Towards a Hand Skeletal Model for Depth Images Applied to Capture Music-like Finger Gestures

[1]Arnaud Dapogny, [1]Raoul de Charette, [1]Sotiris Manitsaris*,
[1]Fabien Moutarde, [1,2]Alina Glushkova
[1]Robotics Lab, Department of Mathématiques et Systèmes, Mines ParisTech, France
[2]Multimedia Technology and Computer Graphics Lab, Department of Applied Informatics, University of Macedonia, Greece
{arnaud.dapogny, raoul_de_charette, sotiris.manitsaris, fabien.moutarde}@mines-paristech.fr
alina.glushkova@uom.edu.gr

Abstract. The Intangible Cultural Heritage (ICH) implies gestural knowledge and skills in performing arts, such as music, and its preservation and transmission is a worldwide challenge according to UNESCO. This paper presents an ongoing research that aims at the development of a computer vision methodology for the recognition of music-like complex hand and finger gestures performed in space. This methodology can contribute both to the analysis of classical music playing schools, such as the European and the Russian, and to the finger gesture control of sound as a new interface for musical expression. An implementation of a generic method for building body subpart classification model applied in musical gestures is presented. A robust classification model from a reduced training dataset, as well as a method for spatial aggregation of the classification results, which provides a confidence measure on each hand subpart location is developed. A 80% pixel-wise classification accuracy and 95% ponctual subpart location accuracy is achieved when musical finger gestures with a semi-closed hand are performed in front of the camera and the rotation around camera axis is not too important.

Keywords: Hand pose, skeletal model, random decision forests, computer vision, machine learning, depth map, finger musical gestures

1 Introduction

Hand and finger gestures have always played an important role in human artistic expression. The human know-how behind this expression constitutes the ICH that should be preserved and transmitted with the contribution of the “i-Treasures” research project. According to the objectives of “i-Treasures” a novel Multimodal Human-Machine Interface for the artistic expression and more precisely for the contemporary music composition should be developed, where natural hand and finger gestures performed should be mapped in real time to sounds. There are significant challenges to address but the first obvious one is to precisely capture and recognize finger gestures. This has been at the heart of many previous researches though often the literature focuses on the most trivial cases where the hand is waved in the air in front of the camera in the open-palm position. In this article we show that an existing model may be applied to more complex hand poses implying a possible use for our future music-like finger gesture recognition. This study and the model produced should contribute to the analysis and identification of the characteristics of different pianistic schools (Russian and European) applied in music playing.

2 State of the Art

Recent research tendencies show an increasing interest in identification and recognition of gestures with the use of different type of motion capture technologies: Wireless motion sensor-based, Marker-based, and marker-less technologies.

Various types of wireless motion sensors [4][5][6] or commercial interfaces, such as the Wii joystick [7], or the IGS-190 inertial motion capture suits from Animazoo, can provide real-time access to motion information. Usually, they are used for the recognition of gestures performed in space or on tangible objects and the provide a rotation representation of the motion.

Marker-based systems are based on optical-markers technology, such as Vicon Peak or Optitrack. In [1][2] for instance, the Vicon system was used to capture the motion of violin players. The aim of this research was the modelling of music performances by understanding different bowing strategies in violin playing. In another case, researchers tried to adapt this method on piano players [3], which has contributed to an off-line analysis.

Marker-less systems do not require subjects to wear special equipment for tracking and are usually based on passive computer vision approaches. In [8], recognition of the musical effect of the guitarist's finger motions on discrete time events is proposed, using static finger gesture recognition based on the "EyesWeb" computer vision webplatform. This approach cannot easily be applied in live performances. In the iMuse project [9], motion of the pianist's hands is used to "follow the music score", which means to synchronize his performance to the music score. A camera was mounted above the piano keys, however, the hand of the musician is globally analyzed, not in a finger level, and consequently finger gestures are not recognized. Another methodology for complex finger musical gesture recognition has been implemented in the PianOrasis system, which is based on marker-less computer vision image analysis techniques to detect and identify the fingertips and the centroid of the hand on a RGB video [10].

As an extension of [10], we present in this short article a hand classification model that permits the recognition/location of different hand subparts while executing music-like finger gestures with the use of a single time-of-flight depth camera (PMD CamBoard Nano).

3 Hand Skeletal Model

3.1 Overview

Several methods exist to retrieve body or hand subparts position either from RGB or from depth maps. The commonly used framework is to fit a skeleton model so that it matches observable features and to apply inverse kinematics to refine the skeleton [14-15]. A major reference in this field is the Kinect body retrieval algorithm described by Shotton et al. [11] where Random Decision Forests (RDF) are trained to perform pixel-wise body classification. This approach has been proven robust though initially restricted to entire body gesture due to sensor limitations. Lately, Keskin et al. [12] investigated the same approach this time on the hand skeleton and exhibit very promising performance. Still, the experimentation was limited to a proof-ofconcept that recognizes American Sign Language digits. Such application is far from our use case since the hand is usually facing the camera in the classical open-palm position and the simultaneous finger gesture recognition has not been evaluated.

Though also derivated from Shotton et al. [11], our development investigates the complex use case where the hand is executing music-like finger gestures that are of higher order of complexity. At first, we built a three-levels hand model with 12 labels that encompasses the hand base (palm and wrist) as well as fingers and fingertips as depicted in Fig. 1. We believe that this model is complex enough to analyse fine hand configurations and at the same time simpler than the 19 labels model used in [12], thus providing less classification errors. Fig 1. shows pixel-wise classification obtained using the discussed method.

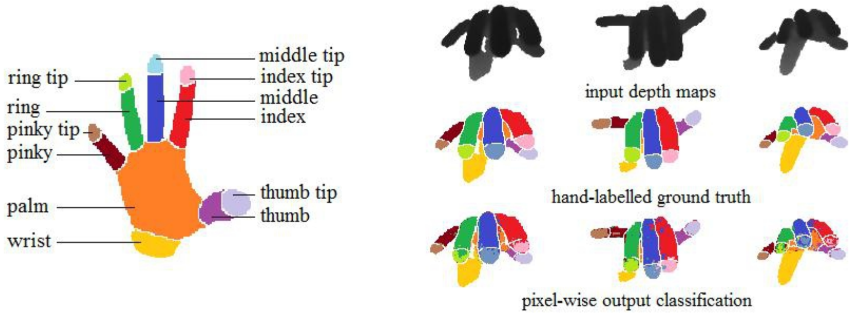


Fig.1. Hand model used with 12 subparts, depth maps, corresponding ground-truth and pixel-wise classification.

3.2 Training with Random Decision Forests (RDF)

As suggested in [11-12], we train Random Decision Forests (RDF) so as to perform pixel-wise classification. RDF [15] is an improvement of the Decision Trees machine learning approach where a complex problem is split in simple decisions to take that are often depicted as the nodes of a tree (the leaves being the final decision). For each tree of an RDF a subset of pixels *x* from images of the training database are used to train the tree. To limit the processing and memory cost, we use up to 3 decision trees with a maximal depth of 20. Then, for each node, we randomly generate 2000 weak classifiers and for each of these 50 candidate thresholds. The weak classifier we use (i.e. feature) compares the depth offsets in a specific neighbour and is similar to [11] as it was proved to be accurate and fast enough. Feature at pixel *x* of depth image *I* is thus computed as a difference of depth levels for offsets *u* and *v* normalised w.r.t. depth at *x,* as it is presented in equation 1.

$$f_{u,v}(I,x) = I(x + \frac{u}{I(x)}) - I(x + \frac{v}{I(x)}) \qquad (1)$$

We then simulate data partition at this node by thresholding this feature response using each candidate threshold and compute the entropy subsequent to the partition of the dataset. Finally, we keep the combination of feature and threshold that maximizes the information gain, which denotes the difference between current node information entropy and the sum of entropy of the sub trees resulting from the data partition. Data partition and weak classifier selection are then computed recursively on the left and right subtrees in a prefix order, until either maximum depth is reached or the information gain goes under a fixed threshold. We eventually store in tree leaves the probability distributions of the different hand subparts. Those distributions can be re-computed afterwards using all the images from the training database which, according to [12], allows slightly higher recognition rates.

In our case, this training stage is performed upon 500 hand-labelled images recorded with our depth camera. This database gathers input from 5 different persons simulating music-like hand and finger gestures in front of the camera such as the “arpeggio”gesture, as depicted in Fig.2.

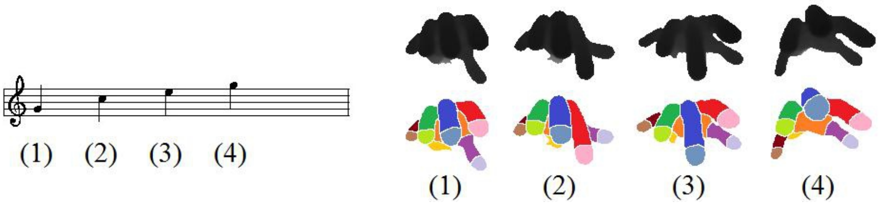


Fig. 2. An “arpeggio” score and piano-like gesture simulation (depth map and ground-truth)

We note that, if [11] and [12] suggests the use of hundreds thousands images for the establishment of a skeletal model relative to a broad spectrum of distinct body or hand postures, we believe that, in our restricted subcase, it is possible to achieve satisfying classification results with a reduced sample.

3.3 Pixel-wise classification

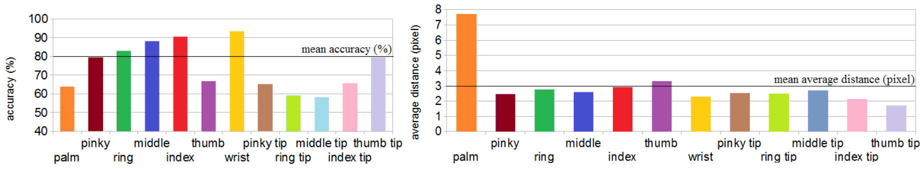
For each pixel of an input depth image, the trained decision trees independently outputs a probability distribution that is relative to the hand subpart assignment. These distributions are then averaged over every tree among the decision forests and form the final pixel probability to belong to each subpart of the model.

Now, from each subpart probability map we estimate the subpart position using the Mean Shift algorithm [13] which has the advantage of converging very quickly. It also allows us to filter the noise on the classification measurement on a pixel level with the thresholding of the probabilies before computing the density estimator.

3.4 Performance

To measure the performance of our proposed method we have compared the output of our pixel-wise classification with 125 images (165x120pixels) that we have previously hand-labelled. In contradiction to [11] and [12] we did not use synthesis images as it is not feasible to reproduce the exact expert gesture. Hence, the hand-labelling may be subject to some variation if the same process was accomplished several times.

Fig. 3.a shows this performance for each of our model subpart with an average classification rate of 80%. Hand subparts performance corresponding to wrist, little finger, ring, middle, index as well as thumb tip is above the average accuracy, which lies slightly above 80%. However, other hand subparts such as thumb or palm are more likely to be occluded by other fingers in the hand configurations we studied. Noteworthy, the fingertips exhibit a lower classification (60-80%) rate which could be explained by the fact that their imaged size is smaller thus implying fewer training data (especially, as only pixels are randomly sampled in the training process). Such assumption could be verified by re-training our model with more images.



(a) (b)

Fig. 3. Performance of our model. (a) Pixel classification rate. (b) Average pixel distance from the joint centroids estimated to the joint centroids ground-truth.

Additionally, to measure the precision of our joint estimation we compare the average distance of the joint retrieved from the Mean Shift algorithm against our ground-truth data. From our 125 test images, the performance reported in Fig. 3.b. shows that all joints position is less than 3 pixels from our ground-truth except the palm joint which is less accurate. Again, this is explained by the fact that the palm is often imaged in several pieces due to finger occlusions making the accurate centroid estimation much more complex.

So far, the performance seems sufficient to be used in a finger gesture recognition pipeline. However, to allow fine pose recognition we lack information about the confidence in the hand subpart position estimation. One solution we did not implement yet would be to compare the zeroth moment score with the output from the mean shift iterations, as it would provide an interesting measure of how important the retrieved local maxima of the probability distribution is.

4 Conclusion and perspectives

This paper presented an implementation of a generic method for building body subpart classification model applied to musical finger gestures for the preservation of the ICH. This methodology can contribute both to the analysis of classical music playing schools but also to the finger gesture control of sound as a new interface for musical expression. 80% pixel-wise classification accuracy and 95% ponctual subpart location accuracy are achieved when musical finger gestures with a semi-closed hand are performed in front of the camera. As a next step in the near future, it is planned to perform data fusion of two depth cameras in order to achieve better results and address the problem of scene and self-occlusions.

Acknowledgements. The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement n° 600676.

References

1. Rasamimanana,N., Bevilacqua, F.: Effort-based analysis of bowing movements: evidence of anticipation effects. The Journal of New Music Research, 37(4): 339 – 351 (2009)
2. Demoucron, M., Askenfelt, A., Caussé, R.: Observations on bow changes in violin performance. In Proceedings of Acoustics, Journal of the Acoustical Society of America, volume 123, page 3123 (1994)
3. Palmer,C. , Pfordresher, P. Q.: From my hand to your ear: the faces of meter in performance and perception. In C. Woods, G. Luck, R. Brochard, F. Seddon& J. A. Sloboda (Eds.) In Proceedings of the 6th International Conference on Music Perception and Cognition.Keele, UK: Keele University (2000)
4. Aylward, R., Lovell, S.D., Paradiso, J. A.: A Compact, Wireless, Wearable Sensor Network for Interactive Dance Ensembles, in Proc. of BSN 2006, The IEEE International Workshop on Wearable and Implantable Body Sensor Networks, Cambridge, Massachusetts, pp. 65-70, April 3-5 (2006)
5. Coduys, T., Henry, C., Cont, A.: TOASTER and KROONDE: High-Resolution and High-Speed Real-time Sensor Interfaces, In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME-04), Hamamatsu, Japan (2004)
6. Todoroff,T.: Wireles digital/analog sensors for music and dance performances. In Proc. NIME ’11, pages 515–518, Oslo, Norway (2011)
7. Grunberg, D.: Gesture Recognition for Conducting Computer Music. Retrieved January 10, 2009, from: http://schubert.ece.drexel.edu/research/gestureRecognition (2008)
8. Burns,A. M., Wanderley,M.: Visual Methods for the Retrieval of Guitarist Fingering. In Proceedings of the International Conference on New Interfaces for Musical Expression. Paris, France (2006)
9. The iMuse project, last retrieved on 25th June 2013 :
   http://debussy.music.ubc.ca/muset/imuse.html
10. Manitsaris, S., Tsagaris, A., Dimitropoulos, K., Manitsaris, A.: Finger musical gesture recognition in 3D space without any tangible instrument for performing arts, International Journal of Arts and Technology (in press)
11. Shotton, J. , Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: “Real-time human pose recognition in parts from single depth images,” Communications of the ACM, vol. 56, no. 1, p. 116 (2013)
12. 12. Keskin, C., Kıraç, F., Kara, Y., Akarun, L.: “Real time hand pose estimation using depth sensors,”, Depth Cameras for Computer (2013)
13. Comaniciu, D.,Meer, P.: “Mean shift: a robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619 (2002)
14. Schwarz, L. A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing, vol. 30, no 3, p. 217-226 (2012)
15. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. In : British Machine Vision Conference. p. 101.1-101.11 (2011)
16. Breiman, L.: “Random forests,” Machine learning (2001)