Analysing Musical Performance in Videos Using Deep Neural Networks

Foteini Simistira Liwicki[1], Marcus Liwicki[1], Pedro Malo Perisé[2], Federico Ghelli Visi[3] and Stefan Östersjö[3]

[1] Machine Learning Group
Luleå University of Technology, Sweden
[2] University of Zaragoza, Spain
[3] GEMM-Gesture Embodiment and Machines in Music
Luleå University of Technology, Sweden
foteini.liwicki@ltu.se

Abstract. This paper proposes a method to facilitate labelling of music performance videos with automatic methods (3D-Convolutional Neural Networks) instead of tedious labelling by human experts. In particular, we are interested in the detection of the 17 musical performance gestures generated during the performance (guitar play) of musical pieces which have been video-recorded. In earlier work, these videos have been annotated manually by a human expert according to the labels in the musical analysis methodology. Such a labelling method is time-consuming and would not be scalable to big collections of video recordings. In this paper, we use a 3D-CNN model from activity recognition tasks and adapt it to the music performance dataset following a transfer learning approach. In particular, the weights of the first blocks were kept and only the later layers as well as additional classification layers were re-trained. The model was evaluated on a set of 17 music performance gestures and reports an average accuracy of 97.9% (F1:77.8%) on the training set and 85.7% (F1:38.6%) on the test set. An additional analysis shows which gestures are particularly difficult and suggest improvements for future work.

1 Introduction

The main purpose of this paper is to investigate if state-of-the-art deep learning techniques can be utilized to perform tedious tasks in labelling of videos. Videos capturing music performance are particularly difficult to analyze, as the individual gestures performed by the artists may be very subtle and difficult to label, even by human experts. This work builds on an existing musical analysis methodology proposed by Coorevits et al. in (Coorevits, Moelants, Östersjö, Gorton, & Leman, 2015). The contribution of this work is adapting a deep-learning based video analysis model to the task of musical performance analysis and thereby enabling it to recognize facial and body expressive gestures. Those gestures were defined by Coorevits et al. in the article "Decomposing composition" (Coorevits et al., 2015) and include, among others, shoulder movement, hands movement, or facials expressions, such as closed eyes. The aim of this paper is to be able to detect and identify automatically those gestures for a given video-recorded performance. The desired solution is a machine learning model which, given a set of videos as input, can generate a list of detected features for each video segment. The chosen architecture is based on Kensho Hara's et al (Hara, Kataoka, & Satoh, 2018) ResNet-34, a 3D convolutional neural network (3D-CNN). It is adapted to suit the task of music performance analysis, i.e., to recognize 17 predefined gestures.

2 State of the Art

Video Analysis with machine learning techniques has emerged in the field of action recognition in video streams with several small datasets, like HMDB-51 (Jhuang, Garrote, Poggio, Serre, & Hmdb, 2011) and UCF-101 (Soomro, Zamir, & Shah, 2012) and most recently with larger, like ActivityNet (Caba Heilbron, Escorcia, Ghanem, & Carlos Niebles, 2015) and Kinetics (Carreira & Zisserman, 2017). The most relevant related work uses 3D-CNN architectures (called space-time CNNs with long-term temporal convolutions) (Varol, Laptev, & Schmid, 2017), they achieve 92.7% on UCF-101 and 67.2% on HMDB-51. For action detection in video streams, such a 3D variant of the CNNs has been proven to be more efficient than the 2D variant, i.e., (Hara et al., 2018) reports an accuracy of 94.5% on the UCF-101,

70.2% on the HMDB-51 and 78.4% on the Kinetics dataset. This architecture is used as underlying base architecture in this present paper.

Music Performance Analysis emerged as a field of study in music psychology and musicology in the late 1980s, with an interest in a wide range of perspectives, including the role of gesture, performer-performer interaction, social contexts, etc. (Bowen, 1996). Theories of embodied music cognition posit that music is a multimodal medium experienced not only through sound, but also through visual and kinematic cues. These theories have given rise to a further study of musical performance through the analysis of movement data captured during performance (Godøy & Leman, 2010). In order to gain a deeper insight, recent multi-method approaches combine quantitative and qualitative data (Coorevits et al., 2015; Gorton & Östersjö, 2019). Gesture analysis so far, requires additional body sensors (see also in the next section) and/or human labelling. Although quantitative analysis of video data of music performance alone appears to be a fruitful field for further development, it remains an under-researched area, which could benefit from the employment of recent deep learning techniques.

3 Dataset

The dataset used in this work takes the videos recorded in the (human expert) study by Coorevits et al. (Coorevits et al., 2015), accompanied by the qualitative coding resulting from the analysis that was carried out as part of this previous study. Two rehearsals and two concert recordings were recorded, with audio, video and movement data. In this paper, only the video data is taken for analysis. The audio data did not help to improve the accuracy of the results. The movement data from the other sensors have been used by the human expert to generate the ground truth.

In total, the analytical process resulted in a total number of 17 codes, which corresponds to the classes discussed below. Most of the codes, such as "nodding" and "expressive shoulder movement", are body movement that has no direct result on technical delivery and does not explicitly produce sound. "Vibrato" was also included as a code even though it does not refer to a particular gesture, and rather has an immediate effect on delivery, particularly in the shaping of material, and obviously it does modulate the sound even if not producing it. But what all the codes have in common is that they represent corporately perceived bodily strategies for the management and communication of the evolving musical structures. For a further description of these coded gesture types, see (Gorton & Östersjö, 2019, pp. 67-77). The labels were selected every 16 video frames, as required by the 3D CNN architecture (see Section 4). The total number of frames in the current video is 11803, therefore there are in total 737 samples. 20% of the data were used as the test set and 80% as the training set. The total number of samples are 1746, 1424 in the training set and 322 in the test set. Table 1 shows the class distribution in the training and test set.

Table 1. Class distribution.

Class	**Train**	**Test**
Vibrato	36	8
Upbeat in head movement	22	2
Repositioning guitar	29	4
Nodding	147	32
Frowning	43	11
Freeze	45	11
Facial expression	256	51
Eyes closed	77	12
Expressive shoulder movement	154	37
Expressive head movement	47	8
Expressive preparation	60	11
Right hand round	131	33
Minimal movement	167	59
Lifting head	77	18
Left hand gesture	92	17
Physical energy	36	6
Sympathetic body movement	5	2

4 Methodology and results

The 3D-CNN network architecture proposed in this paper is an adapted version of (Hara et al., 2018). In this paper, we replace the final layer with a few ReLU layers and a final sigmoid layer having two outputs per class (one for present, one for not present). These two outputs are required as we have a labelling task, i.e., several labels can be present at the same time frame. The variants of ReLU layers investigated in this study are shown in Table 2. In the current work, the Minimum Square Error (MSE) loss function

Table 2. NN architectures and their corresponding accuracy on the training set

No	NN-Architecture	Pre-Trained	Accuracy
1	100 + 50	No	83%
2	300 + 150 + 100 + 50	No	80%
3	200 + 50	No	62%
4	200 + 50	Yes	80%
5	**200 + 50 + ReLUs**	**Yes**	**97.9%**

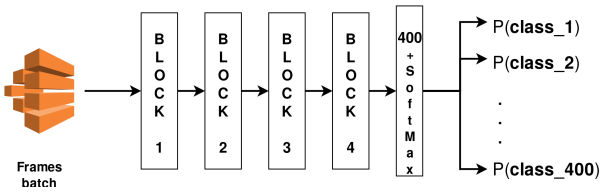

Fig. 1. Original Kensho Hara's et al 3D-CCN model

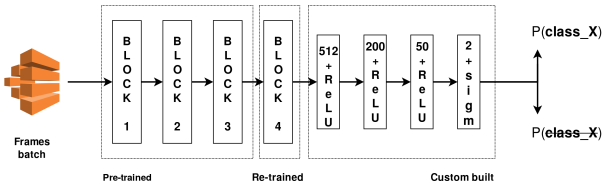

Fig. 2. Proposed architecture

was used. Most of the weights of the original architecture were kept (transfer learning). This is done in order to keep the ability of the model to detect meta-features related to action recognition that can be useful for our goal. The weight that were not kept are:

- **Fourth block;** Instead, the fourth block of the network (formed by six residual convolutional layers) has been retrained. This layer outputs a 512-vector, which can be considered an embedding vector of the 512 most important features for each class.
- **Final block;** Instead, the final block has been customized and substituted by a set dense connected layer using ReLU as activation function. This is performed in order to make the model able to learn the new meanings of the meta features extracted in the previous block. Sigmoid activation function is used instead of the Softmax, because Softmax brings the problem of a decision boundary; two outputs are often a better alternative, as in this case.

The input of the model is 16 frames (as required by the 3D-CNN model), and the output is a 2D-vector. The first element is the likelihood that the input belongs to the given class and the second element is the likelihood that the input does not belong to the given class. An example architecture (architecture

5 from Table 2) can be seen in Fig. 2. Noteworthy, there are in total 18 CNNs. Each of them calculates the likelihood of an input sample belonging or not belonging to a class. This might be due to the big conceptual differences of the individual classes which could be mitigated by a larger architecture (but could lead to overfitting as we only have 1424 training sequences). Notice that we added an extra class named normal play, when all the other classes are not present, leading to 18 classes to be recognized. The experiments performed on a single GPU machine (NVIDIA GTX 1080 TI) using Jupyter notebooks and Anaconda environment.

Table 2 shows the training accuracy for the different architectures. The best architecture reports an average accuracy of 97.9% (F1:77.8%) on the training set and 85.7% (F1:38.6%) on the test set. Tables 3 and 4 show the accuracy for every individual class as well as precision, recall, and the F1 score which combines precision and recall for better comparison. Noteworthy, the precision and recall are rather low for some classes. The main reason for that is the unbalanced dataset: there are much less positive samples than negative ones. Furthermore, some gestures seem to be ambiguous (expressive or sympathetic movements) or would require additional sound (vibrato). Classes with high precision and recall are highlighted in bold. One can see that precision and recall are much higher on the training set, indicating overfitting. Figure 3 shows more details in confusion matrices of selected classes. Measures to overcome this problem are presented in section 5.

Table 3. Evaluation results - training set

Class	Precision	Recall	F1	Accuracy
Vibrato	0.972	0.972	0.972	0.990
Upbeat in head movement	0.950	0.864	0.655	0.990
Repositioning guitar	0.081	0.655	0.144	0.947
Nodding	0.966	0.980	0.973	0.986
Frowning	0.952	0.930	0.941	0.990
Freeze	0.978	0.978	0.978	0.986
Facial expression	0.869	0.988	0.925	0.952
Eyes closed	0.987	0.987	0.987	0.995
Expressive shoulder movement	0.974	0.987	0.981	0.981
Expressive head movement	0.932	0.872	0.901	0.990
Expressive preparation	0.966	0.933	0.949	0.986
Right hand round	0.928	0.977	0.952	0.962
Minimal movement	0.964	0.958	0.961	0.966
Lifting head	0.976	0.532	0.689	0.952
Left hand gesture	0.989	0.989	0.989	0.986
Physical energy	1.000	1.000	1.000	1.000
Sympathetic body movement	0.000	0.000	0.000	0.976
Normal play	0.000	0.000	0.000	0.986
Average			0.778	0.979

5 Conclusion

The most important risk of our approach is the generalization. Our model has a very poor performance with samples never seen before.

In future work, we will consider using a simpler model, such as an SVM, try regularization techniques to achieve better generalization results and data augmentation. The current dataset is small and therefore problematic, but the training accuracy points to this technique being promising for further study.

A limitation of the present study was that the audio data was not taken into account. Previous research on multimodal machine learning (Baltrusaitis, Ahuja, & Morency, 2019), indicates that analyses that relate data from several modalities might capture complementary information that is otherwise not apparent in individual modalities on their own. Thus, future work will integrate the proposed video analysis into a multimodal approach combining quantitative and qualitative methods.

Acknowledgements. The authors would like to thank Killian Murphy for the initial preparation of the dataset.

Table 4. Evaluation results - test set

Class	Precision	Recall	F1	Accuracy
Vibrato	0.167	0.125	0.143	0.920
Upbeat in head movement	0.333	0.500	0.400	0.981
Repositioning guitar	0.021	0.250	0.038	0.660
Nodding	**0.667**	**0.688**	**0.677**	0.858
Frowning	**0.750**	**0.273**	**0.400**	0.939
Freeze	**0.800**	**0.727**	**0.762**	0.967
Facial expression	0.532	0.824	0.646	0.689
Eyes closed	0.600	0.750	0.667	0.939
Expressive shoulder movement	0.634	0.703	0.667	0.821
Expressive head movement	0.000	0.000	0.000	0.929
Expressive preparation	0.500	0.182	0.267	0.925
Right hand round	0.361	0.394	0.377	0.708
Minimal movement	**0.889**	**0.678**	**0.769**	0.835
Lifting head	0.800	0.222	0.348	0.896
Left hand gesture	0.571	0.471	0.516	0.896
Physical energy	0.200	0.167	0.182	0.939
Sympathetic body movement	0.000	0.000	0.000	0.802
Normal play	0.065	0.154	0.091	0.726
Average			0.386	0.857

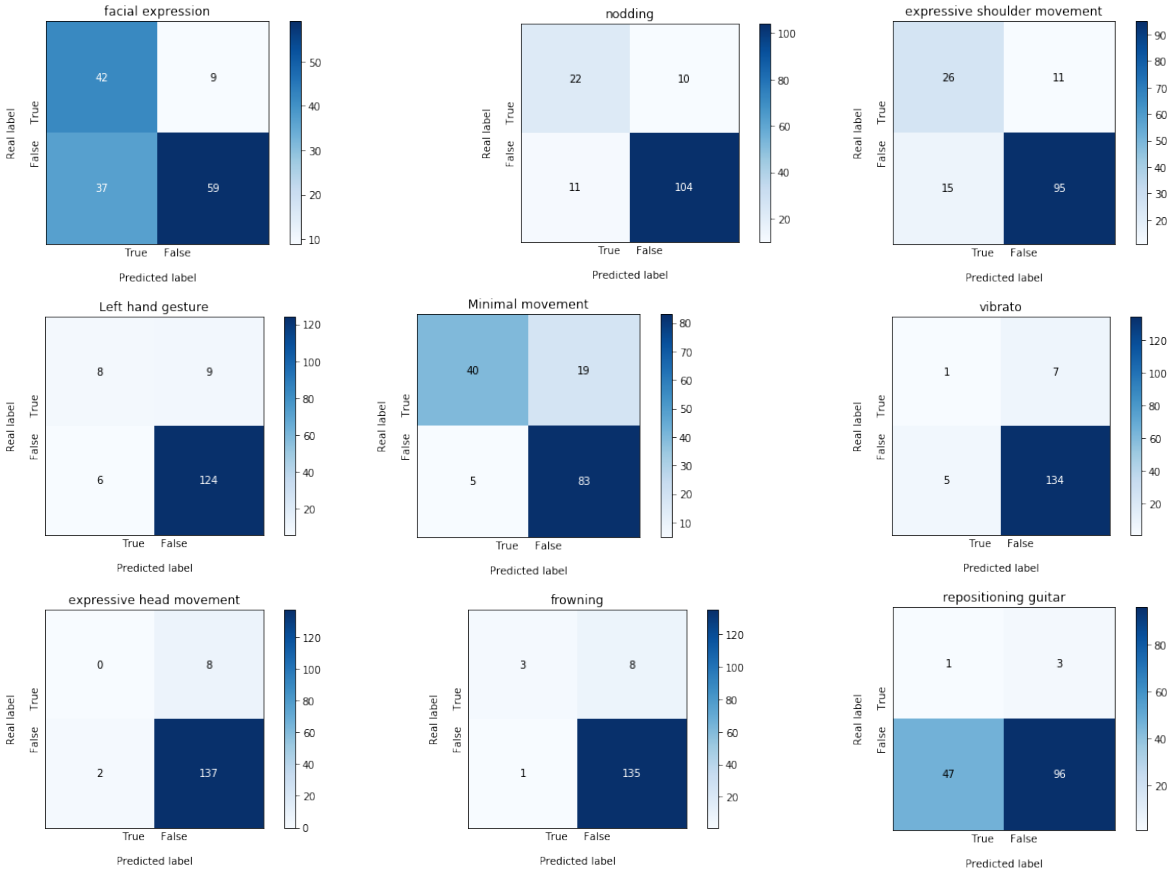

Fig. 3. Confusion matrix per class - Test set

References

Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. doi: 10.1109/TPAMI.2018.2798607

Bowen, J. A. (1996). Performance practice versus performance analysis: Why should performers study performance. *Performance Practice Review*, *9*(1), 3.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 961–970).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6299–6308).

Coorevits, E., Moelants, D., Östersjö, S., Gorton, D., & Leman, M. (2015). Decomposing a composition: On the multi-layered analysis of expressive music performance. In *International symposium on computer music multidisciplinary research* (pp. 167–189).

Godøy, R. I., & Leman, M. (2010). *Musical gestures: Sound, movement, and meaning*. Routledge.

Gorton, D., & Östersjö, S. (2019, nov). Austerity Measures I:. In C. Laws, W. Brooks, D. Gorton, N. T. Thuy, S. Östersjö, & J. J. Wells (Eds.), *Voices, bodies, practices* (pp. 29–80). Universitaire Pers Leuven. Retrieved from `http://www.jstor.org/stable/10.2307/j.ctvmd83kvhttp://www.jstor.org/stable/10.2307/j.ctvmd83kv.6` doi: 10.2307/j.ctvmd83kv.6

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 6546–6555).

Jhuang, H., Garrote, H., Poggio, E., Serre, T., & Hmdb, T. (2011). Hmdb: A large video database for human motion recognition. In *Proc. of ieee international conference on computer vision* (Vol. 4, p. 6).

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Varol, G., Laptev, I., & Schmid, C. (2017). Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1510–1517.