HUMAN POSE ESTIMATION FOR EXPRESSIVE MOVEMENT DESCRIPTORS IN VOCAL MUSICAL PERFORMANCES

Sujoy Roychowdhury **Preeti Rao** **Sharat Chandran**
{214077004, prao, sharat} @ iitb.ac.in
Indian Institute of Technology Bombay

ABSTRACT

Vocal concerts in Indian music are invariably associated with the performers' hand gesticulations that are believed to convey emotion, music semantics as well as the individual style of the performers. Video recordings, with one or more cameras, along with markerless human pose estimation algorithms can be employed to capture such movements, and thus potentially solve music information retrieval (MIR) queries. Nevertheless, off-the-shelf algorithms are built for the most part for upright human configurations contrasting with seated positions in Indian vocal concerts and the upper body movements in the context of performing music. Current state-of-the-art algorithms are black box neural network based and this calls for an investigation of the components of such algorithms. Key decisions involve the choice of one or more cameras, the choice of 2D or 3D features, and relevant parameters such as confidence thresholds in common machine learning methods. In this paper, we quantify the increase in the performance with three cameras on two music information retrieval tasks. We offer insights for single and multi-view processing of videos.

1. INTRODUCTION

Performances of vocal music in the Indian classical traditions involve the use of hand gestures that accompany the singing. We therefore wish to perform the automated analysis of performances with audiovisual recordings. One or more video cameras can be used to record musical performances. Our goal in this work is to explore different Human Pose Estimation (HPE) methods for the computational analysis of expressive movements of upper body limbs of the vocalist.

Markerless Human Pose Estimation algorithms constitute a novel technology that is available to investigate hand gestures by looking at important *keypoints* such as wrists and elbows. These algorithms are trained with artificial deep neural networks on whole body movements, and occasionally on music recordings. One important concern in the use of HPE in musical gesture studies is that the gestures are typically expressive movements and not routine motor movements such as walking, jumping, or performing yoga poses, the latter motor movements being the bulk of the training data used in the development of HPE algorithms. Indian classical music, in both Northern and Southern traditions, is particularly rich in the use of gestures invariably in a seated position.

Figure 1: We analyze seated vocalists with multiple cameras. We identify singers purely based on hand gestures, and predict stable notes.

Paschalidou [1] studied associations between sound and "effort" in gesture in *Dhrupad* performances using an optical motion capture system. Although she finds correspondences, generalizing to multiple singers was challenging. Pearson and Pouw [2] look at vocal-gesture coupling in Karnatak music performance; the Kinect camera and an older, machine learning technique is technique is used for obtaining keypoints. With the current deep-learning HPE technology, Clayton et. al. [3,4] use OpenPose-based wrist keypoints to classify *raga* and identify singers on a dataset of multimodal Hindustani music recordings. Nadkarni et al. [5] also use OpenPose to explore the correspondence between vocal singing and gestures.

However, to the best of our knowledge, there has been no work which uses multiple camera views for studying gesture and vocal correspondence in music. While it is natural to expect that more cameras help HPE, on a careful examination of prior work, we see that the process of estimating keypoints requires multiple design decisions. Several options present themselves in terms of camera position, the number of views, the keypoint detection method and its parameters, and finally, methods for combining information from multiple camera views.

1.1 Scope of this paper

In this paper, we choose 3 recent models for keypoint detection and 3 different HPE methods to obtain information from multiple views for the purposes of analyzing gestures. We restrict our study to wrist and elbow of both hands since they appear to be the ones most relevant when singers are seated. We consider two MIR tasks.

Stable Note Prediction from Gestures This problem was studied by Nadkarni et al. [5]. The authors define a stable note as a region of at least 250 ms duration across which the singer's pitch lies within a 25 *cent* interval of the mean intonation of the raga note.

Gesture-based Singer Identification In this task, the goal is to identify the singer purely from gestures, i.e., without accessing the audio stream, or the face. Rahaim [6] emphasises that gestures in music are not taught or rehearsed and therefore tend to be idiosyncratic. The proposed MIR task attempts to validate the hypothesis that it should be possible to identify the singer from the gesture using 12s randomly chosen snips from the video. Similar problems may be interesting in other MIR settings such as a western music conductor's motions when the face is not visible, or in dance and musical performances where the face is hopelessly masked. A gesture-based singer identification system can also be used to validate a digital avatar system that is attempting to realistically mimic singers.

1.2 Contributions

Although our work is focused on the two tasks mentioned above, we offer insights more generally useful in the HPE analysis of multiview recordings of music performances.

- Every HPE algorithm provides confidence scores. We suggest an approach to the choice of confidence thresholds for fair comparisons across algorithms.
- Multiple cameras lend themselves to 3D reconstruction, and indeed a single camera can also be used in recent state-of-the art methods to infer 3D. We suggest that decision fusing 2D information from multiple cameras can be almost as competitive as using 3D reconstruction.

In term of concrete results for the two problems we report the following:

1. By considering position coordinates and individual coordinates of velocity and acceleration as features, a systematic choice of confidence thresholds, and the best HPE method, we improve the performance of stable note prediction (from gestures) from $\sim$66% [5] to $\sim$83% (single camera).
2. We report the accuracy of the best performing HPE method for gesture-based singer identification (8-way) to be $\sim$93%.

It is to be noted that the two MIR tasks are solved using two different methods. The Stable Note problem uses the classical machine learning method of SVM. There is no artificial neural network here. On the other hand, the gesture-based singer identification problem uses a deep neural network with an inception block.

The rest of the paper is organized as follows. In Sec. 2 we look at different keypoint detection methods, and the reported performance. Sec. 3 describes the dataset used. Task agnostic comparisons of HPE methods is discussed in Sec. 4. Our two suggested methods of consuming information from multiple cameras is described in Sec. 5. The details of our experiments and the results are reported in Sec. 6. A summary appears in Sec. 7.

2. BACKGROUND

We first briefly describe three popular HPE methods, one [7] of which is *proprietary*. Later we describe the applications of these to areas in sports, and medicine in order to understand current understanding of their usage. We are not aware of the direct use of HPE for gesture understanding except the ones mentioned in the introduction.

2.1 Human Pose Estimation Techniques

The pose of a human in HPE methods results in a stick diagram (similar to Fig. 1) of important joints such as the shoulder, wrist, elbow, hip, knee and so on from images and videos. At the turn of the century, the joints, referred to as *keypoints* were obtained with markers placed on different parts of the body — however this can only be set up in controlled experimental settings and may also affect the natural movement of the subject. Subsequently specialized cameras such as the Kinect was employed using classical machine learning techniques. With the advent of deep learning (DL), nowadays standard RGB cameras may be employed for markerless pose estimation.

One of the first DL-based techniques is OpenPose [8] which can identify 25 keypoints in terms of pixel coordinates reported as (x, y). OpenPose is based on estimating confidence heatmaps for keypoints and part affinity fields (PAF) which are vector fields encoding the connection across limbs between different joints. Since their method estimates keypoints and parts directly from the image using a multi-stage Convolutional Neural Network (CNN) their method is called a bottom-up approach in the literature [9]. OpenPose is trained on the MP-II human dataset [10] and the COCO [11] datasets. The MP-II dataset, with 25K images in 20 activity categories like cycling, running, violin playing, etc., has both full body and seating position data. The COCO dataset has 200K annotated images of 17 body keypoints in both seated and full body positions.

Alternatively, approaches based on Mask R-CNN [12] perform semantic segmentation of the image to identify masks on people in the image. Detectron2 [13] uses this mask to identify 17 keypoints for the body parts. As this method uses an identified mask for the prediction of the keypoints, this is often referred to as a top-down keypoint estimation method. Detectron2 is trained on the COCO dataset [11].

Multiple calibrated cameras can use well understood geometric computer vision methods of the 90s for depth estimation from 2D keypoins, thus producing 3D coordinates (x,y,z). Aniposelib [14] is a library which implements the 3D reconstruction from multiple synchronized calibrated cameras. However, the 2D keypoints in Detectron2 can be extended to 3D by a different deep learning based model VideoPose3D [15]. Videopose3D uses two DL models – a temporal dilated convolution for estimating depth per person and a separate 3D trajectory model for the center of the body viz. center-hip coordinate. Depth is estimated as a distance with respect to the center-hip of the body. Videopose3D is trained on Human 3.6M [16] and Human Eva [17] datasets.

There is, however, a direct way of obtaining 3D provided the face of the image is visible. BlazePose [18] identifies thirty-three 3D keypoints from single-view images. This model is trained on a custom dataset consisting of 60K images and is used in the Mediapipe library [7].

2.2 HPE-based Applications

A number of studies (for example, [19]) in HPE have involved evaluation of the accuracy of the HPE models by comparing markerless pose estimation with marker based pose estimation and shown that the Mean Absolute Error to be less than 30mm on 80% of trials.

Markerless systems are evaluated in clinical settings by Zhang et al. [20] and Mroz et al. [21] where they compare Hyperpose [22] vs OpenPose and OpenPose vs BlazePose (Mediapipe) respectively. Zhang et al. [20] establishes that OpenPose is better than HyperPose using manual annotations and then compares OpenPose with BlazePose via Root Mean Squared Error (RMSE) and correlation metrics. Their findings are that while BlazePose is faster, OpenPose provides for more accurate results in their setting. A similar comparison study [23] between three models – OpenPose, BlazePose and AlphaPose looks at a multi-camera setting for estimating biomechanical parameters like Ground Reaction Force (GRF). They observe that the detection rate is dependent on the camera view and the model. Also, they observe the BlazePose has lower detection rate than the other models.

Mehdizadeh et al. [24] look at estimating gait variables comparing OpenPose, AlphaPose [25] and Detectron and do not find any differences between their correlation with gait variables. Since all of the HPE models estimate a confidence score, they choose a confidence threshold for each model independently and discard estimates with a lower confidence than threshold and interpolate the values linearly. They choose the confidence thresholds so that less than 10% of frames were interpolated as a result.

Evaluating athlete anterior cruciate ligament (ACL) injury risk in jumps is important for athletes and the studies by Blanchard et. al [26] and Roygaga et al. [27] look at this using a multi-view camera setting and OpenPose model for HPE. They train models to identify if the jump is erroneous on each view independently and also a fusion model combining the individual models. Their choice of a confidence threshold of 0.3 for OpenPose confidence is validated by an ablation study across different thresholds. Their results indicate that the task performance depends on the view and the type of error. They drop frames below threshold of 0.3 but do not interpolate dropped frames.

2.3 Synopsis

All of these studies bring us to some conclusions which motivate our research in MIR space. First, we are aware of a variety of techniques and approaches for HPE estimation and 3D reconstruction. Second, markerless pose estimation models have acceptable accuracy This is necessary in a musical setting since performers may not be comfortable using markers. Third, we realize the possible benefits of doing multi-view reconstruction in downstream tasks. We understand that performance on any downstream task depends on the view and model of choice. Finally, we are aware of the importance of the choice of thresholds in rejecting or retaining HPE estimates and how these are dependent on the model and view in question.

3. DATASET

We used the dataset from our earlier study [5]. The details of the dataset, data processing and links to download the data are available on github. [1] The dataset consists of 11 professional singers singing 2 *alaps* [2] each of 9 *ragas*. Recordings are captured by 3 synchronized cameras. However, we discovered that the recordings for 3 of the 11 singers were done with uncalibrated cameras and thus, since we are interested in 3D information, Anipose [14] cannot be used. Therefore we base our MIR tasks described in Sec. 1.1 on only the remaining 8 singers. We are left with 143 recordings with about 7 hours 10 mins of recording. These recordings are at 24 fps with a resolution of 1920×1080. The angle between the front and the right camera is approximately 55 degrees and the front and left camera is approximately 47 degrees. We refer to left and right camera based on the singer's point of view. Fig. 1 shows a sample of the singer in the three views.

4. TASK-AGNOSTIC COMPARISONS

We choose the three HPE models because they provide a mix of bottom-up (OpenPose), top-down (Detectron) single view 2D keypoint estimation as well as single view 3D keypoint estimation (Mediapipe). In addition, our reconstruction techniques involve both frame-wise geometric reconstruction via Anipose [14] as well as DL-based methods (Videopose3D [15]) which uses information from neighbouring frames. Thus our methods of estimation and reconstruction are relatively independent of each other.

4.1 Confidence Threshold

All HPE models provide a confidence score for each of the estimated keypoints and it is conventional to choose

[1] Dataset github

[2] Alap is the unmetered introduction in raga performances.

a threshold for the confidence score to ignore predictions with a lower confidence score. Various previous studies [5, 27–29] have used a 0.3 threshold for OpenPose. However, we find that this method is not based on the actual data distribution and also cannot be extended to other HPE models like Detectron and Mediapipe. Due to this, we approach the problem similar to [24]. In every frame, if any of the left and right wrist and elbow keypoints have a confidence score to be less than some value x then we remove the position coordinate in question and interpolate it from the available neighbouring frames. For each model-view combination we change the threshold from 0 to 1 in steps of 0.01 and, in line with [24], choose the threshold so that no more than 10% of frames are dropped in that model-view combination. The corresponding obtained confidence thresholds are given in Tab. 2 and used for all our experiments. Abbreviations used in this paper are in Tab. 1.

OpenPose - OP2	Detectron - DE2	Mediapipe - MP2
Aniposelib - AP3	Videopose3D - VP3	Mediapipe3D - MP3

Table 1: Abbreviations for the various HPE techniques.

View	OP2	DE2	MP3
Front	0.49	0.17	0.27
Left	0.20	0.10	0.01
Right	0.38	0.15	0.12

Table 2: Confidence values obtained when the maximum number of interpolated frames is 10%.

4.1.1 Observations

We observe that thresholds for the left view are lower in all the 3 HPE models and this indicates that the keypoints are predicted with lower confidence for this view. The obtained threshold is particularly low for Mediapipe. Also, we observe that thresholds for OpenPose are higher than the other models indicating that OpenPose predicts keypoints with a higher confidence score. If we use the previously reported threshold of 0.3 for OpenPose then we would have 3.67%, 15.49% and 7.23% of interpolated frames in front, left and right views respectively. However, as seen from Tab. 5, there is no particular advantage of using the previous reported confidence value of 0.3 with its performance lower than what we have in Tab. 4.

4.2 Correspondence between models

Given that the models are attempting to predict the same joints, we expect that that the predictions would be close to each other in the pixel coordinate system. To verify this we consider the Euclidean distance of the predicted keypoints between 2 models in a pair in every frame. We ignore frames where any of the four keypoints have a confidence less than the corresponding threshold as defined in Tab. 2. The results are presented in Fig. 2.

4.2.1 Observations

We observe that the three models correspond well to each other in the front view (noting that the dimensions of the frame to be 1920×1080). However, in other views, while OpenPose and Detectron predictions maintain pair-wise consistency, the same cannot be said for Mediapipe. We repeat these experiments by choosing thresholds corresponding to 5% and 15% interpolated frames and the same trends hold true. These trends also hold when we study the pair-wise correlations (instead of RMSE distances). This analysis, however, can only say that the models concur with each other in the front view but not so in the other views with least concurrence in the left view. We cannot conclude that one model is better than the other based on this analysis. A partial intuition for these results is that the left hand obscures the right hand in the left view, and most singers are right-handed.

5. MULTIPLE CAMERAS

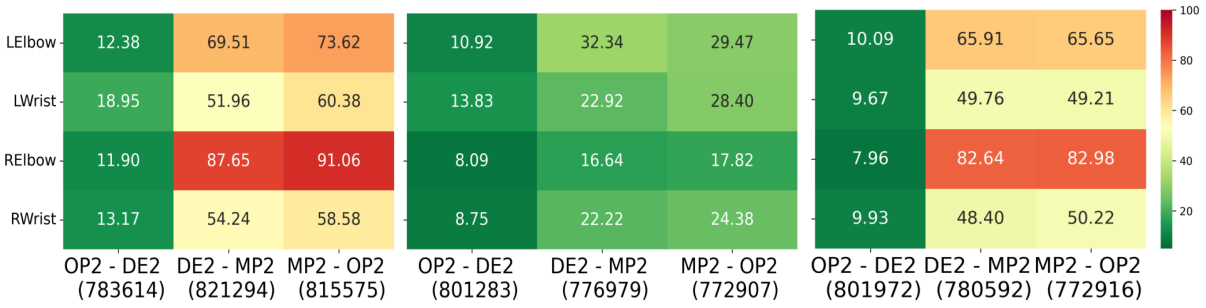
In this section, we provide the details of the use of multiple cameras for the downstream MIR tasks. Fig. 3 shows the algorithms we use to get 2D and 3D coordinates.

5.1 Reconstruction

Reconstruction involves the combination of the data from multiple views to estimate a depth-coordinate either via classical computer vision [14] or DL. The z-coordinate is measured in distance from the camera in geometric reconstruction. On the other hand, with DL the depth is estimated to be a distance with respect to the center hip with larger values indicating further distance from the center. In the recording setting, a higher value of the estimated z-coordinate (e.g., for an outstretched hand) would mean closer to the camera. Mediapipe which predicts the z-coordinate from a single view uses a similar definition of the z-coordinate. Reconstruction using both Aniposelib [14] and Videopose3D can be done using any of the cameras as reference view and information from other cameras used for reconstruction. The results for the downstream tasks can be different.

5.2 Model Fusion

The second method for consuming data from multiple views in a machine learning based MIR setting is to have the downstream task (e.g. classifiers) trained individually on each view and then use the probability predictions of these classifiers as an input to a further classifier. This can be done based on classifiers on three sets (each using 2D data) in which case this will be an alternative to reconstruction. On the other hand, one can use classifiers trained on three reference views using 3D reconstructed data (anipose, videopose3D) or predicted data (e.g. Mediapipe), and then use their probability outputs as an input to a further model. Both of these approaches are examples of multi-view fusion which exploit the complementary information present in different views.



(a) Left view (b) Front view (c) Right view

Figure 2: The average Euclidean distance in pixel coordinates different keypoints for pairwise HPE techniques. Non-interpolated frames considered are shown in parenthesis and four joints are considered. See Table 1 for the legend.

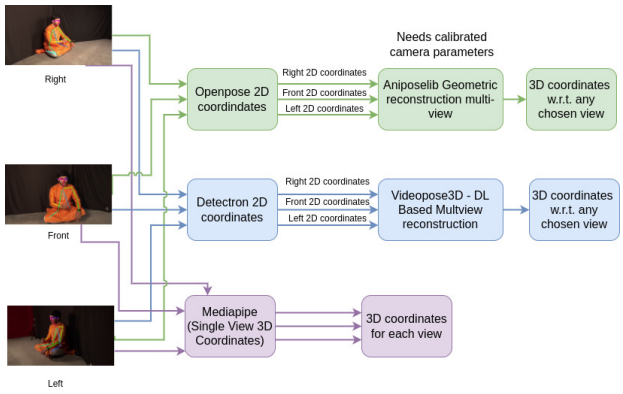


Figure 3: The three HPE methods (see Table 1) in this paper. *Left* and *Right* are views defined with respect to the singer.

6. EXPERIMENT DETAILS

6.1 Kinematic Features

The keypoint detection methods algorithms give us the x and y pixel coordinates for the keypoints and the 3D estimation gives us the z-coordinate to some scale. We linearly interpolate keypoints in frames that having confidence levels lower than the thresholds defined in Tab. 2. We use a low-pass Savitzky-Golay filter [30] to remove any jitter. We next interpolate the gesture data from the video frame rate to 10 ms sampling interval. We use z-score normalization for each keypoint by considering the mean and standard deviation for that keypoint and that coordinate across all frames of the recording. For repeatability, following [5], we estimate velocity and acceleration on each coordinate axis by a 101-point biphasic filter to get a smooth velocity and acceleration profile. We re-use the parameters of the biphasic filter defined in the supplementary link provided in [5].

6.2 Stable Note Detection

Although we replicate from [5] the stable note identification algorithm on audio, we use different features for the gesture classification. Instead of using the velocity and acceleration vector magnitudes, we consider the position, velocity and acceleration along each coordinate axis independently. We thus have 9 kinematic features per wrist. As in [5], we only consider stable and non-stable segments which are at least of 500 ms duration. Using this for the 8 singers, we have 15312 segments with 40.65% of them as stable notes. We use the mean and standard deviation per segment of each of the kinematic features for both wrists as input to our classifier. Thus, for models trained on 3D data (aniposelib, Videopose3d, Mediapipe) we have $9 \times 2 \times 2 = 36$ features considering both wrists. For models trained on 2D data (aniposelib, Videopose3d, Mediapipe2D) we have $6 \times 2 \times 2 = 24$ features. With these features, we train a SVM classifier per singer using 10-fold cross-validation and report the mean cross-validation accuracy. (Note that Mediapipe outputs 3D coordinates but we drop the z-coordinate in the experiments for comparison of 2D results.)

6.3 Gesture-based Singer Identification

We use randomly chosen 12s splits from the video in an attempt to identify the singer. We use a time series for the position, velocity and acceleration (PVA) features along each coordinate axis at (each) 10ms time interval for both wrists and elbows. Thus we have 36 features considering wrist and elbow for models using 3D data and 24 features for models using 2D data. We keep aside data for 3 *ragas* as test data (4103 samples), and train on the rest of the data using a random 80–20 train–validation split. We use a deep neural network consisting of convolutional layers followed by a 2D inception block as shown in Fig. 3 of [3]. We find best hyperparameters separately for each HPE technique, retrain the model with best hyperparameters and then report the results on the test set.

6.4 Fusion models

In the fusion models (for both 2D and 3D classifiers) for the stable note classifier we use the predicted probability for the stable note classifier. Thus we have 3 features in the fusion classifier and we train per singer using 10-fold cross-validation a classification model by a hyperparameter choice over logistic regression, random forests and support vector machines. We report the average of the mean per-singer cross-validation F1-score.

For the fusion models (both 2D and 3D) for the gesture-based identification of 8 singers, we take the softmax output of the final layer of the classifier from all views. Thus we have 24 features in the fusion classifier and we train a classifier by 10 fold hyperparameter tuning across logistic regression, random forests and support vector machines. Our training data for the fusion model consists of the softmax predictions on the train and val data of the neural network. We report the accuracy using features generated from the excerpts corresponding to the 3 held-out ragas.

6.5 Results

The results of the stable note detection appears at the first row of Tab. 3. We report the best result across camera views (or reference view for 3D reconstruction) for the HPE method. We note that the performance for the 2D models, MP2 performs better than either OP2 or DE2. Moving to 3D coordinates, we see a significant improvement in AP3 and VP3.

The results of the (pure) gesture-based singer identification classifier is given in the second row of Tab. 3. We report the best result across camera views (or reference view in the case of 3D). The classification accuracies, compared to a chance accuracy of 12.5%, indicate (given gestures are idiosyncratically singer-specific) that the HPE methods are reliable. We observe to our surprise that the method of classical computer vision (AP3) is the best performing model.

	2D			3D		
	OP2	DE2	MP2	AP3	VP3	MP3
StableNote	77.7	78.6	**83.0**	78.6*	82.5*	**83.5**
SingerID	**83.2**	81.9	79.6	**83.3**	81.4	82.9*

Table 3: F1-score (%) for stable note detection and accuracy (%) for gesture-based singer identification.star(*) indicates significant ($p<0.05$) difference between 2D and 3D, and bold indicates best result for a task in corresponding methods of 2D/3D.

The first row of Tab. 4 shows the results of decision fusion models based on the corresponding models across views for the stable note task. The results show that 2D fusion gives us comparable performance to reconstruction. The results of the models using fusion of classifiers across views are present in the second row of Tab. 4 for the gesture-based singer identification task. We see that when we use fusion instead of reconstruction, the results are much better with every possible technique for both MIR tasks. Accordingly we recommend this method. All fusion results are statistically significantly better than the corresponding best single view results in Tab. 3.

6.5.1 Ablation Study of Thresholds

Tab. 5 has the OpenPose results using a constant threshold of 0.3 for all views and the Anipsoselib result tasks. Tab. 6 has the results ablation study for various levels of interpolated frames.

The results show that our chosen threshold has comparable performance with default 0.3 threshold but our method is extensible to other HPE models. Results for 5%,10% and 20% interpolated frames are very similar. However if we set thresholds corresponding to 30% interpolated frames the performance is poorer.

	2D-Fusion			3D-Fusion		
	OP2	DE2	MP2	AP3	VP3	MP3
StableNote	82.0†	82.1	**83.9**	82.0	85.0*	**86.6***
Singer-ID	91.4†	**93.0**†	92.3†	93.3*	**93.6**	92.7

Table 4: Fusion based results. Values in %. Bold and star have same meaning as Tab. 3. Values with dagger (†) indicate the 2D-fusion model is better ($p< 0.05$) than the corresponding 3D model in Tab. 3

	OP2-Front	OP2-Left	OP2-Right	AP3
Stable Note	77.1	77.8	77.5	78.2
Singer ID	80.8	81.5	82.6	82.3

Table 5: Performance (in %) of OP2 for all views and AP3 using the confidence threshold of 0.3 used in the literature.

Interpolated %	2D Models			3D Models		
	OP2	DE2	MP2	AP3	VP3	MP3
5	78.2	78.6	83.0	78.0	82.6	83.0
10	77.7	78.6	83.0	78.1	82.5	83.0
20	77.1	78.5	82.9	77.4	82.2	82.9
30	75.1	74.8	80.5	75.4	79.0	80.6

Table 6: F1-score (%) across HPE techniques.

7. SUMMARY AND CONCLUSION

Given the importance of reliable joint pose estimation in gesture analysis, we investigated a set of distinct available approaches to the keypoint detection of wrists and elbows for an application of expressive hand movements in two MIR tasks. We showed that the different ways of using multiple camera views, in terms of the single-view pose estimation method and the manner of combining multiple views, can influence task performance significantly. While 3D reconstruction affords a complete description of the gesture movements, the fusion of multiple 2D information is competitive. The fusion of multiple 3D representations is seen to bring in further benefits. The superiority of fusion results over single view is established via statistical significance. The two MIR tasks involve the use of distinctly different machine learning methods (classical SVM, and recent deep-learning) and involve scenes where the action is only in the upper body, providing evidence for the use modern HPE methods. We expect the outcomes of this study therefore to be useful in any application of expressive movement analysis involving upper-body limbs.

Future work would involve the fine-tuning of HPE algorithms with a set of manually labelled keypoints to see whether the optimization with respect to upper body keypoints helps improve the estimates.

8. REFERENCES

[1] S. Paschalidou, "Effort inference and prediction by acoustic and movement descriptors in interactions with imaginary objects during dhrupad vocal improvisation," *Wearable Technologies*, vol. 3, p. e14, 2022.

[2] L. Pearson and W. Pouw, "Gesture–vocal coupling in karnatak music performance: A neuro–bodily distributed aesthetic entanglement," *Annals of the New York Academy of Sciences*, vol. 1515, no. 1, pp. 219–236, 2022.

[3] M. Clayton, P. Rao, N. Shikarpur, S. Roychowdhury, and J. Li, "Raga classification from vocal performances using multimodal analysis," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR, Bengaluru, India, pp. 283-290.*, 2022.

[4] M. Clayton, J. Li, A. Clarke, and M. Weinzierl, "Hindustani raga and singer classification using 2d and 3d pose estimation from video recordings," *Journal of New Music Research*, pp. 1–16, 2024.

[5] S. Nadkarni, S. Roychowdhury, P. Rao, and M. Clayton, "Exploring the correspondence of melodic contour with gesture in raga alap singing," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR, Milan, Italy*, 2023.

[6] M. J. Rahaim, *Gesture, melody, and the paramparic body in Hindustani vocal music*. University of California, Berkeley, 2009.

[7] G. Developers, "Mediapipe pose landmarker," 2020, accessed: 2024-03-09. [Online]. Available: https://developers.google.com/mediapipe/solutions/pose

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[9] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimedia Systems*, vol. 29, no. 1, pp. 167–195, 2023.

[10] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[14] P. Karashchuk, K. L. Rupp, E. S. Dickinson, S. Walling-Bell, E. Sanders, E. Azim, B. W. Brunton, and J. C. Tuthill, "Anipose: A toolkit for robust markerless 3d pose estimation," *Cell reports*, vol. 36, no. 13, 2021.

[15] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7753–7762.

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[17] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International journal of computer vision*, vol. 87, no. 1, pp. 4–27, 2010.

[18] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, USA*, 2020.

[19] N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukashiro, and S. Yoshioka, "Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras," *Frontiers in sports and active living*, vol. 2, p. 50, 2020.

[20] F. Zhang, P. Juneau, C. McGuirk, A. Tu, K. Cheung, N. Baddour, and E. Lemaire, "Comparison of openpose and hyperpose artificial intelligence models for analysis of hand-held smartphone videos," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2021, pp. 1–6.

[21] S. Mroz, N. Baddour, C. McGuirk, P. Juneau, A. Tu, K. Cheung, and E. Lemaire, "Comparing the quality of human pose estimation with blazepose or openpose," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*. IEEE, 2021, pp. 1–4.

[22] R. Ferens and Y. Keller, "Hyperpose: Camera pose localization using attention hypernetworks," *arXiv preprint arXiv:2303.02610*, 2023.

[23] M. Mundt, Z. Born, M. Goldacre, and J. Alderson, "Estimating ground reaction forces from two-dimensional pose data: a biomechanics-based comparison of alphapose, blazepose, and openpose," *Sensors*, vol. 23, no. 1, p. 78, 2022.

[24] S. Mehdizadeh, H. Nabavi, A. Sabo, T. Arora, A. Iaboni, and B. Taati, "Concurrent validity of human pose tracking in video for measuring gait parameters in older adults: a preliminary analysis with multiple trackers, viewing angles, and walking directions," *Journal of neuroengineering and rehabilitation*, vol. 18, pp. 1–16, 2021.

[25] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[26] N. Blanchard, K. Skinner, A. Kemp, W. Scheirer, and P. Flynn, "" keep me in, coach!": A computer vision perspective on assessing acl injury risk in female athletes," in *2019 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1366–1374.

[27] C. Roygaga, D. Patil, M. Boyle, W. Pickard, R. Reiser, A. Bharati, and N. Blanchard, "Ape-v: Athlete performance evaluation using video," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 691–700.

[28] J. Ripperda, L. Drijvers, and J. Holler, "Speeding up the detection of non-iconic and iconic gestures (spudnig): A toolkit for the automatic detection of hand movements and gestures in video data," *Behavior research methods*, vol. 52, no. 4, pp. 1783–1794, 2020.

[29] D. Pagnon, M. Domalain, and L. Reveret, "Pose2sim: An end-to-end workflow for 3d markerless sports kinematics—part 1: Robustness," *Sensors*, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/19/6530

[30] W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, pp. 669–672, 1990.