RAGA CLASSIFICATION FROM VOCAL PERFORMANCES USING MULTIMODAL ANALYSIS

Martin Clayton[1] **Preeti Rao**[2] **Nithya Shikarpur**[2] **Sujoy Roychowdhury**[2] **Jin Li**[1]
[1]Department of Music, Durham University, United Kingdom
[2]Department of Electrical Engineering, Indian Institute of Technology Bombay, India
martin.clayton@durham.ac.uk, prao@ee.iitb.ac.in

ABSTRACT

Work on musical gesture and embodied cognition suggests a rich complementarity between audio and movement information in musical performance. Pose estimation algorithms now make it possible (in contrast to traditional Motion Capture) to collect rich movement information from unconstrained performances of indefinite length. Vocal performances of Indian art music offer the opportunity to carry out multimodal analysis using this information, combining musician's body movements (i.e. pose and gesture data) with audio features. In this work we investigate raga identification from 12 s excerpts from a dataset of 3 singers and 9 ragas using the combination of audio and visual representations that are each semantically salient on their own. While gesture based classification is relatively weak by itself, we show that combining latent representations from the pre-trained unimodal networks can surpass the already high performance obtained by audio features.

1. INTRODUCTION

In this article we explore the potential for multimodal analysis of Hindustani classical vocal performance. It is well known that Hindustani vocalists use a wide range of manual gestures to accompany their singing: the relationship between their hand movements and the acoustic content of their music has been compared to that between gesture and speech [1–5]. Empirical studies have related gesture to perceived effort and apparent manipulation of imagined objects by singers [6], and have demonstrated the increase of coordinated head movement between soloists and accompanists at cadential moments [7]. The sound-gesture relationship has also been explored in the related Carnatic (South Indian) music tradition [8, 9]. These studies have been based on both empirical analysis and observation of gestures, together with the ethnographic enquiry: empirical study of movement is made possible by various combinations of motion-capture and video-based tracking of individual body parts.

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

With raga serving as the melodic framework in Indian art music, raga characteristics have been extensively explored via melodic features computed from the predominant pitch contour extracted from performance audio [10, 11]. There has not been any work that has similarly employed the visual data of performances, let alone the combination. In their survey paper [12], Duan et al. reviewed past research categorised by type of instrument and analysis task; no work was found on audiovisual analysis for singing. The nearest instrument task they reviewed was automatic transcription aided by visual analysis such as hand and finger tracking, helping play/non-play detection and note onset localization [13].

In the case of motion capture, the difficulty of data collection, particularly in natural contexts, limits the scope of research. Capture of full-body position information from video, such as is now possible using pose estimation algorithms, significantly increases the potential scope of multimodal analysis, with the possibility of collecting movement data from natural performance contexts extending over long durations. This makes it possible to explore sound-movement relationships of many kinds. Many different aspects of hand movement have been linked to musical sound and structure: for example, the tapping or beating of hands against knees may indicate the tempo and serve to instruct and coordinate performers; continuous hand movements seem to match aspects of the flow and organisation of sung phrases; some gestures seem to indicate analogies with the manipulation of physical objects such as elastic bands, or to describe dimensions of the sound (an open hand shape is understood to be linked to an open-throated voice production; hands gradually moving apart match an increase in volume). Gesture is observed

Raga	Bag	Marwa	Bahar	Kedar	Shree	Nand	MM	Jaun	Bilas	**Sum**
Sum	9	10	8	10	9	8	11	10	11	**86**

Table 1. Number of pieces from each raga in our dataset.

to be idiosyncratic, and yet there may be common features of the sound-movement relationship: Leante argued that specific gestures such as the deliberate vertical raising of a hand linked melodic aspects to visual imagery in Shree Rag [2]. Each of these possibilities suggests different empirical multimodal studies. The study presented here investigates the possibility that manual gesture is sufficiently closely related to the melodic movement of Hindustani ragas (melodic modes) that movement data may be used to help predict the identity of the raga being sung. We achieve this with suitable deep learning models applied to pose data and, further, to the audio and combined audio and visual streams.

In the next section, we present the data set. This is followed by a description of the audio, visual and combined-modality methods explored in this work. Experiments that compare the distinct approaches in terms of raga prediction performance for two different training conditions are discussed next, followed by the presentation of the results and key insights.

2. DATASET AND PROCESSING

We exploit a dataset comprising solo recordings of a common set of 9 Hindustani ragas recorded by 3 Hindustani singers [14]. For each singer and raga, we have 2 distinct takes of alap singing (duration of a take ranges from 165-221 s, with a median duration of 187 s) for 55 recordings in all (for one combination we have only one take, and for two combinations we have three takes). An alap is the improvised opening section of a concert and introduces the raga. We also have a set of 31 pakad (catch phrases) recordings (duration ranging from 18-96 s with a median duration of 19 s). Table 1 shows the raga distribution of the 86 pieces across the 3 singers combined. The ragas, as listed in Table 2, offer a cross-section of raga features in aspects such as the mood or character with which they are associated (serious, joyful, etc.), typical speed and complexity of melodic movement, and predominant melodic range (i.e. favouring the upper or lower tetrachord). The singers, as captured in Figure 1 are all professional performers, two female (Apoorva Gokhale and Sudokshina Chatterjee, here abbreviated AG and SCh) and one male (Chiranjeeb Chakraborty, CC). The pose of the singer's upper body skeleton is estimated for each frame in the corpus using the OpenPose system [15] for skeleton extraction from the video at 25 frames per second. The selected 11 key-points are from the upper body as shown as video input in Figure 2.

Raga	**Scale**
Bageshree (Bag)	S R g m P D n
Bahar	S R g m P D n N
Bilaskhani Todi (Bilas)	S r g m P d n
Jaunpuri (Jaun)	S R g m P d n
Kedar	S R G m M P D N
Marwa	S r G M D N
Miyan ki Malhar (MM)	S R g m P D n N
Nand	S R G m M P D N
Shree	S r G M P d N

Table 2. The pitch sets employed by the nine ragas. Lower case letters refer to the lower (flatter) alternative and upper case to the higher (sharper) pitch in each case.

The recordings are split into clips of 12 s each with starting times separated by a randomly chosen value in the interval [0.8, 2.4] s. The 12 s duration was set in order to encompass the typical duration of vocal phrases in this music. We then investigate the task of raga identification for each clip – first with each modality separately and then with combining the audio and video modalities using different methods to test whether this can enhance overall classification accuracy. This task is attempted under two conditions. In the first (termed 'seen singer'), each singer contributes to both the training and test data, as explained later; in the second (termed 'unseen singer'), we attempt classification of one singer's clips based purely on training using the other singers' clips. Table 3 summarises the number of examples (clips) in the training and validation sets which are distributed almost equally across the 9 ragas for each singer.

3. METHODS

Figure 2 depicts our overall system for raga classification from the audiovisual data. As discussed next, each of the multiple pathways from the input data to the final prediction represent distinct approaches that differ in which modality is exploited or in how the two are brought together.

3.1 Audio Features

Melodic aspects that distinguish ragas include the tonal material, the hierarchy of notes and their sequences leading to characteristic phrases [10]. Phrases and motifs are recognised by their melodic shape which can be represented by the computed pitch contour. Melodic phrases cannot be regarded simply as sequences of the distinct pitch classes, since raga also helps determine features such as oscillations around certain pitches and distinctive pitch transitions (i.e. slides): thus, the pitch contours contain rich information. Our dataset of alap and pakad pieces comprises solo singing with drone. We apply source separation [16] followed by pitch and voicing detection at 10 ms intervals using short-time autocorrelation analysis

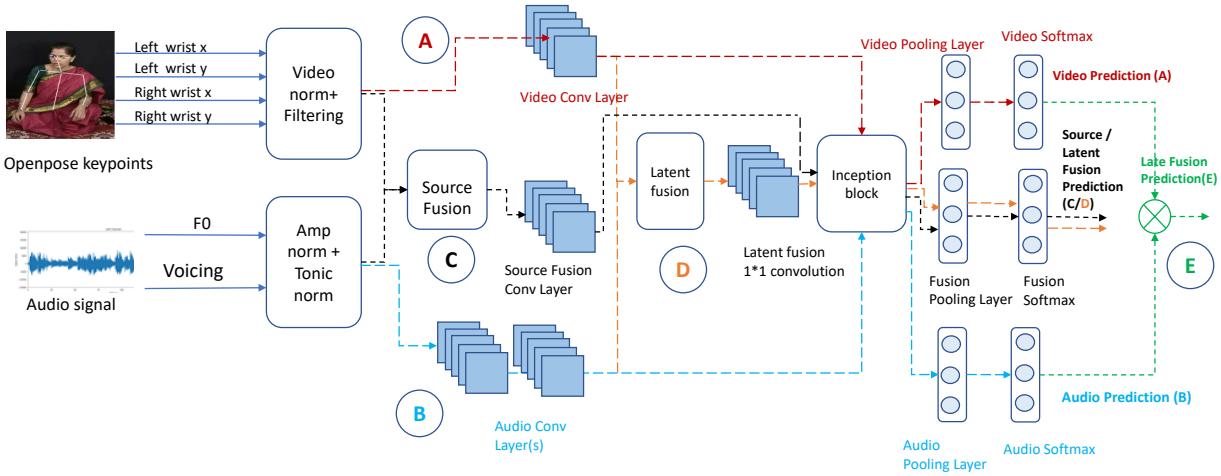

Figure 2. Proposed system for multimodal classification from pose and audio time series extracted from 12s video examples. This diagram shows 5 such configurations. A and B represent unimodal classification for video and audio streams respectively. C and D represent multimodal classification using data fusion at different layers. E represents model fusion using unimodal classification results.

[17, 18]. An important analysis parameter (that also dictates the window size) is the pitch search range, limiting which can help minimize octave errors. The tonic, automatically detected and manually verified [19, 20], was obtained for each performance and used to achieve the needed tonic-normalization of the extracted pitch contours. The detected tonic is also used to define the expected 2 octave pitch range for each piece. The voicing is a binary variable per frame that is set to zero in detected silence frames corresponding to singing pauses. Brief unvoiced regions (less than 500 ms) arising from short silences and consonant utterances are filled in via cubic spline interpolation to obtain the continuous pitch contours associated with melodic movements. The resulting time series from each clip thus comprises of 1200 samples (12s x 100/s).

3.2 Video Features

Real-time skeleton data, such as that obtained via OpenPose, can be handled by a graph model such as Graph CNN. We treat it instead as multivariate time-series data, each time series corresponding to one of the position coordinates of a tracked joint, similar to the processing of data from body-worn sensors in human activity recognition tasks. The 2D positions of the 11 keypoints of the upper body (eyes, nose, neck, shoulders, mid-hip, elbows and wrists) are recorded and used to obtain normalised (x,y) coordinates for each of the two wrists at the frame rate of 25/s. The position data for the singer's two wrists alone is retained for the analysis since (a) these points are more reliably estimated by OpenPose than others such as the elbows or shoulders, and (b) hand gestures most clearly relate to raga expression, and this data is most likely to contribute to raga classification. We thus have 4 time-series representing the x,y positions of each wrist. Any missing data are interpolated and each of the wrist position time series is low-pass filtered to remove any jitter. The length of each of the time series is 300 samples (12s x 25 fps).

3.3 Network Architectures and Hyperparameters

Given that our music audio and video time series data embed information at multiple time scales, we choose an inception network for its multiple kernel sized filters [21,22]. Inception networks have been previously used for multivariate time series classification [23, 24]. We empirically observed that preceding the inception block with two convolutional layers led to superior audio performances while a single convolutional layer worked best for the video input. The relatively high frame rate of audio also necessitated greater stride choices through the convolutional layers that helped reduce the time series dimension to 200 at the input to the inception block. The convolutional layers help in learning audio features which may be relevant for processing via the inception network which further has a range of receptive fields for the convolution. A similar approach to use prior convolution layers with inception blocks was adopted in the original work introducing the inception network for image classification [21]. We further exploit these prior convolutional layers learned for the individual unimodal (audio and video) channels in our latent representation fusion model described in the following section. The inception block is followed by pooling and softmax layers.

Figure 3 shows the architecture of the inception block. Overall, the hyperparameters tuned include the kernel size and number of filters in the convolution layers, the common kernel size, number of filters, pool size and pool type in the inception block and the pool type in the final pooling layer. Hyperparameter tuning was carried out with sweeping across the chosen ranges using Bayesian optimization [25, 26].

3.4 Multimodal Analysis

It is widely believed that integrating features from multiple modalities can lead to more robust classification due

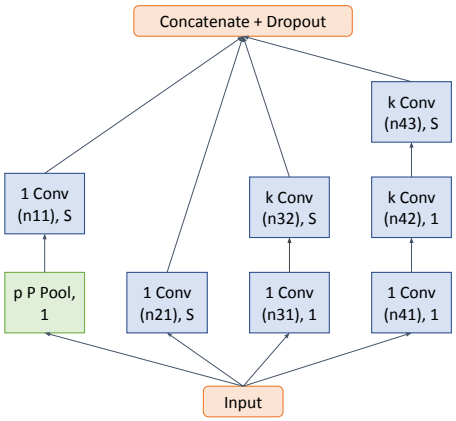

Figure 3. Structure of the inception block. 'k Conv (n), S' indicates a convolution layer with n k-sized kernels and a stride S. 'p' indicates the pooling size of the pool layer and 'P' the type of pooling. k, p, P and the number of filters were determined by hyperparameter tuning. S=1 for the video and combined models and S=2 for the audio model.

to potentially complementary information across the different modalities, all observing the same phenomenon. Multimodal classification has been an active research area in machine learning [27] where the precise approaches to combining information have been broadly categorised as early (feature-based), late (decision-based) and hybrid (their combination). While late fusion builds on combining the decisions of the individual unimodal predictors, early fusion can potentially exploit the low-level correlations between features across modalities. In our task, we have audio and video time series that actually co-vary as the singer makes continuous movements with her hands while singing. This relationship between melodic pitch variation and wrist movements makes it attractive to investigate the combination of the time series for classification.

In Figure 2, the flow-graphs labeled A and B depict the individual video and audio classification paths respectively. While the basic audio-only and video-only models differ only in the number of convolutional layers employed before the inception block, the hyperparameters assume values that are influenced by the time resolution of the corresponding time series with their different sampling rates. Early fusion is then obtained by first downsampling the pitch and voicing time series to the lower rate of 50/s and interpolating the wrist position data from its original 25/s to 50/s. The six time series form a 6-channel input to the convolutional layers thus realizing a form of source fusion. This is shown in position C in Figure 2 and a separate convolutional layer is used to learn the joint features before passing them onto the inception block.

Late fusion is achieved at the point labeled E (decision stage) in Figure 2 by combining the softmax outputs of the ensemble of the best model for each modality. We use soft voting [28] between the two classifiers by averaging the softmax outputs and then choosing the class with the maximum average as the predicted class. In addition, we learn

Data split	Seen split			Unseen split		
	AG	CC	SCh	AG	CC	SCh
Train	5590	5487	5588	4715	4105	4304
Validation	972	1075	974	1847	2457	2258

Table 3. Number of 12 s duration examples in each singer's train-validation data split

classifiers using the concatenated softmax outputs of the best models of each modality. This is a common approach in model stacking [29,30] and has been used in multimodal fusion earlier [27,31]. We try multiple models for stacking viz. Random Forests [32, 33] and logistic regression [34] and hyperparameter tune each of them using scikit-learn's GridSearchCV routine [35].

While decision fusion is the most straightforward approach to combining modalities, we also consider fusion at an intermediate stage given that the dynamic correspondence between the movement in the video and the aligned audio is expected to persist in the ealier convolutional layers. We achieve this by fusing the latent representations from each of the subnetworks where each has been optimised for a simpler task, viz. audio-only or video-only classification. Here we use the pre-trained weights of the convolutional layers of the best models of each modality and freeze the weights. This ensures that the features learnt for the individual modalities are maintained and the following inception block is tuned to amplify the interrelations between the two modalities. We call this novel method of fusion as latent fusion and depict it by D in Figure 2. Given that the input audio and video streams are at different effective temporal rates (lengths of 200 and 300 respectively) due to their original sampling and/or the different convolutional layer strides, we need to do a pooling on each individual modality convolutional output to bring the number of steps in the output sequence of the frozen models in sync. In addition, post the fusion we apply a convolution with filter size of 1 and learn the number of required filters via hyperparameter tuning thus letting the model learn the number of channels needed for the inception network.

4. EXPERIMENTS AND RESULTS

The training and validation sets are designed separately for the two tasks. We report experimental results for different model architectures (including one or both modalities) for each singer. In the seen singer task, we create training and validation data splits for each singer. Raga classification accuracies are reported for each method on the validation dataset of each singer and also averaged across the singers. For a given singer, the validation set comprises the set of examples from one of the singer's alap takes for each raga. The corresponding training dataset is then all the *remaining* pieces by the singer plus all the pieces of the other 2 singers. We have thus a validation and train set for each of the 3 singers. The similar exercise is carried out for the unseen singer task. Here, all the pieces by a given singer

are placed in the validation set and the pieces sung by the other two singers in the training set. Table 3 summarises the number of 12 s examples in each split as used in the experiments presented next.

For each task, we carry out model hyperparameter tuning on each individual singer's train-val dataset to obtain 3 sets of hyperparameters in all. These 3 hyperparameter-tuned models next have their weights recomputed on the training data of a given singer (say, AG) to obtain 3 trained models. Finally, the evaluation of Singer AG validation data is obtained via the ensemble of the three models. Similarly, we obtain the ensembled models for each of the singers CC and SCh and evaluate the methods on their respective validation datasets.

Our first set of experiments tests separately the performances of the audio modality and the video modality for the seen and unseen singer raga classification tasks. Table 4 presents the obtained validation accuracies. We observe that the unseen singer task is more challenging as expected. The audio based accuracies are significantly higher than those from video data on both tasks. The video based accuracies in the unseen singer task are close to chance (in the 9-way raga classification) indicating that the association between raga identity and gesture is highly singer dependent.

Our next set of experiments involve different approaches to multimodal classification. Given the non-informative nature of video cues in the unseen singer task, we restrict our attention here to the seen singer task. Table 5 presents the obtained validation accuracies across the different classification methods for each singer and the resulting mean across singers. We observe that early fusion is on the average at the performance level of the weaker modality. This is similar to the observations of Oramas [36] where the learning over very unequal modalities can be overwhelmed by either one. The results of late fusion appear in the final two rows and we find that they fall slightly short of the audio performances, all pointing to the challenge of actually realizing the benefits of the complementary information.

We look for another opportunity to combine audio and video information streams at the output of the convolutional layers. The latent representations at this stage are generated from convolutional layers that are frozen in pre-trained unimodal classification tasks. Rather than simple concatenation of the two representations, we use pooling to first align the representations along the time axis reducing both audio and video to a length of 100 from 200 and 300 respectively. We obtain a performance that is slightly, but consistently, superior to that from audio alone (D vs B in Table 5) for every one of the 3 singers, indicating that the combination of aligned latent representations successfully exploits the joint information. This is also borne out by the histogram summary provided in Figure 4.

5. DISCUSSION

As expected, prediction accuracy is significantly higher for the audio modality than the video, and higher in the 'seen singer' than the 'unseen singer' condition. Prediction accuracy is increased slightly when audio and video data are combined in comparison with audio data alone (from 82.8 to 85.1 %). It should be noted that some extracts score much more highly than others: SCh's Shree scores highly for prediction accuracy in both modalities; CC's MM is unusual in that video prediction seems more accurate than the audio; some extracts score very poorly on prediction from video only, such as AG's Kedar and SCh's Bahar. Some of the wrong predictions can probably be attributed to lack of data: for example, if the singer is silent for a significant part of the 12 second clip, or their hands do not move significantly. Other mismatches may be explained by features of either the melodic movement or the singers' gestures.

We present confusion matrices for the audio, video and the latent fusion bimodal classification in Figure 5. The video-only matrix has significant off-diagonal dispersion.

Data Split	Seen Singer		Unseen Singer	
	Audio	Video	Audio	Video
AG	92.1	36.3	76.9	14.3
CC	79.4	31.8	60.4	13.8
SCh	77.0	39.2	67.2	10.0

Table 4. Validation accuracy (%) of only audio and only video modalities on seen/unseen singer train-val splits.

Model Type	Model Name	AG	CC	SCh	Mean
A	Video	36.3	31.8	39.2	35.8
B	Audio	92.1	79.4	77.0	82.8
C	Source fusion	30.1	42.4	35.8	36.1
D	Latent fusion	**93.3**	**82.7**	**79.2**	**85.1**
E1	Equal voting	85.9	73.7	67.9	75.8
E2	Stacking classifier – RF	81.9	74.2	76.3	77.5

Table 5. Validation accuracy (%) from each singer's split for different model architectures in the seen singer task.

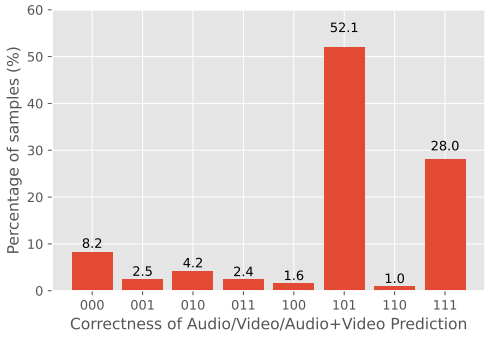

Figure 4. Histogram indicating percentage of the 3021 validation data examples predicted correctly (1) or incorrectly (0) by audio, video and the latent fusion based methods. For example, 011 indicates incorrect prediction by audio but correct predictions by video and bimodal classifiers.

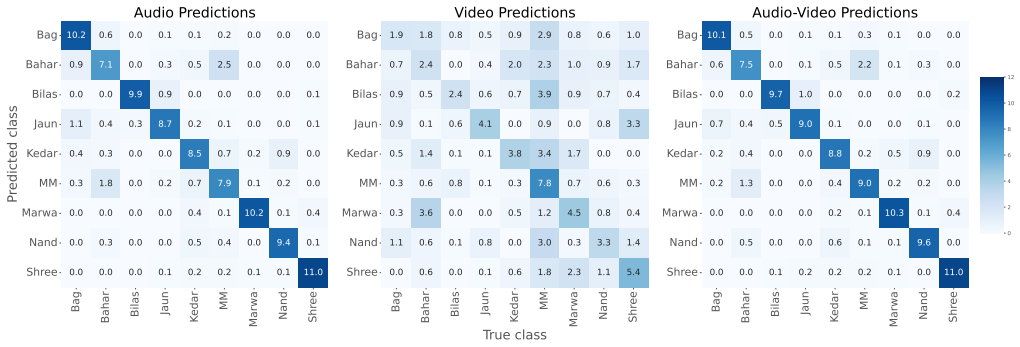

Figure 5. Confusion matrices of predictions made from audio, video and audio-video modalities. Numbers are represented in percentages of the total number of test examples across the three singers combined.

On the other hand, the multimodal matrix visibly improves upon the audio-only by further moving examples into the diagonal. The similar behaviour was noted in the individual singer confusion matrices. We also observe that the most common mis-predictions in the audio modality occur when the scale is the same or similar, and particularly where both pitch material and melodic movements are similar, as is the case between Miyan ki Malhar (MM) and Bahar. The main differences between these two closely-related ragas is that Bahar favours a higher pitch range and faster movement [37]: qualitative review of the prediction results shows that these factors are at play (e.g., in the faster portions of the extracts MM is more likely to be mis-classified as Bahar, at least for the two female singers). Other confusions are observed between Kedar and Nand (same scale), between Bageshree and Bahar (the latter uses one additional note), or between Jaunpuri and Bageshree (one note is different). Relatively few cases occur where the pitch material is very different (e.g. in SCh between MM and Nand or Kedar), which are harder to interpret. Some prediction errors seem to be related to the use of a low pitch (below Pa, the fifth degree, in the lower octave), as for example CC's Shree between 38-53s: this can be explained by the fact that the pitch extraction was constrained to a range of two octaves, with the low Pa as its limit. We find in this case clear instances of video and multimodal predictions doing well.

Confusion matrices for the video prediction are harder to read, given the lower accuracy overall, but some patterns can be observed. The Bahar/MM confusion is also present in the video domain (for the female singers), perhaps related to the fact that these two ragas share not only a scale but also many aspects of their melodic movement. Bahar can sometimes be confused with Nand and Marwa. The Bahar/Nand confusion is not surprising, since both are associated with lively, complex melodic movement and a joyful mood. The confusion between these two ragas and Marwa is less expected, since Marwa's mood contrasts strongly (being serious and often described as unsettled or restless). It may be that there is a similarity in the singers' hand movements between Bahar and Nand's liveliness and Marwa's restlessness. MM, Kedar and Bilaskhani seem to get confused when the singers make rounded, bimanual gestures, as when the hands seem to be moving round each other. These gestures are associated with specific melodic movements, an andolan (slow oscillation) on the Ga (3) in MM and distinctive crooked (vakra) pitch movements in the other two ragas. For SCh, several ragas are wrongly predicted as Shree when she makes a direct upward hand movement, which is often distinctive of this raga [2]. Such qualitative observations suggest that the video prediction system may be classifying the hand movements in meaningful ways even when the predictions are wrong, pointing to similarities between the ways each singer gestures in different ragas.

When we look at the proportion of clips correctly identified in at least one of the two modalities in Figure 4, the most common result is to be correctly classified from the audio but incorrectly classified from the video. The small percentage of clips for which the opposite is true (c. 6.6 %) suggests there is some scope for the video information to improve the prediction accuracy achieved through audio alone, although this seems like a difficult challenge as the audio prediction accuracy is already high. Even so, we note that the 6.6 % where the audio is incorrect but video correct, the multimodal condition actually recovers about a third of this. Further, we have 2.5 % of the total set of clips (75/3021) correctly predicted only in the multimodal condition (i.e. audio and video data are combined using latent fusion). This amounts to a quarter of all clips that were wrongly classified in both audio-only and video-only conditions. It is interesting to note that of the 75 clips correctly predicted only in the multimodal condition, the most common ragas represented were Bahar (19) and MM (14), which as noted above share the same scale and many melodic movements. This study suggests that the combination of coordinated audio and gesture features can improve the classification of ragas from short clips. This approach could be extended to other musicological investigations such as the musical expression of mood or character, melodic phrase segmentation, and interpersonal coordination.

Supplementary material: https://dap-lab.github.io/multimodal-raga-supplementary/.

6. REFERENCES

[1] M. Clayton, "Time, gesture and attention in a khyāl performance," *Asian Music*, vol. 38, no. 2, pp. 71–96, 2007.

[2] L. Leante, "The lotus and the king: Imagery, gesture and meaning in a hindustani rāg," *Ethnomusicology Forum*, vol. 18, no. 2, pp. 185–206, 2009.

[3] ——, "Gesture and imagery in music performance: Perspectives from north indian classical music," in *The Routledge Companion to Music and Visual Culture*, T. Shephard and A. Leonard, Eds. Routledge, 2013, pp. 145–152.

[4] ——, "The cuckoo's song : imagery and movement in monsoon ragas." in *Monsoon feelings : a history of emotions in the rain.*, I. Rajamani, M. Pernau, and K. R. B. Schofield, Eds. New Delhi: Niyogi Books, 2018.

[5] M. Rahaim, *Musicking Bodies: Gesture and Voice in Hindustani Music*. Wesleyan University Press, 2012.

[6] S. Paschalidou, T. Eerola, and M. Clayton, "Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical indian singing," in *Proc. of the 3rd Int. Symposium on Movement and Computing*, 2016.

[7] M. Clayton, K. Jakubowski, and T. Eerola, "Interpersonal entrainment in indian instrumental music performance: Synchronization and movement coordination relate to tempo, dynamics, metrical and cadential structure," *Musicae Scientiae*, vol. 23, no. 3, pp. 304–331, 2019.

[8] L. Pearson, "Gesture and the sonic event in karnatak music," *Empirical Musicology Review*, vol. 8, no. 1, pp. 2–14, 2013.

[9] ——, "Coarticulation and gesture: An analysis of melodic movement in south indian raga performance," *Music Analysis*, vol. 35, no. 3, pp. 280–313, 2016.

[10] G. Koduri, S. Gulati, P. Rao, and X. Serra, "Rāga recognition based on pitch distribution methods," *Journal of New Music Research*, vol. 41, no. 4, pp. 337–350, 2012.

[11] K. K. Ganguli and P. Rao, "On the distributional representation of ragas: Experiments with allied raga pairs," vol. 1, no. 1, pp. 79–95, 2018.

[12] Z. Duan, S. Essid, C. C. Liem, G. Richard, and G. Sharma, "Audiovisual analysis of music performances: Overview of an emerging field," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 63–73, 2019.

[13] A. Bazzica, J. C. van Gemert, C. C. S. Liem, and A. Hanjalic, "Vision-based detection of acoustic timed events: a case study on clarinet note onsets," 2017. [Online]. Available: https://arxiv.org/abs/1706.09556

[14] M. Clayton, J. Li, A. R. Clarke, M. Weinzierl, L. Leante, and S. Tarsitani, "Hindustani raga and singer classification using pose estimation," 2021. [Online]. Available: https://doi.org/10.17605/OSF.IO/T5BWA

[15] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[16] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, p. 2154, 2020.

[17] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," http://www.praat.org/, 2021, version 6.1.38, retrieved May 2022.

[19] J. Salamon, S. Gulati, and X. Serra, "A multipitch approach to tonic identification in indian classical music," in *Proc. of the 13th Int. Soc. for Music Information Retrieval Conference*, Porto, Portugal, 2012.

[20] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor *et al.*, "Essentia: an audio analysis library for music information retrieval," in *Proc. of the 14th Int. Soc. for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] C.-L. Liu, W.-H. Hsaio, and Y.-C. Tu, "Time series classification with multivariate convolutional neural network," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4788–4797, 2018.

[24] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[25] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, 2011.

[26] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[28] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[29] N. Hatami and R. Ebrahimpour, "Combining multiple classifiers: diversify with boosting and combining by stacking," *Int. Journal of Computer Science and Network Security*, vol. 7, no. 1, pp. 127–131, 2007.

[30] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, no. 3, pp. 255–273, 2004.

[31] J. H. Koo, S. W. Cho, N. R. Baek, M. C. Kim, and K. R. Park, "Cnn-based multimodal human recognition in surveillance environments," *Sensors*, vol. 18, no. 9, p. 3040, 2018.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152.

[34] R. E. Wright, "Logistic regression." 1995.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the Int. Soc. for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[37] S. Rao and W. van der Meer, "Miyan ki malhar," https://autrimncpa.wordpress.com/miyan-ki-malhar/, accessed: May-2022.