Journal of New Music Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/nnmr20

Hindustani raga and singer classification using 2D and 3D pose estimation from video recordings

Martin Clayton, Jin Li, Alison Clarke & Marion Weinzierl

To cite this article: Martin Clayton, Jin Li, Alison Clarke & Marion Weinzierl (03 Apr 2024): Hindustani raga and singer classification using 2D and 3D pose estimation from video recordings, Journal of New Music Research, DOI: 10.1080/09298215.2024.2331788

To link to this article: https://doi.org/10.1080/09298215.2024.2331788

Published online: 03 Apr 2024.

JOURNAL OF NEW MUSIC RESEARCH
https://doi.org/10.1080/09298215.2024.2331788

Routledge
Taylor & Francis Group

Hindustani raga and singer classification using 2D and 3D pose estimation from video recordings

Martin Clayton [a], Jin Li [a,b], Alison Clarke[c] and Marion Weinzierl[c]

[a]Department of Music, Durham University, Durham, UK; [b]School of Computer Science, Shaanxi Normal University, Xi'an, People's Republic of China; [c]Advanced Research Computing, Durham University, Durham, UK

ABSTRACT

Using pose estimation with video recordings, we apply an action recognition machine learning algorithm to demonstrate the use of the movement information to classify singers and the ragas (melodic modes) they perform. Movement information is derived from a specially recorded video dataset of solo Hindustani (North Indian) raga recordings by three professional singers each performing the same nine ragas, a smaller duo dataset (one singer with tabla accompaniment) as well as recordings of concert performances by the same singers. Data is extracted using pose estimation algorithms, both 2D (OpenPose) and 3D. A two-pathway convolutional neural network structure is proposed for skeleton action recognition to train a model to classify 12-second clips by singer and raga. The model is capable of distinguishing the three singers on the basis of movement information alone. For each singer, it is capable of distinguishing between the nine ragas with a mean accuracy of 38.2% (with the most successful model). The model trained on solo recordings also proved effective at classifying duo and concert recordings. These findings are consistent with the view that while the gesturing of Indian singers is idiosyncratic, it remains tightly linked to patterns of melodic movement: indeed we show that in some cases different ragas are distinguishable on the basis of movement information alone. A series of technical challenges are identified and addressed, with code shared alongside audiovisual data to accompany the paper.

ARTICLE HISTORY
Received 15 October 2021
Accepted 10 March 2024

KEYWORDS
Indian music; gesture; movement; machine learning; pose estimation; action recognition

Introduction

Music and gesture studies (Godoy & Leman, 2009; Gritten & King, 2011) explores the relationship between manual gesture and musical performance in ways analogous to the relationship between gesture and spoken language (Goldin-Meadow, 2003; Kendon, 2004; McNeill, 1992, 2005). Indian vocal performance offers a rich field for such study since singers typically use a wide range of expressive and/or functional hand gestures to accompany their own singing. Rahaim's *Musicking Bodies* (2012) takes an ethnographic and phenomenological approach to the khyal genre of Hindustani music,[1] exploring the idea that gesture can offer information complementary to sound and help us to understand singers' musical intentions. Clayton's (2007b) empirical study of performance video recordings addresses the various functions of gesture in a khyal performance and the relationship between melodic and gesture phrases. Further studies by Leante and Clayton have combined observation and interview data in the interpretation of khyal singers' gestures (Dahl et al., 2009; Fatone et al., 2011; Leante, 2009, 2013a, 2013b, 2018). Paschalidou used motion capture recordings of Hindustani dhrupad vocalists to explore relationships between sound and gesture, particularly in terms of perceived effort and apparent manipulation of imagined objects by the singers (Paschalidou & Clayton, 2015; Paschalidou et al., 2016). Pearson (2013) explored the sound-gesture relationship in teaching contexts in the related Karnatak (South Indian) music tradition: Pearson and Pouw (2022) explore gesture–vocal coupling in the same tradition using motion capture data: although not the main focus of the paper, they report an attempt to classify Karnatak ragas using a machine learning approach with motion capture data; they find more success classifying singers than their ragas based on movement data, supporting the idea that singer idiosyncrasy is an important factor. Clayton et al. (2022) explore the same dataset considered in this paper using multimodal machine learning methods.[2]

CONTACT Martin Clayton martin.clayton@durham.ac.uk Department of Music, Durham University, Palace Green, Durham DH1 3RL, UK

[1] Khyal and dhrupad are two of the major vocal genres of Hindustani (North Indian) classical music.
[2] The analysis for Clayton et al. (2022) was carried out after that reported here.

It is clear from these studies that gesture is almost ubiquitous in Indian classical singing; that it has many and overlapping functions and referents in relation to the musical sound (i.e. it can accompany the 'flow' of melodic movement but also punctuate the music rhythmically or afford inter-performer communication); and that there are few or no standardised gestures. Gesturing is quite idiosyncratic, but nonetheless common themes do emerge from qualitative study – Leante's work with Rag Shree, for example, highlights the pervasive use of emphatic hand-raising gestures accompanying melodic slides between scale degrees Re and Pa (^2b–^5) (2009). Given the labour-intensiveness and subjectivity of manual annotation of performance videos and the paucity of motion capture data, this is an area of research that should benefit from the use of computer vision to study movement in video recordings.

Work on movement in Indian music has also looked at topics such as entrainment and performer interaction (Clayton, 2007a; Clayton et al., 2019, 2020; Moran, 2013). This work features both manual annotation and automated movement tracking in the study of video recordings. However, to date no studies have been produced using the pose estimation algorithms that have emerged in recent years, which can produce full skeleton movement, richer data than that employed, for instance, for the tracking of musicians' head movements by Jakubowski et al. (2017).

Action recognition is a basic application in computer vision. Given a video that contains people performing a specific action, e.g. walking or jumping, action recognition models identify the correct class label of the action. Early research focused on extracting hand-crafted features which contain the necessary signature of different actions, for example, using the temporary trajectories of the sparse feature points (Ghamdi et al., 2012) and the optical flow (Ji et al., 2013). Recent research has found that deep neural networks are able to extract robust visual features from images and videos. For example, 2-dimensional Convolutional Neural Network (CNN) architecture can be utilised to compute the feature map for a single RGB or optical flow image, and then temporal features can be extracted from the feature map sequence (Simonyan & Zisserman, 2014); using 3-dimensional CNN, models can process an image sequence directly (Tran et al., 2015).

The main drawback of recognising actions directly from videos is that non-salient RGB information may cause bias. For example, when identifying a person playing football or riding a bicycle, the model may pay more attention to the background than the movement itself. To avoid this, skeleton-based action recognition methods are attracting increasing attention from researchers.

Pose estimation applies machine learning approaches to estimate the position of a human skeleton on the basis of a single photographic image: by repeating the process on successive video frames it is possible to estimate the movement of a human body over the course of a video. Skeleton data such as those derived from pose estimation algorithms are represented as a sequence of point coordinates, where each point denotes a body part of the detected person. A direct way to use the skeleton is drawing the points (body parts, many of them joints) as well as the edges connecting these points (bones) in one frame into an image with a black background, and then processing the image sequence using deep learning-based methods (Liu et al., 2017; Zhang et al., 2017). However, these methods still have large costs in terms of computational complexity since every pixel needs to be computed.

The more efficient way to deal with the skeleton data is using a graph model, where the body parts and bones[3] can be regarded as the vertices and edges. Graph Convolutional Networks can efficiently access the irregular skeleton key points and extract their features in the spatial–temporal domain (Yan et al., 2018). Liu et al. (2020) proposed a multi-scale spatial–temporal graph convolutional network (MS-G3D), which considers different orders of neighbour for each key point in the graph in action recognition.[4]

A number of research groups are currently working with pose estimation algorithms in music-related projects (e.g. Potempski et al., 2021), but to date no other large-scale studies on musicians' movement derived in this way have been published.[5] The current paper asks whether a combination of pose estimation algorithms and MS-G3D networks can be used to classify video clips of khyal singers to enable the singer and/or raga performed to be identified. On the basis of the existing literature cited above, we predict that:

(1) Given the idiosyncratic nature of khyal singers' gestures, it should be possible to recognise individual singers on the basis of their extracted movement alone.
(2) Given the qualitative observations that gestures are related to melodic motion – particularly in alap sections which involve neither explicit rhythmic structure nor co-performer interaction[6] – the differences

[3] In many cases these 'bones' represent actual bones, as between the wrist and elbow; exceptions would include the link between the 'nose' and 'ear' points.
[4] For example, the right elbow is the first order neighbour of the right wrist, while the right shoulder is the second order neighbour of the right wrist.
[5] See however Clayton et al. (2022), which analyses the same dataset as the current paper.
[6] Alap is the free-rhythm introduction to a raga performed at the start of a rendition of Hindustani music. In a khyal concert situation this would be performed with melodic accompaniment on harmonium or sarangi, which

in melodic movements between ragas mean that for a given singer, ragas may be identified from the singer's movement alone. However, since we believe that gestures *typical* of ragas (at least for a particular singer) occur only occasionally, and much of the gesture in between these moments may be more generic, the accuracy of raga recognition may be only slightly above chance level.

(3) Gesture idiosyncrasy means that the task of identifying ragas sung by one singer using a model trained on other singers' movements will be much more difficult, and results are likely to be no better than chance.

These hypotheses and the analyses reported below represent the first step in exploring singers' gestures using machine learning approaches with video data. The study offers a proof of concept, demonstrating that movement data extracted using pose estimation software can enable further explorations of singers' gestures through video recordings. We demonstrate the use of 2D and 3D pose estimation data to study the upper-body movement of Indian classical singers: three singers each perform a common set of nine ragas (melodic modes) in the khyal style. Movement data is extracted using OpenPose (2D pose estimation) and occlusion-robust pose-maps (3D pose estimation). A version of the MS-G3D model is applied in a set of action-recognition tasks: building on the original MS-G3D we add to the position data velocity vectors for each body part, testing whether this improves the accuracy of classification. Here, each raga is regarded as an action class, and all the video clips created from one video share the same class label. In classifying singers from the movement data, each singer is regarded as an action class.

The specific tasks attempted are as follows:

(A) Identify a singer from their skeleton movement (which of three singers is performing in an unseen clip?)
(B) Identify which raga a particular singer is performing (if the system is trained on a set of recordings by the same singer)
(C) Identify a raga, irrespective of which singer is performing
(D) Compare the accuracy of classification models between 2D and 3D pose data
(E) Test whether adding velocity vectors to the input data increases recognition accuracy, using two different strategies

The intention is therefore both to explore the nature of gestures made accompanying Hindustani vocal music, and to explore the robustness of different approaches to the application of pose estimation and action recognition deep learning methods in the investigation of musicians' gestures. The aim of this research is not to produce recognition systems based on video data alone, since in real-life situations audio data would almost always be available alongside video. Rather, our intention is to explore empirically the complementary notions that singers' gestures may contain raga-specific information, and that singers' gestures are idiosyncratic. Is there enough raga-specific information for this to potentially aid automatic classification (for example distinguishing between two ragas with similar pitch profiles)? On the other hand, are they sufficiently idiosyncratic that their main usefulness would be in searching for videos of particular singers? For either or both of these future tasks, what kind of approach is likely to be most effective? Beyond the question of possible applications of such a system, of course, our aim is to shed light on musical gesturing that can feed back also into qualitative studies.

Data

Music recordings

Clayton and colleagues have to date published 17 complete raga performances Hindustani (North Indian) raga music, featuring multitrack audio, static video shots and annotations (Clayton et al., 2021a). These recordings have the advantage that they are taken from real-life performances, mostly in front of live audiences. They do not however allow a systematic comparison between singers' gesturing while performing a common set of ragas. The new solo recording collection created for this study fills this lacuna: three professional singers were asked to perform unaccompanied alap renditions of nine Hindustani ragas in khyal style, for approximately 3 min each. They were also asked to perform a set of shorter 'snapshots' of each raga, labelled pakad ('catch'). In most cases two takes of each raga alap were recorded, producing a collection of 55 3-minute raga performances (duration 165–221 s) and 27 pakad clips (9–96 s).[7] Three cameras were used: one central and one each to the right and left of centre (only the central view is used in this study). The singers are Apoorva Gokhale (henceforth AG), Chiranjeeb Chakraborty (CC) and Sudokshina Chatterjee (SCh). Recordings were made and edited in the studios of the Durham University Music Department by Simone Tarsitani. Musicians were informed of the purpose for

is absent in this solo dataset. (Singing in a similar style but with drum accompaniment, as is common in khyal, is also referred to as singing alap.)

[7] In one case only one take was usable, while in a couple of cases three usable takes of one raga were captured. See Table 1 for details.

Figure 1. Singers, from left to right: Apoorva Gokhale (AG), Chiranjeeb Chakraborty (CC), and Sudokshina Chatterjee (SCh). Video stills (detail).

which their performances were recorded; they gave their informed consent in writing for the recordings to be used for academic research, including publication for educational and not-for-profit purposes and deposit in data repositories or archives for not-for-profit educational and research use only (Figure 1).

These solo recordings are complemented by a set of duo recordings of the same nine ragas by SCh with tabla player Subrata Manna. The purpose of these recordings was to test whether the same singer and ragas could be identified by the action recognition models in the duo format, where we would expect SCh's gesture to be affected by the different musical context (the presence of a regular beat and a co-performer), and the camera angle and distance from the singer would also be different. These recordings, made by the musicians themselves in Kolkata, follow a common format of a brief alap followed by a short performance of a khyal in slow ektal.[8]

Finally, we also make use of concert recordings made in Durham in a public concert format, featuring the singers accompanied by the most common instruments used in khyal: tabla and harmonium. SCh's performance is of Miyan ki Malhar, one of our set of nine ragas. AG's is of Raga Yaman and CC's of Raga Bhatiyar, which are not part of the 9-raga set and therefore do not allow us to test the raga recognition model.

The ragas selected were, in alphabetical order, Bageshree, Bahar, Bilaskhani Todi, Jaunpuri, Kedar, Marwa, Miyan ki Malhar, Nand, and Shree. The selection, decided by Clayton and Laura Leante, was intended to cover a wide range of possibilities in terms of time of performance (morning to night),[9] mood (light to serious), typical velocity and directness of melodic movement, and favoured pitch range (upper or lower tetrachord) (Table 1).

Pose estimation and post-processing

2D pose estimation was carried out using OpenPose (Cao et al., 2021) to extract the key points (hands, head, shoulders, wrists, mid hip, elbows, etc.) of the musicians (see Figure 4 below).

As OpenPose runs most efficiently on Graphics Processing Units (GPUs), a Google Colab[10] notebook was created to enable OpenPose to be run on a remote GPU. This allows the software to be run from any location via a web browser, automating the installation of OpenPose on the remote server and removing the need for users to have a GPU. The Colab notebook also provides a user interface to allow parameters for post processing to be selected. All of the relevant code is shared on a GitHub repository (Clarke et al., 2021) linked to the Open Science Framework project containing the audiovisual recordings (Clayton et al., 2021b).

The extracted data was then further processed as follows:

- As the musicians were seated on the floor, the lower body parts were removed from the data.
- The key points were connected by lines representing the 'bones' to allow more intuitive visualisation.
- In order for the further post-processing steps to be able to rely on previous frames, the person numbering had to be made consistent by sorting individuals

[8] Ektal is the most common tala (metre) used for slow tempo khyals. Its 12 time units (matras) typically span 40–60 s.

[9] Hindustani ragas are associated with particular times of day, or in some cases seasons.

[10] Google Colab, website (last visited August 2021) https://research.google.com *'colaboratory*

Table 1. Solo, duo and concert videos.

Code	Singer	Raga	Type
AG_1a_Jaun	Apoorva Gokhale	Jaunpuri	Solo alap
AG_1b_Jaun	Apoorva Gokhale	Jaunpuri	Solo alap
AG_2a_Marwa	Apoorva Gokhale	Marwa	Solo alap
AG_2b_Marwa	Apoorva Gokhale	Marwa	Solo alap
AG_3a_Bag	Apoorva Gokhale	Bageshree	Solo alap
AG_3b_Bag	Apoorva Gokhale	Bageshree	Solo alap
AG_4a_Nand	Apoorva Gokhale	Nand	Solo alap
AG_4b_Nand	Apoorva Gokhale	Nand	Solo alap
AG_5a_MM	Apoorva Gokhale	Miyan ki Malhar	Solo alap
AG_5b_MM	Apoorva Gokhale	Miyan ki Malhar	Solo alap
AG_6a_Bilas	Apoorva Gokhale	Bliaskhani Todi	Solo alap
AG_6b_Bilas	Apoorva Gokhale	Bliaskhani Todi	Solo alap
AG_7a_Bahar	Apoorva Gokhale	Bahar	Solo alap
AG_7b_Bahar	Apoorva Gokhale	Bahar	Solo alap
AG_8_Kedar	Apoorva Gokhale	Kedar	Solo alap
AG_9a_Shree	Apoorva Gokhale	Shree	Solo alap
AG_9b_Shree	Apoorva Gokhale	Shree	Solo alap
CC_1a_Bilas	Chiranjeeb Chakraborty	Bliaskhani Todi	Solo alap
CC_1b_Bilas	Chiranjeeb Chakraborty	Bliaskhani Todi	Solo alap
CC_2a_Jaun	Chiranjeeb Chakraborty	Jaunpuri	Solo alap
CC_2b_Jaun	Chiranjeeb Chakraborty	Jaunpuri	Solo alap
CC_3a_MM	Chiranjeeb Chakraborty	Miyan ki Malhar	Solo alap
CC_3b_MM	Chiranjeeb Chakraborty	Miyan ki Malhar	Solo alap
CC_4a_Nand	Chiranjeeb Chakraborty	Nand	Solo alap
CC_4b_Nand	Chiranjeeb Chakraborty	Nand	Solo alap
CC_5a_Shree	Chiranjeeb Chakraborty	Shree	Solo alap
CC_5b_Shree	Chiranjeeb Chakraborty	Shree	Solo alap
CC_6a_Kedar	Chiranjeeb Chakraborty	Kedar	Solo alap
CC_6b_Kedar	Chiranjeeb Chakraborty	Kedar	Solo alap
CC_7a_Marwa	Chiranjeeb Chakraborty	Marwa	Solo alap
CC_7b_Marwa	Chiranjeeb Chakraborty	Marwa	Solo alap
CC_8a_Bag	Chiranjeeb Chakraborty	Bageshree	Solo alap
CC_8b_Bag	Chiranjeeb Chakraborty	Bageshree	Solo alap
CC_9a_Bahar	Chiranjeeb Chakraborty	Bahar	Solo alap
CC_9b_Bahar	Chiranjeeb Chakraborty	Bahar	Solo alap
SCh_1a_Bilas	Sudokshina Chatterjee	Bilaskhani Todi	Solo alap
SCh_1b_Bilas	Sudokshina Chatterjee	Bilaskhani Todi	Solo alap
SCh_2a_Jaun	Sudokshina Chatterjee	Jaunpuri	Solo alap
SCh_2b_Jaun	Sudokshina Chatterjee	Jaunpuri	Solo alap
SCh_3a_MM	Sudokshina Chatterjee	Miyan ki Malhar	Solo alap
SCh_3b_MM	Sudokshina Chatterjee	Miyan ki Malhar	Solo alap
SCh_3c_MM	Sudokshina Chatterjee	Miyan ki Malhar	Solo alap
SCh_4a_Nand	Sudokshina Chatterjee	Nand	Solo alap
SCh_4b_Nand	Sudokshina Chatterjee	Nand	Solo alap
SCh_5a_Shree	Sudokshina Chatterjee	Shree	Solo alap
SCh_5b_Shree	Sudokshina Chatterjee	Shree	Solo alap
SCh_6a_Kedar	Sudokshina Chatterjee	Kedar	Solo alap
SCh_6b_Kedar	Sudokshina Chatterjee	Kedar	Solo alap
SCh_6c_Kedar	Sudokshina Chatterjee	Kedar	Solo alap
SCh_7a_Marwa	Sudokshina Chatterjee	Marwa	Solo alap
SCh_7b_Marwa	Sudokshina Chatterjee	Marwa	Solo alap
SCh_8a_Bag	Sudokshina Chatterjee	Bageshree	Solo alap
SCh_8b_Bag	Sudokshina Chatterjee	Bageshree	Solo alap
SCh_9a_Bahar	Sudokshina Chatterjee	Bahar	Solo alap
SCh_9b_Bahar	Sudokshina Chatterjee	Bahar	Solo alap
AG_P1_MM	Apoorva Gokhale	Miyan ki Malhar	Solo pakad
AG_P2_Jaun	Apoorva Gokhale	Jaunpuri	Solo pakad
AG_P3_Kedar	Apoorva Gokhale	Kedar	Solo pakad
AG_P4_Bahar	Apoorva Gokhale	Bahar	Solo pakad
AG_P5_Shree	Apoorva Gokhale	Shree	Solo pakad
AG_P6_Nand	Apoorva Gokhale	Nand	Solo pakad
AG_P7_Bag	Apoorva Gokhale	Bageshree	Solo pakad
AG_P8_Marwa	Apoorva Gokhale	Marwa	Solo pakad
AG_P9_Bilas	Apoorva Gokhale	Bliaskhani Todi	Solo pakad
CC_P1a_Bilas	Chiranjeeb Chakraborty	Bliaskhani Todi	Solo pakad
CC_P1b_Bilas	Chiranjeeb Chakraborty	Bliaskhani Todi	Solo pakad
CC_P2_Jaun	Chiranjeeb Chakraborty	Jaunpuri	Solo pakad
CC_P3_MM	Chiranjeeb Chakraborty	Miyan ki Malhar	Solo pakad

(*continued*)

Table 1. Continued

Code	Singer	Raga	Type
CC_P4_Nand	Chiranjeeb Chakraborty	Nand	Solo pakad
CC_P5_Shree	Chiranjeeb Chakraborty	Shree	Solo pakad
CC_P6_Kedar	Chiranjeeb Chakraborty	Kedar	Solo pakad
CC_P7_Marwa	Chiranjeeb Chakraborty	Marwa	Solo pakad
CC_P8_Bag	Chiranjeeb Chakraborty	Bageshree	Solo pakad
CC_P9_Bahar	Chiranjeeb Chakraborty	Bahar	Solo pakad
SCh_P1a_Bilas	Sudokshina Chatterjee	Bliaskhani Todi	Solo pakad
SCh_P1b_Bilas	Sudokshina Chatterjee	Bliaskhani Todi	Solo pakad
SCh_P2a_Jaun	Sudokshina Chatterjee	Jaunpuri	Solo pakad
SCh_P2b_Jaun	Sudokshina Chatterjee	Jaunpuri	Solo pakad
SCh_P3a_MM	Sudokshina Chatterjee	Miyan ki Malhar	Solo pakad
SCh_P3b_MM	Sudokshina Chatterjee	Miyan ki Malhar	Solo pakad
SCh_P4a_Nand	Sudokshina Chatterjee	Nand	Solo pakad
SCh_P4b_Nand	Sudokshina Chatterjee	Nand	Solo pakad
SCh_P5a_Shree	Sudokshina Chatterjee	Shree	Solo pakad
SCh_P5b_Shree	Sudokshina Chatterjee	Shree	Solo pakad
SCh_P6a_Kedar	Sudokshina Chatterjee	Kedar	Solo pakad
SCh_P6b_Kedar	Sudokshina Chatterjee	Kedar	Solo pakad
SCh_P7a_Marwa	Sudokshina Chatterjee	Marwa	Solo pakad
SCh_P7b_Marwa	Sudokshina Chatterjee	Marwa	Solo pakad
SCh_P8a_Bag	Sudokshina Chatterjee	Bageshree	Solo pakad
SCh_P8b_Bag	Sudokshina Chatterjee	Bageshree	Solo pakad
SCh_P9a_Bahar	Sudokshina Chatterjee	Bahar	Solo pakad
SCh_P9b_Bahar	Sudokshina Chatterjee	Bahar	Solo pakad
SCh_Duo_Bilas	Sudokshina Chatterjee	Bilaskhani Todi	Duo ektal
SCh_Duo_Jaun	Sudokshina Chatterjee	Jaunpuri	Duo ektal
SCh_Duo_MM	Sudokshina Chatterjee	Miyan ki Malhar	Duo ektal
SCh_Duo_Nand	Sudokshina Chatterjee	Nand	Duo ektal
SCh_Duo_Shree	Sudokshina Chatterjee	Shree	Duo ektal
SCh_Duo_Kedar	Sudokshina Chatterjee	Kedar	Duo ektal
SCh_Duo_Marwa	Sudokshina Chatterjee	Marwa	Duo ektal
SCh_Duo_Bag	Sudokshina Chatterjee	Bageshree	Duo ektal
SCh_Duo_Bahar	Sudokshina Chatterjee	Bahar	Duo ektal
NIR_SCh_Malhar	Sudokshina Chatterjee	Miyan ki Malhar	Concert

by x-coordinate (which generally did not change due to the seated position of the musicians).[11]

- A confidence threshold was used to improve the smoothness of the prediction. Any key point candidate with a confidence level, given by the pose detection software, that was lower than a certain threshold was replaced by the same key point from the previous frame if that had a higher confidence.
- Existing jitter in the detected movements, due to inaccuracies in the pose detection, was smoothed using a Savitzky–Golay filter (Savitzky & Golay, 1964).[12]
- The output was rendered as videos for visual inspection, with skeletons either overlaid on the video, or shown on their own against a black background.

The output files from OpenPose, which creates one JSON file per frame, were combined into a single CSV file for each person detected in the video, with one row per frame. Parameters for the post-processing were selected manually, and aim to produce a smoothness of movement which matches that observable in the videos themselves: in other words, to remove the noise inherent in the pose estimation process without eliminating smaller movements that are visible to the naked eye. (It is also found that if movement is smoothed over too long a window, a time lag between the video and the skeleton movement becomes apparent to the viewer.)

Visual inspection suggests that the combination of OpenPose's pose detection and post-processing produces an accurate result. Areas of weakness include some instability in the estimation of the elbow positions compared to the more stable shoulder and wrist points, and a tendency for the Mid Hip point to be estimated slightly off-centre.

To extract the 3D skeleton from the monocular RGB input, the method proposed by Mehta et al. (2018) is introduced. A CNN-based model is trained to predict the 3D skeleton coordinates of multiple persons in one input image. The post-processing operation of the 3D results is exactly the same as that described above for 2D results from OpenPose.

Normalisation

If the movement data extracted by pose estimation algorithms is used for the action recognition model, the

[11] A more complex script for identifying multiple people is shared as part of our code as 'run_openpose_adaptive.py'. In the current study, since musicians remained seated and there were no occlusion issues, this was not required.

[12] Using a smoothing window of 13 frames and second order polynomial.

Figure 2. Directly normalising the absolute coordinates to [0, 1] will distort the skeleton and amplify the horizontal movement with respect to the vertical.

absolute size of the skeleton (in pixels) may bias the classification: the differences between movements are subtle, thus a skeleton of the same size may be a more salient factor in classification. This is likely to make singer identification more accurate when working within the solo dataset, for example, but less accurate when mixing solo, duo and concert videos. To counteract this problem and make the model more robust, the following normalisation process is implemented.

Here the normalisation of 2-dimensional skeleton data is described as an example. We use the minimum and maximum absolute coordinates to obtain a bounding box including all body parts over a whole long video. A simple normalisation can be obtained by dividing the absolute coordinates of each body part by the width and height of the bounding box. However, this may change the ratio of the height and width of the body shape, which is illustrated in the Figure 2. Moreover, the horizontal movement is magnified compared to the vertical movement if we normalise in this way.

In order to preserve the aspect ratio of in the original video, we extend the bounding box from the rectangle to the square. In detail, the coordinates of the k-th body part in one video are denoted as $X_k = \{x_{k,1}, x_{k,2}, \ldots, x_{k,T}\}$, $Y_k = \{y_{k,1}, y_{k,2}, \ldots, y_{k,T}\}$, where T is the maximum frame index of this video. Then the set of all key points in this video can be denoted as $X = \{X_1, X_2, \ldots, X_K\}$ and $Y = \{Y_1, Y_2, \ldots, Y_K\}$, where K is the number of body parts detected. The bounding box of the musician can be determined by the coordinate of the top-left and bottom-right points.

$$\begin{cases} x_{\max} = \max(x_{k,t}), x_{k,t} \in X \\ x_{\min} = \min(x_{k,t}), x_{k,t} \in X \\ y_{\max} = \max(y_{k,t}), y_{k,t} \in Y \\ y_{\min} = \min(y_{k,t}), y_{k,t} \in Y \end{cases} \tag{1}$$

The width and height of the bounding box can be calculated by

$$\begin{cases} w = x_{\max} - x_{\min} \\ h = y_{\max} - y_{\min} \end{cases} \tag{2}$$

To ensure that the ratio of horizontal or vertical movements will not change, we let

$$l = \max(w, h) \tag{3}$$

and

$$\begin{cases} \tilde{x}_{\min} = x_{\min} - \frac{(l-w)}{2} \\ \tilde{y}_{\min} = y_{\min} - \frac{(l-h)}{2} \end{cases} \tag{4}$$

A square bounding box can be denoted by $(\tilde{x}_{\min}, \tilde{y}_{\min}, l, l)$, where the person stays in the centre of the box. The normalised coordinates can be calculated as

$$\begin{cases} \tilde{x}_{k,t} = \frac{(x_{k,t} - \tilde{x}_{\min})}{l} \\ \tilde{y}_{k,t} = \frac{(y_{k,t} - \tilde{y}_{\min})}{l} \end{cases} \tag{5}$$

By transforming the coordinates using Equations (1)–(5), we obtain the relative coordinate values, while the ratio of the movement in horizontal and vertical dimensions is also preserved. More importantly, the influence of resolution of the different videos will be avoided. The only difference with the 3-dimesional skeleton is computing an additional coordinate $\tilde{z}_{k,t}$ following Equations (1)–(5). Examples of the resulting square boxes are illustrated in Figure 3.

Division into clips

The machine learning model used for action recognition (MS-3GD) requires video clips with a fixed number of frames. In many action recognition problems, clips are quite short: for example, an 'action' lasting 2–3 s might be cut into clips of 1-second each, using a sliding window

Figure 3. The bounding box in different videos of SCh. As we normalise the coordinates into [0, 1] by dividing the length of the square, the influence of the different size of the image in the video frame is avoided.

advancing one frame at a time (Tao & Papadias, 2006). In khyal singing, vocal phrases often last 10 s or more, thus we chose to use longer clips in case the identifying movement patterns evolved over a longer time-scale. Therefore, each long video is split into 12-second clips (25 fps video, thus each includes 300 frames). To improve the diversity of the training data, the temporal stride between two clips is set at 40 frames, where a small random value (from −20 to 20) is added to adjust the index of the start frame. Following this procedure, more than 7,200 clips were generated, about 80% of which are from solo videos and the rest are from duo or concert videos. (In practice, the musicians' pose information is extracted from the whole videos, and the 'clips' are created by segmenting the CSV files of movement data by using a sliding window.)

Shared dataset

The full dataset is publicly shared on OSF (Clayton et al., 2021b) and includes the following:

(1) All solo, duo and concert raga videos, as recorded and with final movement data (2D skeleton, upper body parts from OpenPose) overlaid[13]
(2) 3D movement data visualised with black background for all solo, duo and concert raga takes
(3) Both 2D and 3D JSON files containing the output of pose estimation algorithms (before post-processing)
(4) Both 2D and 3D CSV files (post-processed, single file per take per musician)
(5) Duo and concert videos with predicted labels for singer and raga printed in the top-left corner
(6) CoLab notebook and entimement-openpose python library (linked to GitHub)
(7) Details of parameters and initial values for models

[13] Although only one view is analysed here, all three views are shared in the OSF dataset.

Method

Training and tests sets (splits)

In order to explore our five main tasks, after the set of movement data for each of the 12-second video clips is generated, six different splits of training and test set are considered.

(1) Singers separate: The model is initialised for each of the three singers separately. One take of each raga is used for the test set and all other takes (including the Pakad clips that are longer than 12 s) are used to train the model.[14] In this way, we test whether the MS-G3D model can successfully classify ragas as sung by a specific singer, and whether this task is easier with some singers than others. If successful, this would demonstrate that singers' gesturing while performing different ragas is different, and that there is some consistency in this distinction between takes. This split is used to address task B (can we identify which raga a specific singer is singing?).
(2) Unseen singer: All takes for one singer are used for the test set, while all other takes are used for the training data. This model tries to classify the raga sung by a new musician who was totally unseen before. If successful, this approach would demonstrate consistency between the raga-specific gesturing of different singers. Given that Indian singers' gestures are understood to be highly idiosyncratic, this is expected to be much more challenging that split 1. This split helps us to address task C (can we identify a raga irrespective of the singer?), in the particularly difficult case in which the singer used in the test stage has not been employed to train the model.
(3) Duos: All solo clips are used for training data, and the singer's movement in the Duo clips (SCh) are

[14] We only have one usable take of AG's Raga Kedar, which is included in the test data: this is likely to affect the accuracy of identification for this singer/raga combination.

used as the test data. If successful, this would demonstrate that distinctions between gesturing with different ragas is robust not only between takes, but also in the presence of an accompanist (and thus when gestures communicating to the co-performer would be expected), and with a different recording set-up (camera angle, etc.). This split is used to further address task B (can we identify which raga a specific singer is singing?), but does not restrict the data to a single camera set-up.

(4) Concert: All solo clips are used for training data, and the singer in the concert clips is used for the test data. All concert clips are used for singer recognition, but only SCh's concert clips are used for the raga recognition. The singers now have two accompanists, and the camera angles are different (although they are still frontal shots). Like the Duo split, this one is used to further address task B, but does not restrict the data to a single camera set-up.

(5) Random solo: A split of the solo video clips only, for which we randomly selected nine takes (one take for each raga and three for each musician) for the test. This split can be used for evaluating the performance of both raga and musician identification. This split enables us to address task A (can we identify the singer?) and C (can we identify a raga irrespective of the singer?).

(6) Random all: A random split of solo, duo and concert clips. Since the training set now includes duo and concert as well as solo clips, the model is expected to be more robust for musician identification (task A).

Tasks D (comparing 2D and 3D data) and E (testing the impact of using velocity data) are explored across the various splits.

Training the action recognition model

In this paper, a skeleton-based action recognition method is proposed to identify the ragas and musicians. Firstly, the output of our post-processed data from OpenPose comprises 17 key points (from index 0 to 16) for one person, which are shown on the left side of Figure 4. Specifically, the related body parts of the key points are nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, middle hip, right eye, right ear, left eye, left ear, right knee, right ankle, left knee and left ankle. Since the singers sit cross-legged on the floor, the movement of lower body parts (the knees and ankles) contains little useful information, and in any case is estimated very poorly by OpenPose. Moreover, the detection of the ears is unreliable because of occlusion. Therefore,

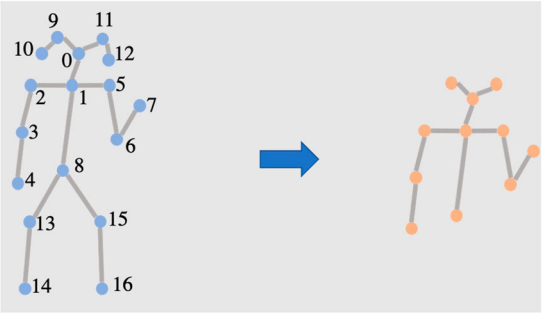

Figure 4. The skeleton detected (left) and input into the model (right).

these six parts are dropped when we train the skeleton-based model. The simplified skeleton is shown on the right in the Figure 4, where the 11 orange circles represent the key points that are finally considered when we generate the position and velocity vectors.

The input of the model consists of two channels, the position and velocity. The input in the position channel is a tensor with dimension (C_0, T, K), where C_0 equals 3 in a 2-dimensional skeleton, containing the normalised horizontal and vertical coordinates, and the confidence value returned by OpenPose. For the 3-dimensional skeleton, the depth coordinate is included, thus C_0 equals 4. T is the total number of frames in each video clip, which equals 300. K denotes the total number of body parts, which is 11.

The distinction between movement patterns in the ragas is much less obvious than that between actions such as running and jumping used in general action recognition tasks, and it may be that velocity of movement is sometimes a distinguishing factor in our case. In this paper therefore, velocity information is added as an independent input, so that the model can use this information.

The velocity vector is therefore considered in the second channel. Specifically, the horizontal, vertical and depth velocity values of the k-th body part in t-th frame is defined as

$$\begin{cases} \dot{x}_{k,t} = \tilde{x}_{k,t} - \tilde{x}_{k,t-\Delta} \\ \dot{y}_{k,t} = \tilde{y}_{k,t} - \tilde{y}_{k,t-\Delta} \\ \dot{z}_{k,t} = \tilde{z}_{k,t} - \tilde{z}_{k,t-\Delta} \end{cases} \tag{6}$$

respectively, where Δ is the frame interval for calculating the velocity vector. When $t \leq \Delta$, we simply let

$$\begin{cases} \dot{x}_{k,t} = \dot{x}_{k,\Delta+1} \\ \dot{y}_{k,t} = \dot{y}_{k,\Delta+1} \\ \dot{z}_{k,t} = \dot{z}_{k,\Delta+1} \end{cases} \tag{7}$$

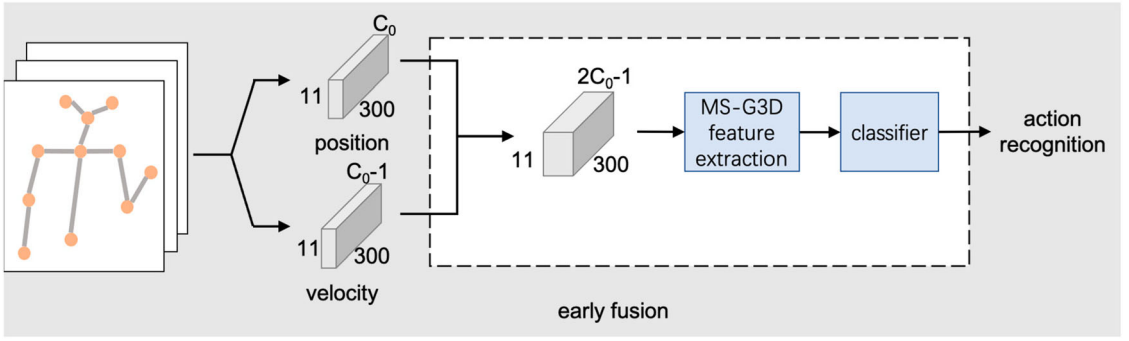

Figure 5. The early fusion strategy.

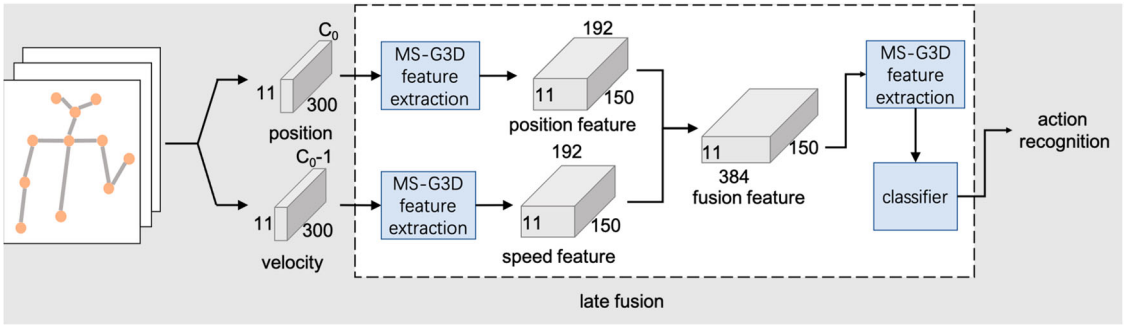

Figure 6. The late fusion strategy.

Therefore, the input of the velocity channel is a tensor with dimension $(C_0 - 1, T, K)$, similar to the position channel.

In order to combine the position and velocity information, two fusion strategies are evaluated in the paper, termed 'early fusion' and 'late fusion', and compared to the original MS-G3D model. For the early fusion, the tensors of position and velocity are concatenated to form a new tensor with dimension $(2C_0 - 1, T, K)$. This tensor is fed into the MS-G3D module, which captures complex regional spatial–temporal joint correlations as well as long-range spatial and temporal dependencies. In the following stage, a linear classifier is implemented to predict the class label (raga or musician) of the short clips (Figure 5).

For the late fusion strategy, tensors of position and velocity are processed by two separate MS-G3D modules to extract semantic visual features, where the output dimension in both channels is $(C_1, T/2, K)$. Features from the two channels are concatenated to a new tensor with dimension $(2C_1, T/2, K)$. Next, the fusion feature is fed to another MS-G3D module as well as the classifier, to predict the score of all the actions that are required to be identified. The illustration of the late fusion model is shown in Figure 6.

To train the model, the mini-batch gradient descent (Li et al., 2014) is used for optimising the parameters. Specifically, one batch that contains several short clips is input to the model, and the predictions of these clips are calculated by the forward inference process. Then the training loss is computed between the ground truth (i.e. the real raga labels of these clips) and the prediction, which is further utilised to update the parameters by a backward propagation process. For each training epoch, the training data is fed to the model once in a random order. In the experiments, 16 samples are processed in parallel and the number of training epoch is set to 50 empirically.

Results

In order to evaluate the performance of methods for an action recognition task, classification accuracy is used: the percentage of correctly predicted clips in the test set is calculated. In the results reported below, the original MS-G3D method is compared to the proposed two-pathway models, including early fusion and late fusion strategies. Both 2D and 3D skeletons are fed into these three models, allowing us to compare six sets of results for the two classification tasks (raga and musician) across six different splits of the data.

Singer classification

The performance of the musician classification for the different methods are reported for splits 3–6 (Table 2).

Table 2. Accuracy of singer classification on splits 3–6.

	Split 3 Duo	Split 4 Concert	Split 5 Random solo	Split 6 Random all
2D MS-G3D	48.6%	61.0%	100.0%	77.5%
2D early fusion	**93.6%**	**66.0%**	99.8%	**86.7%**
2D late fusion	87.6%	55.0%	100.0%	57.1%
3D MS-G3D	**59.0%**	**65.3%**	100.0%	72.9%
3D early fusion	56.8%	61.0%	100.0%	**83.3%**
3D late fusion	50.6%	54.7%	100.0%	77.1%

The easiest case is split 5 (random solo), where the model is trained and tested on clips from the solo dataset: in this case all the models show very good performance (close to 100% accuracy). In splits 3 and 4, the model is trained on solo clips and tested on the other datasets (duo and concert clips). In split 3 (duos) the accuracy varies significantly between the different models, with 2D early fusion the best performing (93.6%). Only SCh appears in Duo videos, and the unbalanced test data may account for the variability in the results. In split 4 (concerts) the number of clips of different musicians in the test set was balanced by randomly excluding some clips of SCh, since a longer video clip for SCh was used; a similar process was used to balance the sample in split 6. The variation is much less for split 4, with the same model (2D early fusion) again performing best. Results on split 6 (random all) are better than that on split 4, as expected since all camera angles are represented in both training and test sets. In this case the 2D data gives slightly higher accuracy (86.7 vs 83.3%).

Raga classification

Results of the raga classification for different musicians in split 1 – in which the model is trained and tested on each singer's solo clips separately – are shown in Table 3. The average classification accuracy using 2D pose data is about 8% better than that using 3D data. The mean accuracy for the raga classification ranges from 26.9 to 38.2% according to different models; the mean accuracy of all models is 32.6%, compared to the random guess accuracy of 11.1% (one of nine ragas).[15] The early fusion strategy in the two-pathway method achieves the highest mean classification accuracy, from which two conclusions are supported. Firstly, adding the velocity information only marginally improves the mean accuracy of classification. Secondly, the early fusion performs better than the late fusion in this task. Last but not least, the accuracy is highest for SCh, suggesting greater consistency between takes than is the case for AG and CC.

The confusion matrices (Stehman, 1997) of the classification results for musicians (Figure 7) show that the singers differ somewhat in which ragas are easier and which harder to classify. Specifically, Bilaskhani Todi and Jaunpuri sung by AG are successfully classified, but Marwa, Kedar and Shree not so. CC's Bahar, Bageshree and Miyan Malhar are easier to identify; Marwa and Kedar are again poorly classified alongside Bilaskhani Todi. For SCh, most ragas are well identified except the Miyan Malhar and Shree.

[15] Clayton et al. (2022), attempting the same task using a different approach and data from the two wrists only, report similar results: 36.3, 31.8 and 39.2% accuracy for the respective singers (mean accuracy 35.8%).

Table 3. Accuracy of raga classification for different musicians in split 1 (singers separate).

	AG	CC	SCh	Mean
2D MS-G3D	29.0%	**33.7%**	**49.0%**	37.2%
2D early fusion	35.3%	33.0%	46.4%	**38.2%**
2D late fusion	**37.7%**	25.1%	45.1%	34.6%
3D MS-G3D	28.0%	20.7%	**39.1%**	29.2%
3D early fusion	**30.2%**	**25.4%**	33.3%	**29.6%**
3D late fusion	27.1%	19.5%	34.2%	26.9%

Results of the raga classification on splits 2–5 are displayed in Table 4. According to results using split 2 (unseen singer), identifying the raga performed by a new singer is challenging since the prediction is only slightly better than the random guess. Classification accuracies on split 3 (duos) are higher than random guess but lower than that on split 1. The training set of split 4 (concert) is the same as that of the split 3; therefore, we tested the model directly with parameters trained in the split 3. Split 4 results show that the models perform well in identifying the raga of the SCh concert video (Miyan Malhar). Split 5 (random solo) achieves a higher accuracy than split 1, where training and testing is restricted to the same singer.

Discussion

Pose estimation seems to offer a robust way to extract movement information from music performance videos, which can then be used to explore music performance. Machine learning approaches based on existing action detection systems proved capable of classifying both ragas and musicians, with varying degrees of accuracy. We outlined three specific tests above:

(1) Given the idiosyncratic nature of khyal singers' gestures, it should be possible to recognise individual singers on the basis of their extracted movement alone.

This was demonstrated. In the easiest split (using solo videos only) accuracy was close to 100%. In the hardest split, where the system is trained on solo videos and tested on group videos, accuracy was still high, at 68.8–73.0%, using the early fusion approach (random guess would be 33%). This indicates that the model is able to generalise to different camera angles and video resolution

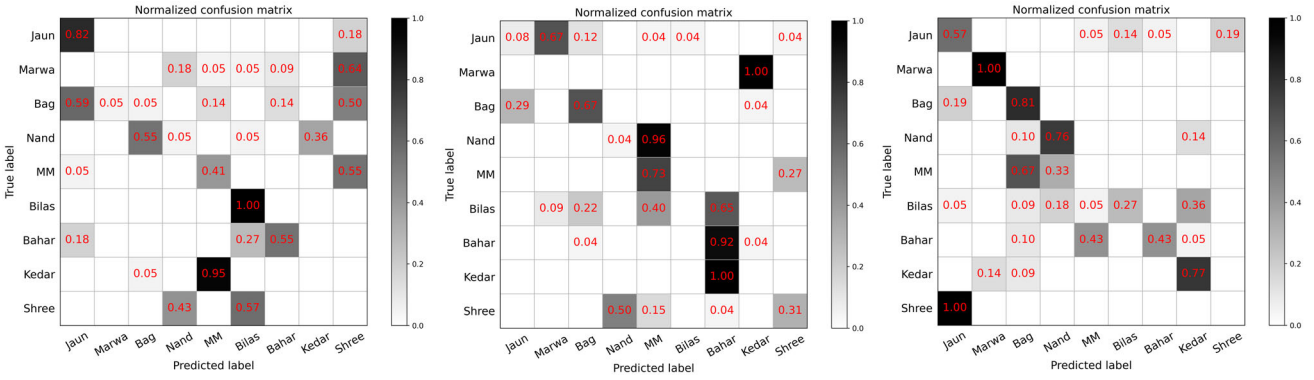

Figure 7. Confusion matrices, from left to right: AG, CC and SCh.

Table 4. Accuracy of the raga classification on splits 2–5.

	Split 2 Unseen singer	Split 3 Duo	Split 4 Concert	Split 5 Random solo
2D MS-G3D	10.0%	25.9%	**96.0%**	**42.7%**
2D early fusion	16.7%	**28.4%**	95.2%	39.7%
2D late fusion	**16.8%**	20.7%	84.7%	40.5%
3D MS-G3D	**17.2%**	16.1%	76.7%	**38.8%**
3D early fusion	15.8%	**24.0%**	**100.0%**	32.2%
3D late fusion	15.9%	21.2%	94.5%	26.3%

(at least after normalisation). It is worth noting that for SCh, the models worked very well in identifying her in the duo clips, but very poorly in the concert clips. This could be due either to the bigger difference in camera angle (the concert clips are taken from a high camera looking down on the stage), and/or due to some differences in her gesturing (in the concert clips she interacts much more with her accompanists). Both of these factors apply also to the AG and CC concert clips, however, where the model nonetheless performed much better in singer identification.

(2) For a given singer, ragas may be identified from the singer's movement alone, but accuracy may be only slightly above chance level.

The possibility of identifying ragas from a singer's movements has been demonstrated. In the best case in split 1 the classification accuracy reached 49%, more than slightly above the random guess level of 11%. However, the results are patchy: they are better for some singers than others; for each singer some ragas are classified with very high accuracy, and others very poorly. For each singer, then, movement consistency between takes seems to be greater for some ragas than others.

To help to interpret the results, we created versions of the duo and concert videos with the model predictions printed in the top-left corner (all visualisation use the output of the '2D early fusion' model, which achieves the best accuracy in most cases). We note the following:

In some cases, accuracy is affected by the inclusion of passages where the singer is moving very little or not at all. In AG's concert video, for example, the first 47 s (in which she has hardly begun to gesture) are mis-labelled as CC: this alone accounts for 26% of the AG concert clips. In the duo recordings, the start of each video is labelled as Jaunpuri: the model seems to default to this raga when the singer is moving minimally, and this helps to explain the remarkably high prevalence of Jaunpuri (as shown in the confusion matrix in Figure 8).

The case of SCh in the concert video (which featured Rag Miyan Malhar only) is instructive: Miyan Malhar was very poorly classified when one solo take was used for testing and the other two for training, it was better for the duo videos (see Figure 8), but when all solo takes were used for training and concert clips for testing the success was very high. Given that the system failed to identify the singer in this video, it is possible that whatever caused this also affected the raga classification, and that the excellent raga identification was due to chance.

(3) Identifying ragas sung by one singer using a model trained by other singers' movements will be much more difficult, and results are likely to be no better than chance.

As expected, the classification accuracy was much lower in this case (split 2). However, five of the six models performed a little above random guess level (15.8–17.2%, as opposed to 11%).

The results do not tell us directly how the model achieves its classification accuracy, and this is not trivial to interpret. The t-distributed stochastic neighbour embedding (t-SNE) method (Maaten & Hinton, 2008) is introduced here to visualise how the machine learning system works. This is a powerful tool for compressing

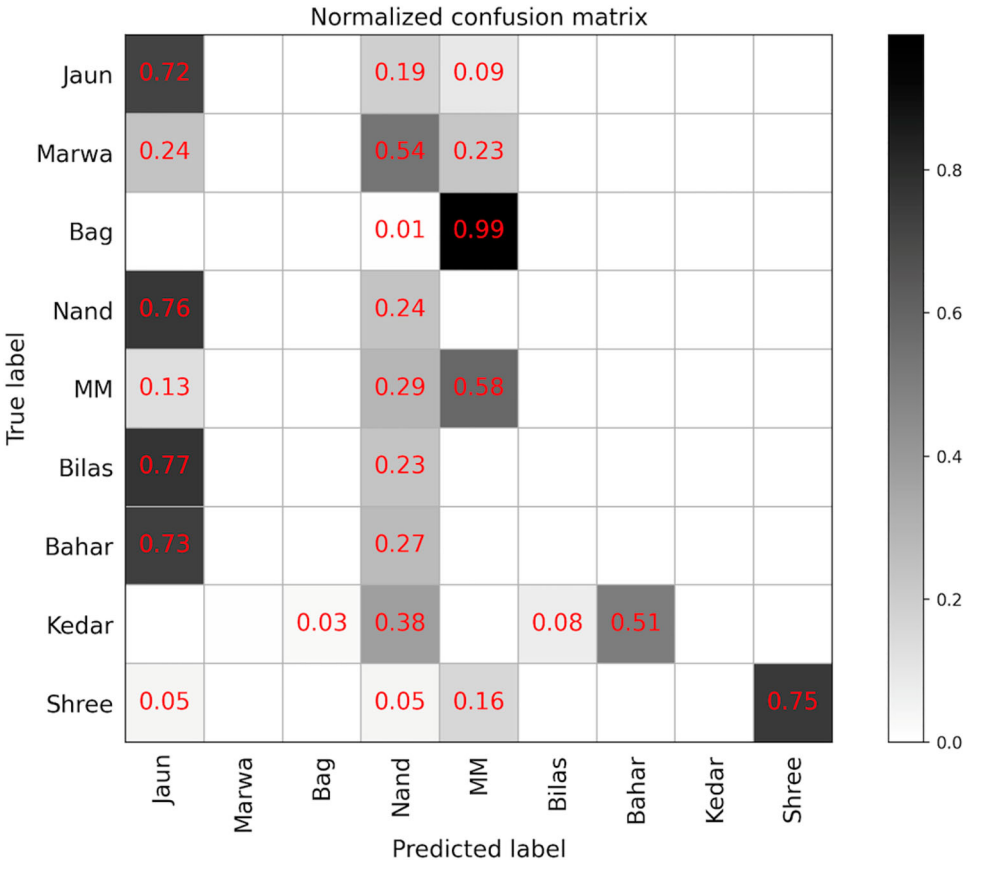

Figure 8. Confusion matrix of the duo videos.

the high dimensional features into a low (e.g. 2) dimensional space, where the data structure is retained as much as possible. In Figure 9, each point in the 2-dimensional space denotes the skeleton data for a short solo clip, and different ragas or musicians are represented in different colours. Machine learning methods train linear or non-linear classifiers that can divide inputs into classes. From the distributions in Figure 9, it is clear that there are large overlaps between different ragas, which are very difficult to discriminate. In contrast, clips of different musicians form obvious clusters, thus identifying the musicians is much easier than the ragas.

This exploration also allows us to evaluate which approach might be most effective for future studies of this nature: 2D or 3D pose, using the original MS-G3D or an early or late fusion 2-path model? Regarding the difference between 2D and 3D the results are somewhat mixed, but the 2D data gives better prediction accuracy in the majority of cases. This may indicate that the 2D data extraction is more reliable than the 3D, although we are not able to check that directly in the absence of a ground truth.

As for the usefulness of the velocity vector information, incorporating the velocity information in the early fusion model gives the best results for singer classification, For raga classification the picture is more mixed, but the early fusion model again performs better in the majority of cases. One possible explanation for this is that the number of parameters in the late fusion model is twice as that of the early fusion, thus the late fusion model is more like to cause overfitting. This means the parameters tend to converge to a local optimum that performs well only for the training set but cannot generalised to the test set.

The results reported here demonstrate that the use of pose estimation to gather movement information, together with action recognition models, can be productive in the analysis of musical performance. They confirm the findings of qualitative studies that Indian singers' gestures are idiosyncratic, but that singers show a degree of consistency in the ways they distinguish different ragas through their movement. The fact that classification accuracy varied very considerably between singer-raga pairings reflects the fact that

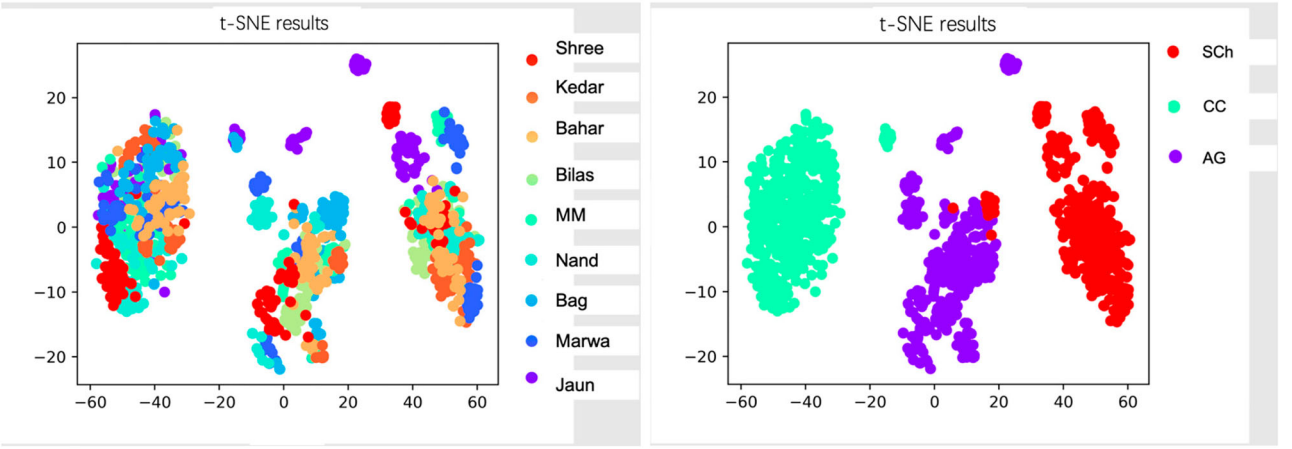

Figure 9. The visualisation of the features of different ragas (left) or musicians (right).

the approach taken here avoided using any prior knowledge about the different ragas, and did not attempt to incorporate human observation (for example, applying an observation that singers may raise their hands more quickly, or more directly, in a certain raga and trying to hand-craft a feature to capture this computationally).

Future studies would benefit from the use of even larger datasets to train the machine learning models. Given the impracticality of creating datasets an order of magnitude larger, a more realistic approach may be to use specially-crafted datasets such as our solo recordings to develop approaches that can then be generalised to publicly-available data such as YouTube videos. Future research may also start to build on our simple bottom-up approach to classification in different ways. One way to do so would be data-driven, using clustering techniques to identify common movement patterns. Another would be to use expert knowledge to identify features visually, and then train computational systems to recognise them. In these ways future research should be able to ask questions of video performance data that are driven by musicological, psychological and movement science concerns.

Acknowledgements

The authors would like to thank Apoorva Gokhale, Chiranjeeb Chakraborty, Sudokshina Chatterjee and Subrata Manna for their help and cooperation in making original recordings for this project, Simone Tarsitani for technical assistance and Laura Leante who helped to design the original solo dataset.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Natural Science Foundation of China: [Grant Number 62206162]; financed by the 'EnTimeMent' project funded by EU Horizon 2020 FET PROACTIVE: [Grant Number 824160]; and CCF-Tencent Open Fund: [Grant Number RAGR20220127].

Ethical approval

Approval for the recording and sharing of the solo dataset and new concert recordings was granted by the Durham University Music Department Ethics Committee on 17th February 2020 (application MUS-2020-02-03T13_53_06-fghk75). This was extended to cover the duo recordings and their use, approved by the same committee on 19th April 2021 (MUS-2021-03-31T09_31_25-fghk75).

ORCID

Martin Clayton http://orcid.org/0000-0002-9670-5077
Jin Li http://orcid.org/0000-0002-0260-3169

References

Al Ghamdi, M., Zhang, L., & Gotoh, Y. (2012). Spatio-temporal SIFT and its application to human action classification. In Fusiello A., Murino V., & Cucchiara R. (Eds.), *Computer Vision – ECCV 2012. Workshops and Demonstrations. ECCV 2012. Lecture Notes in Computer Science* (Vol. 7583, pp. 301–310). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33863-2_30

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Clarke, A., Weinzierl, M., & Li, J. (2021). Pose estimation for raga (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.5526676

Clayton, M. (2007a). Observing entrainment in music performance: Video-based observational analysis of Indian musicians' *tanpura* playing and beat marking. *Musicae Scientiae*, *11*(1), 27–59. https://doi.org/10.1177/102986490701100102

Clayton, M. (2007b). Time, gesture and attention in a Khyāl performance. *Asian Music*, *38*(2), 71–96. https://doi.org/10.1353/amu.2007.0032

Clayton, M., Jakubowski, K., & Eerola, T. (2019). Interpersonal entrainment in Indian instrumental music performance: Synchronization and movement coordination relate to tempo, dynamics, metrical and cadential structure. *Musicae Scientiae*, *23*(3), 304–331. https://doi.org/10.1177/1029864919844809

Clayton, M., Jakubowski, K., Eerola, T., Keller, P. E., Camurri, A., Volpe, G., & Alborno, P. (2020). Interpersonal entrainment in music performance: Theory, method and model. *Music Perception*, *38*(2), 136–194. https://doi.org/10.1525/mp.2020.38.2.136

Clayton, M., Leante, L., & Tarsitani, S. (2021a). North Indian raga performance. OSF. May 14. https://doi.org/10.17605/OSF.IO/NKJGZ

Clayton, M., Li, J., Clarke, A. R., Weinzierl, M., Leante, L., & Tarsitani, S. (2021b). Hindustani raga and singer classification using pose estimation. OSF. October 14. https://doi.org/10.17605/OSF.IO/T5BWA

Clayton, M., Rao, P., Shikarpur, N., Roychowdhury, S., & Li, J. (2022). Raga classification from vocal performances using multimodal analysis. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, Bengaluru, India. https://dap-lab.github.io/multimodal-raga-supplementary/

Dahl, S., Bevilacqua, F., Bresin, R., Clayton, M., Leante, L., Poggi, I., & Rasamimanana, N. (2009). Gestures in performance. In R. I. Godoy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 36–68). Routledge.

Fatone, G. A., Clayton, M., Leante, L., & Rahaim, M. (2011). Imagery, melody and gesture in cross-cultural perspective. In A. Gritten & E. King (Eds.), *New perspectives on music and gesture* (pp. 203–220). Ashgate.

Godoy, R. I., & Leman, M. (Eds.). (2009). *Musical gestures: Sound, movement, and meaning*. Routledge.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Gritten, A., & King, E. (Eds.). (2011). *New perspectives on music and gesture*. Ashgate.

Jakubowski, K., Eerola, T., Alborno, P., Volpe, G., Camurri, A., & Clayton, M. (2017). Extracting coarse body movements from video in music performance: A comparison of automated computer vision techniques with motion capture data. *Frontiers in Digital Humanities*, *4*, 9. https://doi.org/10.3389/fdigh.2017.00009

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Leante, L. (2009). The lotus and the king: Imagery, gesture and meaning in a Hindustani *Rāg*. *Ethnomusicology Forum*, *18*(2), 185–206. https://doi.org/10.1080/17411910903141874

Leante, L. (2013a). Gesture and imagery in music performance: Perspectives from North Indian classical music. In T. Shephard & A. Leonard (Eds.), *The Routledge companion to music and visual culture* (pp. 145–152). Routledge.

Leante, L. (2013b). Imagery, movement and listeners' construction of meaning in North Indian classical music. In M. Clayton, B. Dueck, & L. Leante (Eds.), *Experience and meaning in music performance* (pp. 161–187). Oxford University Press.

Leante, L. (2018). The cuckoo's song: Imagery and movement in monsoon ragas. In I. Rajamani, M. Pernau, & K. R. Butler Schofield (Eds.), *Monsoon feelings: A history of emotions in the rain* (pp. 255–290). Niyogi Books.

Li, M., Zhang, T., Chen, Y., & Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 661–670). https://doi.org/10.1145/2623330.2623612

Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, *68*, 346–362. https://doi.org/10.1016/j.patcog.2017.02.030

Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 143–152). https://openaccess.thecvf.com/content_CVPR_2020/papers/Liu_Disentangling_and_Unifying_Graph_Convolutions_for_Skeleton-Based_Action_Recognition_CVPR_2020_paper.pdf

Maaten, L., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. The University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., & Theobalt, C. (2018). Single-shot multi-person 3d pose estimation from monocular RGB. In *2018 International Conference on 3D Vision* (pp. 120–130). https://arxiv.org/abs/1712.03453v3

Moran, N. (2013). Social co-regulation and communication in North Indian duo performances. In M. Clayton, B. Dueck, & L. Leante (Eds.), *Experience and meaning in music performance* (pp. 64–94). Oxford University Press.

Paschalidou, S., & Clayton, M. (2015). Towards a sound-gesture analysis in Hindustani Dhrupad vocal music: Effort and raga space. In *International Conference on the Multimodal Experience of Music (ICMEM)*, Sheffield. https://www.researchgate.net/publication/312029966_Towards_a_sound-gesture_analysis_in_Hindustani_Dhrupad_vocal_music_effort_and_raga_space

Paschalidou, S., Eerola, T., & Clayton, M. (2016). Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical Indian singing. In *Proceedings of the 3rd International Symposium on Movement and Computing (MOCO '16)*, Association for Computing

Machinery, New York, NY, USA, Article 45 (pp. 1–2). https://doi.org/10.1145/2948910.2948914

Pearson, L. (2013). Gesture and the sonic event in Karnatak music. *Empirical Musicology Review*, *8*(1), 2–14. https://doi.org/10.18061/emr.v8i1.3918

Pearson, L., & Pouw, W. (2022). Gesture–vocal coupling in Karnatak music performance: A neuro-bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences*, *1515*(1), 219–236. https://doi.org/10.1111/nyas.14806

Potempski, F., Sabo, A., & Patterson, K. K. (2021). Technical note: Quantifying music-dance synchrony with the application of a deep learning-based 2D pose estimator. *bioRxiv* 2020.10.09.333617. https://doi.org/10.1101/2020.10.09.333617

Rahaim, M. (2012). *Musicking bodies: Gesture and voice in Hindustani music*. Wesleyan University Press.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *arXiv* Preprint. https://arxiv.org/abs/1406.2199

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Tao, Y., & Papadias, D. (2006). Maintaining sliding window skylines on data streams. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 377–391. https://doi.org/10.1109/TKDE.2006.48

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4489–4497). https://doi.org/10.1109/ICCV.2015.510

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, North America. April 2018. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2117–2126). https://arxiv.org/abs/1703.08274