ON THE PRECISION OF MARKERLESS 3D SEMANTIC FEATURES: AN EXPERIMENTAL STUDY ON VIOLIN PLAYING

Matteo Moro[⋆†] *Maura Casadio*[⋆] *Leigh Ann Mrotek*[‡]
Rajiv Ranganathan[§] *Robert Scheidt*[‡] *Francesca Odone*[⋆†]

[⋆] Department of Informatics, Bioengineering, Robotics and Systems Engineering, (DIBRIS), University of Genova, Italy
[†] Machine Learning Genoa (MaLGa) Center, Genova, Italy
[‡] NeuroMotor Control Laboratory, Marquette University, Milwaukee, WI
[§] Department of Kinesiology, Michigan State University, East Lansing, MI

ABSTRACT

Human motion analysis is an essential task in several domains and, depending on the application field, it requires different level of accuracy. In the motor control field it is commonly performed with motion capture systems and infrared markers that guarantee a high accuracy. However, these systems are expensive, cumbersome, and may induce bias. An alternative to marker-based technologies are image-based marker-less systems, that are cheaper and do not affect the naturalness of the motion. Although their accuracy level seems to limit their use in motor control field, a thorough quantitative comparison with marker-based techniques does not appear to be available yet. We provide such an analysis, comparing the estimates of a 3D image-based marker-less pipeline we propose, with a standard marker-based system; the analysis is carried out on a multi-sensor dataset acquired to study the motion of violin players. The results we obtain on the precision level are suggesting that marker-less systems may successfully track performances in real-world settings.

Index Terms— Marker-less, 3D Reconstruction, Human Motion Analysis, Semantic Features Detection

1. INTRODUCTION

Measuring quantitative information about human motion is fundamental to understand how our central nervous system controls and organizes movements and is essential to many fields including motor control/learning and rehabilitation engineering [1]. Currently, for scientific use, the study of human motion in the rehabilitation and motor control fields is commonly done through marker-based techniques, motion capture systems and wearable sensors [2]. These methods are the gold standard because of their high accuracy (usually in the order of few millimeters) and for their reliability [2]. However, these technologies are expensive, time consuming and the use of markers or wires may severely affect the naturalness of the motion, especially in real-world scenarios [3]. Image-based marker-less techniques are an alternative to these methods [4]: they are less expensive, can be used to record performers in their natural settings and they do not affect the naturalness of the motion. Unfortunately, marker-less techniques have long thought to be less reliable and less precise [3]. Therefore, in a field where accuracy is essential, their use is limited. However, a systematic and detailed comparative analysis of recent image processing methods [2] has yet to be explored.

The long term goal of this project is the study of motor learning and motor re-learning: how people acquire motor skills and how these skills change with practice and experience. The specific aim of this paper is to implement an image-based multi-view marker-less pipeline to quantitatively study 3D human motion and compare it with a gold standard marker-based procedure in a complex task like playing a music instrument.

In order to evaluate the performance of the marker-less approach, we acquire the kinematics of 58 violinists repeatedly playing a G scale arpeggio. For the marker-less system, we rely on a 3-view camera system. The choice of three viewpoints allows us to geometrically reconstruct the 3D information while reducing the numbers of self-occlusions, which are quite frequent in moving human bodies. As a gold standard reference, we employ a motion capture system (Optotrak 3020) with active markers placed on the violin and on the bow. The synchronous recording allows us to validate the marker-less system with respect to the marker-based one.

The multi-view pipeline we propose includes three main steps: in the first one, in each frame of each acquired video, we detect pre-defined semantic features, chosen on the anatomic landmarks mostly involved during the playing

This work is supported by the US National Science Foundation (Grant 1823889). M. Moro is supported by Italian multiple sclerosis foundation - FISM – 2019/PR-single050. The authors thank Blake Brasch and staff of the Music Institute of Chicago for their kind assistance with data collection at the 2019 Summer Suzuki Institute. Thanks to erasmus+ K107 action.

session (shoulder and arm) and on the instrument. Semantic feature extraction is formulated as a semantic segmentation problem and it is carried out with an architecture based on Residual Neural Network [5]. Secondly, the $(x, y)_t$ coordinates of landmarks are filtered to enforce spatio-temporal consistency. Lastly, the 3D position of each landmark is reconstructed following a N-view geometry approach [6].

In the literature there are algorithms for the 3D reconstruction specific for the human pose. Some of them rely on single images [7], but do not appear to be appropriate for precise localization; others work under the hypothesis calibration is not available, and generally require very large datasets [8, 9]. Others approach the problem using a similar prior as calibrated reconstruction (multi-view calibrated inputs during training) but in a data-driven fashion [10]. The choice of a general purpose geometric approach is motivated by the simplicity and high generalization potential. Experiments will also speak in favour of its accuracy.

To compare our implemented pipeline with our gold standard, since the two systems are not mutually calibrated, we compare Euclidean distances between pairs of 3D landmarks, in the marker-based and the marker-less approach respectively. The distributions of distances show that the measures computed with our marker-less pipeline are very close to the one computed with the marker-based system (with an error on the pair-wise distances below 6 mm in at least 70% of the cases). Adopting marker-less methods in this application domain may provide significant benefits with respect to participant setup time and reduced invasiveness.

2. THE DATASET

The motions of 58 violinists of a wide range of ages and capabilities has been acquired during 5 days of the 2019 Summer Suzuki Institute organized by the Music Institute of Chicago. The violinists signed an informed consent form approved by the Marquette University Institutional review board.

The setup includes a multi-view camera system [3 RGB Mako G125 GigE cameras with Sony ICX445 CCD sensor, resolution 1292 X 964, 30 frames per second] and a motion capture system [Optotrak 3020, Northern Digital Inc., 100 Hz] with 6 active markers on the violin and 4 on the bow - see Figure 1. The RGB cameras have been calibrated in order to obtain intrinsic and extrinsic parameters. Each violinist was asked to sit on a chair, at a fixed distance from the acquisition sensors, and to perform 50 repetitions of a 13-note arpeggio (G-scale arpeggio). Apart from that, no other instructions were given to the violinists: their pose is variable and clothing differs across participants. They all played the same instrument. To reduce the acquisition time and limit the discomfort of the volunteers, no markers were attached to the players.

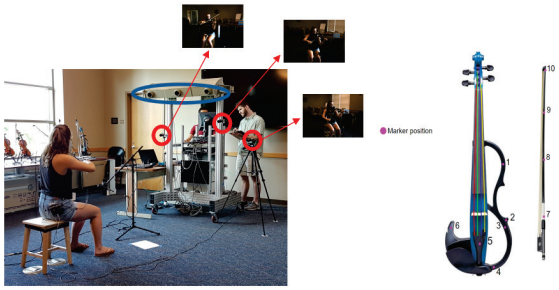

Fig. 1. Data acquisition setup (video cameras marked in red, motion capture system marked in blue) and markers location on violin and bow

3. PROPOSED PIPELINE

3.1. 2D Landmarks detection

The first step of the proposed pipeline is based on the detection of the positions in the image plane of some landmark points. In order to be able to compare the marker-less analysis with the marker-based one, we focus on the positions of the infrared markers on the violin and on the bow. Moreover, since the long term goal of the work is the analysis of the human motion, we consider also some human joints (the right shoulder, elbow, and wrist). To this purpose, we rely on a semantic-feature detection method [5], and we train it on labelled features extracted from the three image views. The architecture is a variant of Residual Deep Network (ResNet) pre-trained on ImageNet[11], and it allows for the extraction of semantic features of choice after an appropriate fine tuning. The choice of a semantic feature detector, instead of a classical human pose estimation algorithm [12, 13] is due to the fact we are not interested in the full-body pose, but we are instead interested in including semantic features that belong to objects (violin and bow).

To fine tune the network on our data, we consider 45 subjects and we randomly select 15 frames for each viewpoint (45 frames for each subject), then we manually label the position in the image plane of the landmarks: 4 markers on the bow, 5 markers on the violin (the 6th one is excluded because almost always occluded), 3 anatomic landmarks on the body - see Figure 2. The parameters used to train the network are the ones suggested in other applications, see [14].

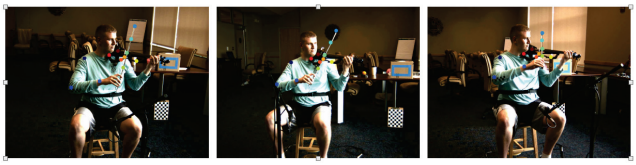

Fig. 2. Landmark points detected in the image plane.

Once the network is trained, for each test frame it provides a set of 2D landmarks $\mathbf{p}_i^V = (x_i^V, y_i^V, c_i^V)$, where $i \in \{$*shoulder, elbow, wrist, violin1, ..., violin5, bow7, ..., bow10*$\}$ characterises the semantic features, $V = \{1, 2, 3\}$ describes the view-point; (x_i^V, y_i^V) is the 2D landmark position on the image acquired from view V, c_i is a value in the interval $[0, 1]$ that quantifies the detection confidence. The latter is derived by an output layer of the model, representing probability score-maps of the semantic feature considered. The predicted position (x_i, y_i) is chosen as the pixel with the highest probability value.

3.2. 3D landmarks reconstruction

The semantic features extracted from the three viewpoints in each time instant, are combined to compute their corresponding points in the 3D space by means of multi-view geometric reconstruction. During calibration we estimate intrinsic matrices K_1, K_2, K_3 and extrinsic parameters between camera pairs [15], (R_{ij}, t_{ij}), $i, j = 1, 2, 3 \;\; i \neq j$. To synchronize the systems, we apply rotation averaging [6] that takes the relative rotations R_{ij} and computes the absolute rotations R_i in order to satisfy the compatibility constraint

$$R_{ij} * R_i = R_j.$$

In the presence of noise the problem can be solved through the minimization of:

$$\min_{R_1,\ldots,R_3} \sum_{(i,j)} ||R_{ij} - R_i * R_j^T||^2.$$

If the first view is chosen as reference, we have that $R_1 = I$. Similarly, it is possible to synchronize the translation vectors obtaining the absolute translations t_i starting from the t_{ij} and satisfying the compatibility constraint

$$t_{ij} = t_i - R_{ij} * t_j.$$

Once rotations and translations are synchronized [1], considering $\tilde{\mathbf{p}}_i^V$ the 2D landmarks expressed in mm ($\tilde{\mathbf{p}}_i^V = K_V \mathbf{p}_i^V$), then for each corresponding triplet $(\tilde{\mathbf{p}}_i^1 \; \tilde{\mathbf{p}}_i^2 \; \tilde{\mathbf{p}}_i^3)$ we apply a linear triangulation algorithm followed by a non-linear refinement based on the Gauss-Newton method [16], obtaining $\mathbf{P}_i$ in the 3D space. Figure 3 shows examples of the reconstructed $\mathbf{P}_i$ landmarks.

4. EXPERIMENTS

4.1. Landmarks detection evaluation

Firstly, we process all 58 videos acquired through our trained model. In order to evaluate the quality of the detection of each landmark in the image plane, we analyze the confidence number c_i returned by the model. Both for training and test subjects we count for each landmark the number of frames where confidence is lower than 0.75; in this way we are identifying the number of times that we can not trust the detection. Figure 4 shows the percentage - with respect the total number of frames for each video - of cases detected with $c < 0.75$. Occlusions are included in this analysis.

[1] Our rotation and translation synchronization is based on http://www.diegm.uniud.it/fusiello/demo/toolkit/

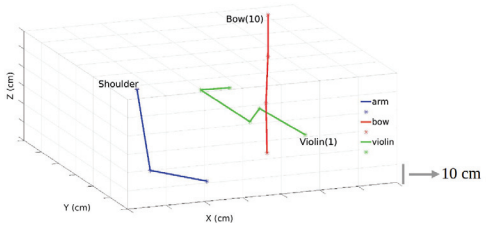

Fig. 3. The reconstructed landmarks: the right arm in blue (shoulder, elbow and wrist), the 4 markers on the bow in red and the 5 ones on the violin in green.

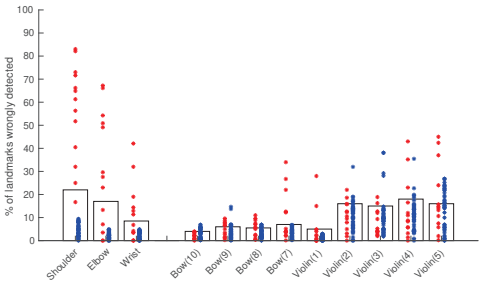

Fig. 4. % of frames (y axis) in which the landmarks (x axis) are detected with a confidence < 0.75. In blue we show the results for the training subjects (45), in red the test ones (13). These results show that the the number of points detected with $c < 0.75$ in the violin and in the bow is balanced in training and test subjects.

As we can see from Figure 4 the % of frames with points in the bow and in the violin detected with low confidence is balanced in test and training videos; these cases are mainly due to occlusions that can occur during the performance depending on the pose of the violinist with respect to the violin itself. Different considerations can be done for shoulder, elbow and wrist where the % of cases detected with low confidence is higher in test subjects. This is mainly due to the high variability of body landmarks, as confirmed in Figure 5: the figure compares the appearance variability of the *shoulder* landmark with *violin1*. The higher variability of shoulder, mainly due to different clothes worn by the volunteers, is apparent. Because of that, we may conclude body landmark detection would need to be trained on a larger dataset [17]. These points are not considered in our comparative analysis, as we do not possess a 3D gold standard for them.

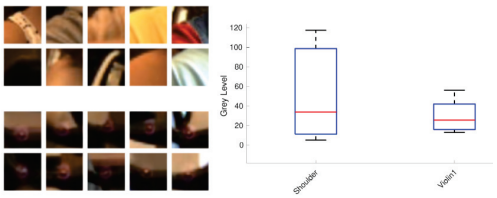

Fig. 5. Left: examples of textures for *shoulder* (top) and *violin 1* (bottom). Right: average grey level variability.

4.2. Marker vs marker-less performance comparison

We now evaluate the precision of the reconstructed 3D landmarks. Since we do not posses the relative position between the cameras and the motion capture reference systems, we compare Euclidean distances in the 3D space between pairs of landmarks estimated by the marker-based method and the marker-less one. Let dM_t^j be the distances computed with the marker-based system for each t-th frame and for each j-th pair of markers, with $j = \{violin1 - violin2, violin2 - violin3, violin3 - violin4, violin4 - violin5\}$ as numbered in Figure 1. dML_t^j are the corresponding marker-less distances. We then evaluate the difference between the measures computed with the two techniques: $(dM_t^j - dML_t^j)$. A difference close to 0 mm means that our marker-less measure is very close to the gold standard. In Figure 6 we report the errors for 4 different pairs of points. As we can notice the majority of the samples has a very small difference. The distributions of the errors are approx Gaussian centered in 0 and with a mean standard deviation of 6 mm.

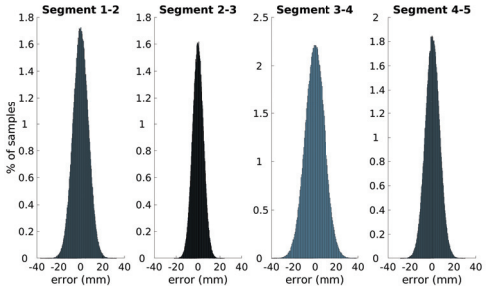

Fig. 6. Difference in mm (x axis) between Euclidean distances computed with marker-based signal (gold standard) and the geometric 3D reconstructed marker-less one. The 4 plots refer to 4 different distances between pairs of markers on the violin numbered as in Figure 1. The results show that the errors distribution are approx Gaussian centered in 0 mm and with a mean standard deviation around 6 mm, meaning that the error is very low for the majority of cases.

4.3. Marker-less 3D reconstruction comparative analysis

As a final evaluation of the 3D reconstruction algorithm adopted, we compare its accuracy with a recent alternative [10]. This method is a self-supervised learning method for 3D human pose estimation, which does not need any 3D ground-truth and makes use of multiple viewpoints and epipolar geometry. Figure 7 reports a consistently larger error with respect to the geometric approach. This is confirmed by Table 1, where we report mean and standard deviation for both techniques with respect to the gold standard.

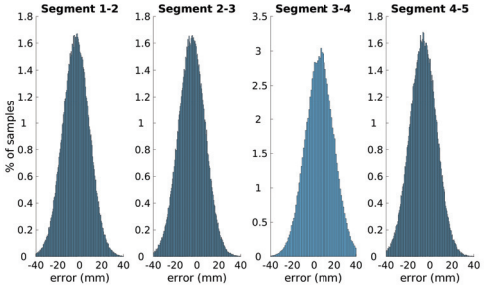

Fig. 7. Difference in mm (x axis) between Euclidean distances computed with marker-based signal (gold standard) and the CNN-based 3D reconstructed marker-less one. A comparison with Fig. 6 shows that the error with [10] is significantly larger.

	1-2	**2-3**	**3-4**	**4-5**
Geom	0.7 ± 5.7	0.6 ± 4.9	1.9 ± 8.3	0.8 ± 5.5
CNN	3.5 ± 9.1	4.8 ± 9.3	9.1 ± 12.2	5.3 ± 8.9

Table 1. Absolute value of mean $\pm$ standard deviation **in mm** of the error reported in Figure 6 and 7 for geometrical (Geom) and CNN-based (CNN) 3D reconstruction. The pairs of markers are numbered as shown in Figure 1.

5. CONCLUSION

In this work we proposed a novel multi-view image-based marker-less pipeline that can be adopted to study human motion avoiding the use of expensive and intrusive marker-based motion capture systems. The pipeline is organized in three steps: 2D landmarks detection, temporal filtering, and 3D reconstruction. We evaluate the accuracy of the pipeline on a dataset of violin players synchronously acquired with both a 3-view cameras system and a motion capture system. The results show that the error that we have by adopting the implemented pipeline is in the order of few millimeters. This open the possibility of adopting video-based marker-less systems also in application field, like motor learning, where a high level of precision is required.

6. REFERENCES

[1] Jennifer L McGinley, Richard Baker, Rory Wolfe, and Meg E Morris, "The reliability of three-dimensional kinematic gait measurements: a systematic review," *Gait & posture*, vol. 29, no. 3, pp. 360–369, 2009.

[2] Matteo Zago, Matteo Luzzago, Tommaso Marangoni, Mariolino De Cecco, Marco Tarabini, and Manuela Galli, "3d tracking of human motion using visual skeletonization and stereoscopic vision," *Frontiers in Bioengineering and Biotechnology*, vol. 8, pp. 181, 2020.

[3] Bruce Carse, Barry Meadows, Roy Bowers, and Philip Rowe, "Affordable clinical gait analysis: An assessment of the marker tracking accuracy of a new low-cost optical 3d motion analysis system," *Physiotherapy*, vol. 99, no. 4, pp. 347–351, 2013.

[4] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports medicine-open*, vol. 4, no. 1, pp. 24, 2018.

[5] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281, 2018.

[6] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li, "Rotation averaging," *International journal of computer vision*, vol. 103, no. 3, pp. 267–305, 2013.

[7] Denis Tome, Chris Russell, and Lourdes Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2500–2509.

[8] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Thompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt, "Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 3, pp. 501–514, 2016.

[9] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson, "3d pictorial structures for multiple view articulated pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3618–3625.

[10] Muhammed Kocabas, Salih Karagoz, and Emre Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1077–1086.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser 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.

[13] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[14] Matteo Moro, Giorgia Marchesi, Francesca Odone, and Maura Casadio, "Markerless gait analysis in stroke survivors based on computer vision and deep learning: a pilot study," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 2097–2104.

[15] Zhengyou Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[16] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[17] Luis Perez and Jason Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.