the video datasets were originally collected as a secondary measure to MoCap and the performance settings were not optimised with video data collection or computer vision analysis in mind. This suggests that the performance of these computer vision techniques might improve even further when working with video data that is optimised for the present research purposes, but also that existing video corpora that have been compiled for other aims could still provide promising data sources for subsequent research in which quantification of movement from video is required.

Our results also extend previous research (e.g., Paxton and Dale, 2012; Romero et al., 2016) by suggesting that the more recently developed, two-dimensional tracking techniques (optical flow and KCF) tend to outperform the older method of frame differencing. In addition, tracking of the head within the head ROI was more accurate overall than tracking of the torso within the upper body ROI. The KCF technique in particular displayed marked performance improvements in comparison to the other two techniques when constrained to the head ROI as compared to the upper body. A plausible explanation for the improved performance within the head ROIs is that the larger ROIs set around the upper body contain a variety of sources of movement, including not just torso movement but head movement and, in some cases, hands, bows of stringed instruments, etc., thereby resulting in decreased tracking accuracy of the torso. Researchers aiming to make use of larger ROIs (such as the upper body ROI from our study) to address particular research questions in the future might note that we were still able to provide a reasonable approximation of overall movement of musical performers as compared to MoCap data. However, it can be difficult to differentiate between various sources of movement within a large ROI, for example, sound-producing/instrument-specific movements (e.g., movement of the violin bow or shifting of the left hand up and down the neck of a cello) versus sound-facilitating/ancillary gestures (e.g., head nods or swaying together in time). Thus, ROI size should be taken into account in future research when the objective is to track movement from specific body parts or to measure only specific types of movement. On the other hand, if the objective is to provide an overall estimate of a performer’s movement and there is no need to clarify the body part from which the movement originates or its expressive/functional purpose a larger ROI could still be suitable.

Within the present study the two-dimensional computer vision techniques exhibited greater precision in tracking horizontal than vertical movement. This seems to be at least partially explained by the missing dimension (depth) that cannot be precisely tracked by video analysis methods in the same way as afforded by MoCap. The implication of this finding is that studies which aim to track precise directionality of vertical movement such as head nods might encounter a certain degree of measurement error, whereas horizontal movements such as side-to-side swaying can be tracked with a greater degree of spatial precision. However, combining these two tracking dimensions into polar coordinates (as in Table 2) tends to provide a good approximation of the overall movement of a performer, with median correlations above .80 for the upper body ROI and above .90 for the head ROI in both optical flow and KCF. Another possible avenue for future research would be to record video of musical performances using multiple camera angles in an attempt to recover the missing third dimension that cannot be measured from the present video data.

Some differences between the three datasets emerged, particularly in regard to the upper body ROI. In general, measurements of performers' movements were more accurate for the Piano Duo than the String Quartet and, in some cases, the Mixed Instrument Duos. This may be due, at least in part, to the fact that within the String Quartet dataset and certain examples from the Mixed Instrument Duos (cellist and double bassist), the bows of the violinist/violist and the left hands of the cellist/double bassist often entered the ROIs and created an extra source of motion that could be picked up by the computer vision techniques. This was the case even when the ROI was focused around the head, as the bow or left hand sometimes occluded the face. These cases provide examples of a discrepancy in differentiating the sound-producing, instrumental movements of a performer from ancillary movements of the head, and highlight that the specific demands and idiosyncrasies of performing on certain instruments should be taken into account when conducting research that aims to quantify musicians' movements from video. In the case of the string quartet, a different camera angle could be considered to avoid occlusions within the ROI. Or, depending on the research question of interest, other body parts could be tracked that do not present this occlusion problem, for instance, the tapping of performers' feet in time to the music.

It should also be noted that some of the differences in movement tracking/quantification accuracy between the three datasets could have arisen from differences in the video source material, such as lighting, camera angle, and distance of the performers from the camera. Future research should aim to test the independent contributions of each of these factors. Additionally, there may have been fundamental differences between the *types* of ancillary movements that performers in the different datasets made, which could be affected both by the instrument being played and the musical style itself (e.g. free jazz improvisation and notated string quartets might require different types of communicative gestures for different purposes). Although classifying movement types is beyond the scope of the present study, future research could also test whether certain classes of body movements are more accurately tracked than others.

These results open new avenues for researchers of musical movement. In our own future research we aim to apply some of these computer vision techniques to examine how the relationships between the movements of co-performers stabilise or change over time and how these corporeal relationships affect audience appraisals of a performance. We also aim to conduct cross-cultural comparisons of what it means to "play in time together" within different musical traditions, using music that is performed for a variety of different functions (e.g., rituals, dance, concert performance, etc.; Clayton, 2013). Additional possible applications of these computer vision techniques for future research could include the study of leader-follower relationships, the relationship between visual movement coordination and synchrony/asynchrony in the auditory modality, and studies of movement coordination differences between expert versus novice performers.

References

Alborno, P., Volpe, G., Camurri, A., Clayton, M., and Keller, P. (2015). Automated video analysis of interpersonal entrainment in Indian music performance. In *7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)* (pp. 57-63). IEEE. doi: 10.4108/icst.intetain.2015.259521

Bolme, D. S., Draper, B. A., and Beveridge, J. R. (2009). Average of synthetic exact filters. Proceedings from *Computer Vision and Pattern Recognition* (pp. 2105-2112). IEEE. doi: 10.1109/CVPR.2009.5206701

Burger, B. and Toiviainen, P. (2013). MoCap Toolbox-A Matlab toolbox for computational analysis of movement data. In *Proceedings of the Sound and Music Computing Conference 2013*. Berlin: Logos Verlag Berlin. ISBN 978-3-8325-3472-1

Camurri, A., Canepa, C., Coletta, P., Cavallero, F., Ghisio, S., Glowinski, D., and Volpe, G. (2010). Active experience of audiovisual cultural content: The virtual binocular interface. In *Proceedings of the Second Workshop on eHeritage and Digital Art Preservation* (pp. 37-42). Firenze, Italy. doi: 10.1145/1877922.1877934

Clayton, M (2013). Entrainment, ethnography and musical interaction. In M. Clayton, B. Dueck, & L. Leante (Eds.), *Experience and Meaning in Music Performance* (pp. 17-39). Oxford: Oxford University Press.

Clayton, M., Sager, R., and Will, U. (2005). In time with the music: The concept of entrainment and its significance for ethnomusicology. In *European Meetings in Ethnomusicology* 11: 1-82.

Davidson, J. W. (2009). Movement and collaboration in musical performance. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford Handbook of Music Psychology* (pp. 364-376). Oxford: Oxford University Press.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In J. Bigun and T. Gustavsson (Eds.), Proceedings from *13th Scandinavian Conference on Image Analysis* (pp. 363-370). Heidelberg: Springer Berlin.

Fleet, D. and Weiss, Y. (2006). Optical flow estimation. In N. Paragios, Y. Chen, & D. Olivier (Eds.), *Handbook of mathematical models in computer vision* (pp. 237-257). Springer US.

Forsyth, D. A. and Ponce, J. (2002). *Computer vision: A modern approach*. Englewood Cliffs, NJ: Prentice Hall Professional Technical Reference.

Glowinski, D., Gnecco, G., Piano, S. and Camurri. A. (2013). Expressive non-verbal interaction in string quartet. In Proceedings of *Conference on Affective Computing and Intelligent Interaction (ACII 2013).* Geneva, Switzerland.

Henriques, J. F., Caseiro, R., Martins, P., and Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37: 583-596. doi: 10.1109/TPAMI.2014.2345390

Hester, C. F. and Casasent, D. (1980). Multivariant technique for multiclass pattern recognition. *Applied Optics* 19: 1758-1761.

Jensenius, A. R., Godøy, R. I., and Wanderley, M. M. (2005). Developing tools for studying musical gestures within the Max/MSP/Jitter environment. In *Proceedings of the International Computer Music Conference* (pp. 282-285). ISSN 2223-3881

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., and Leman, M. (2010). Concepts and methods in research on music-related gestures. In R.I. Godøy & M. Leman (Eds.), *Musical Gestures: Sound, Movement, and Meaning* (pp. 12–35). New York: Routledge.

JVC Video Anthology of World Music and Dance (1990). Tokyo: JVC, Victor Company of Japan.

Keller, P.E. (2014). Ensemble performance: Interpersonal alignment of musical expression. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures* (pp. 260-282). Oxford: Oxford University Press.

Keller, P.E. and Appel, M. (2010). Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. *Music Perception* 28: 27-46. doi: 10.1525/mp.2010.28.1.27

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S. and Munhall, K. G. (2014). Movement coordination during conversation. *PloS One* 9: e105036.

Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding* 81: 231-268. doi: 10.1006/cviu.2000.0897

Moran, N., Hadley, L. V., Bader, M. and Keller, P. E. (2015). Perception of ‘back-channeling’ nonverbal feedback in musical duo improvisation. *PloS One* 10: e0130070.

Paxton, A. and Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods* 45: 329-343. doi: 10.3758/s13428-012-0249-2

Ragert, M., Schroeder, T., and Keller, P.E. (2013). Knowing too little or too much: The effects of familiarity with a co-performer’s part on interpersonal coordination in musical ensembles. *Frontiers in Auditory Cognitive Neuroscience* 4: 368. doi: 10.3389/fpsyg.2013.00368

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In J. Sloboda (Ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition* (pp. 70-90). Oxford: Oxford University Press.

Romero, V., Amaral, J., Fitzpatrick, P., Schmidt, R. C., Duncan, A. W., and Richardson, M. J. (2016). Can low-cost motion-tracking systems substitute a Polhemus system when researching social motor coordination in children? *Behavior Research Methods*. doi: 10.3758/s13428-016-0733-1

Shaffer, L. H. (1984). Timing in solo and duet piano performances. *The Quarterly Journal of Experimental Psychology* 36: 577-595. doi: 10.1080/14640748408402180

Solberg, R. T. and Jensenius, A. R. (2016). Optical or inertial? Evaluation of two motion capture systems for studies of dancing to electronic dance music. Proceedings from *Sound and Music Computing*, Hamburg, Germany. ISSN 2518-3672

Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., and Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research* 34: 97-113. doi: 10.1080/09298210500124208

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 780-785. doi: 10.1109/34.598236

Conflict of Interest Statement

This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding Statement

This work was supported by the Arts and Humanities Research Council [grant number AH/N00308X/1].

Acknowledgments

The authors would like to thank Peter Keller and Nikki Moran for sharing video and motion capture data for use in this project and providing feedback on an earlier version of this paper. Thanks to Simone Tarsitani for help with the data preparation and storage. Recording credit for the Piano Duo goes to Marie Ragert, Kerstin Traeger, Maria Bader, and Jan Bergmann, and recording credit for the Mixed Instrument Duos goes to Kerstin Traeger, Maria Bader, and Jan Bergmann. Recording credit for the String Quartet goes to Corrado Canepa, Paolo Coletta, Nicola Ferrari, Simone Ghisio, Donald Glowinski, Giorgio Gnecco, Maurizio Mancini, Stefano Piana, Roberto Sagoleo, and Giovanna Varni. Thanks also to the musicians of the Piano Duo, Mixed Instrument Duos, and the Quartetto di Cremona string quartet, for their crucial contribution to the creation of the video and motion capture datasets.

Figure captions

Figure 1. Screenshots of one example video from each of the three datasets.

Figure 2. An example of the data preparation and extraction process from the Mixed Instrument Duos. The left panel shows the selection of ROIs for the head of each performer and a corresponding head marker from the MoCap data. The right panel shows the KCF data and MoCap trajectories for the x- and y-coordinates of each performer's head as time series.

Table 1. Summary of Performance Details for Each Dataset

Dataset	No. of video recordings	No. of different performers	No. of trials analysed*	Instrumentation	Mean Duration in seconds (SD)
Piano Duo	7	2	14	two pianists/ vocalists	119.68 (1.78)
Mixed Instrument Duos	3	5	5	cellist, soprano saxophonist, double bassist, two pianists	76.08 (43.14)
String Quartet	8	3	14	violinist, violist, cellist	125.74 (22.92)
Total	**18**	**10**	**33**	**6 instruments**	**115.11 (27.70)**

*Note: A trial was defined as one video-recorded performance by one performer.

Table 2. Median Correlations between the Computer Vision and Motion Capture Data

Region of Interest	Dataset	Number of trials	FD: Median correlation (SD)	OF: Median correlation (SD)	KCF: Median correlation (SD)
Upper Body	Piano Duo	14	.80 (0.14)	.98 (0.07)	.89 (0.09)
	Mixed Instrument Duos	5	.71 (0.16)	.85 (0.06)	.80 (0.07)
	String Quartet	14	.77 (0.08)	.75 (0.26)	.77 (0.25)
	All Datasets	**33**	**.75 (0.13)**	**.87 (0.22)**	**.84 (0.20)**
Head	Piano Duo	14	.73 (0.18)	.94 (0.03)	.95 (0.06)
	Mixed Instrument Duos	5	.72 (0.17)	.92 (0.13)	.92 (0.18)
	String Quartet	14	.83 (0.13)	.80 (0.21)	.94 (0.07)
	All Datasets	**33**	**.79 (0.16)**	**.91 (0.17)**	**.94 (0.10)**

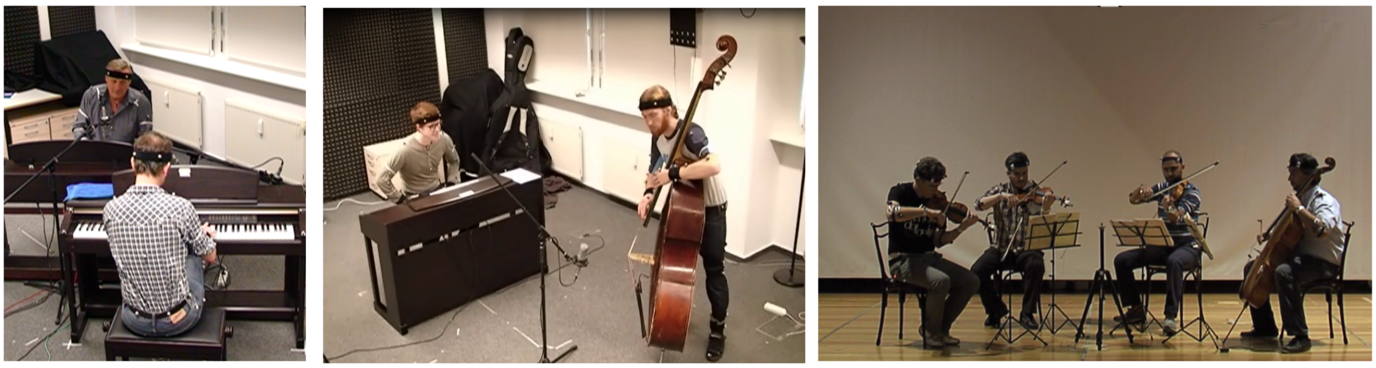
Note: FD = Frame Differencing, OF = Optical Flow, KCF = Kernelized Correlation Filters; x- and y-coordinates are combined into polar coordinates for motion capture, OF, and KCF data

Table 3. Median Correlations between the OF/ KCF and Motion Capture Data, by Dimension

Technique	Region of Interest	Dataset	Number of trials	Median correlation, x-dimension (SD)	Median correlation, y-dimension (SD)
OF	Upper Body	Piano Duo	14	.98 (0.06)	.93 (0.45)
		Mixed Instrument Duos	5	.85 (0.05)	.66 (0.55)
		String Quartet	14	.74 (0.25)	.39 (0.28)
		All Datasets	**33**	**.87 (0.22)**	**.65 (0.41)**
	Head	Piano Duo	14	.94 (0.03)	.82 (0.22)
		Mixed Instrument Duos	5	.91 (0.11)	.85 (0.05)
		String Quartet	14	.81 (0.20)	.57 (0.17)
		All Datasets	**33**	**.92 (0.16)**	**.75 (0.21)**
KCF	Upper Body	Piano Duo	14	.86 (0.10)	.62 (0.34)
		Mixed Instrument Duos	5	.79 (0.07)	.45 (0.51)
		String Quartet	14	.75 (0.24)	.33 (0.41)
		All Datasets	**33**	**.80 (0.19)**	**.55 (0.41)**
	Head	Piano Duo	14	.96 (0.04)	.60 (0.28)
		Mixed Instrument Duos	5	.91 (0.17)	.78 (0.09)
		String Quartet	14	.93 (0.07)	.80 (0.17)
		All Datasets	**33**	**.94 (0.09)**	**.78 (0.22)**

Note: OF = Optical Flow, KCF = Kernelized Correlation Filters

Figure 01.TIF



Piano Duo Mixed Instrument Duos String Quartet

Figure 02.TIFF