CONVOLUTIONAL NETWORKS FOR VISUAL ONSET DETECTION IN THE CONTEXT OF BOWED STRING INSTRUMENT PERFORMANCES

Grigoris BASTAS (Γρηγόρης Μπάστας)[1,3], **Aggelos GKIOKAS** (Άγγελος Γκιόκας)[2],
Vassilis KATSOUROS (Βασίλης Κατσούρος)[1], and **Petros MARAGOS** (Πέτρος Μαραγκός)[3]

[1]*Institute for Language and Speech Processing (ILSP)*, **Athena R.C.**, Athens, Greece
[2]*Music Technology Group*, **Universitat Pompeu Fabra**, Barcelona, Spain
[3]*School of Electrical and Computer Engineering*, **National Technical University of Athens**, Athens, Greece

ABSTRACT

In this work, we employ deep learning methods for visual onset detection. We focus on live music performances involving bowed string instruments. In this context, we take as a source of meaningful information the sequence of movements of the performers' body and especially the bowing motion of the (right) hand. Body skeletons for each video frame are extracted through OpenPose and are then used as input for Temporal Convolutional Neural Networks (TCNs). TCNs prove capable of handling such temporal information by conditioning outputs on an adequately long history (i.e. variable receptive field), ensuring highly parallelizable lightweight computations and a multitude of trainable parameters that provide robustness. As another source of information for our task, we consider the more subtle movements of the (left) hand fingers which are responsible for pitch changes. Detections in this case rely directly on pixel data from specifically chosen regions of interest. Here, a 2D Convolutional Neural Network (CNN) is applied on the input in order to learn the features to be fed to the TCN. The models were trained and evaluated on single-player string recordings from the University of Rochester Multi-Modal Music Performance (URMP) Dataset. We show that these two approaches provide some complementary information.

1. INTRODUCTION

Onset times is one of the most fundamental elements of the temporal organization of a music piece. Onsets are placed relatively to the metrical grid of a music piece and their positions on this grid (along with duration and velocity) greatly determines the rhythm structure of a music performance. Moreover, the micro-timings, i.e. small time deviations of the onset with respect to the ideal metrical grid are very related to aspects of human expressivity. Consequently, onset detection comprises a fundamental problem in the field of Music Information Retrieval and it is related to many other tasks including tempo estimation, beat tracking, music transcription, source association.

Music is often experienced by humans in a visual context. The visual perception plays a key role whether the stimulus involves an album cover, a video clip or a live music performance [1]. Research in psychology has documented an impact of the visual information on the judgment over musical live performances [2] or even on how musical structure is perceived [3]. During the past few years, audiovisual analysis has drawn the attention of the music information retrieval community. Visual information can be important for deducing a performer's stylistic techniques, recognizing the playing instruments, capturing the emotional variations in a piece, etc. Innovative information extraction techniques employed in such a multimodal context have been evolved [4]. Traditional image processing, pattern recognition and deep learning methods have been used to deal with tasks relevant to this emerging field.

The development of convolutional networks has been decisive for the advancements in computer vision during the last decades. Convolutional Neural Networks (CNNs), which apply two-dimensional convolutions, play a crucial role in machine learning because they enable learning latent features from images, adaptable according to each specific task. During the last few years, convolutions have also been employed to handle sequential data using learnable filters to convolve over the axis of time. These types of architectures which involve one-dimensional convolutions are called Temporal Convolutional Networks (TCNs) and have exhibited some advantages over the often employed Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) units.

For many musical instruments, the produced sounds correspond to certain visible movements and specific positioning of the instrument player's hands. More particularly, with regard to the bowed string instruments, the bowing motions lead to the articulations of music notes (see Fig. 1). Such visual cues are detectable by the human eye quite easily. Pitch change requires altering the positioning of the fingers on the neck of the instrument. Each fingering transition is strongly correlated to note onsets.

Capturing such visual information content with computational methods can be challenging, that is why state-of-the-art computer vision techniques need to be employed. Occlusions and irrelevant or subtle movements are not easy to cope with. Naturally, not all right hand movements correspond to onsets, and several onsets can be produced by legatos (i.e. left hand pitch change) without changing the

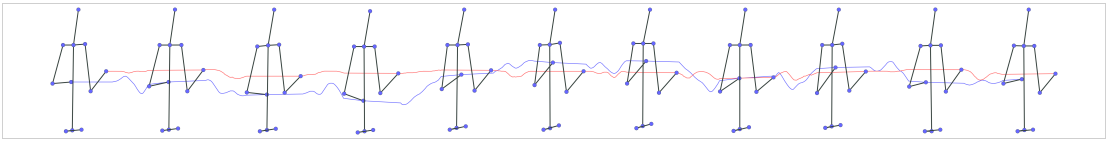

Figure 1. Instances of a moving skeleton extracted from a violin performance. Trajectories of the right hand are displayed with blue color and for the left hand with red.

direction of the bowing motion. Similarly, fingering transitions of the left hand may not bear musical information, or might not signal the exact time location of an onset. Such motion can also be difficult to discriminate from vibrato. Also, onsets of the same note can be produced simply by new bow strokes without any change in the fingering. Difficulties arise since part of the left hand is occluded by the neck of the instrument and occlusions might occur between fingers, depending also on the relative position of the hand to the camera. Additionally, discriminating between the motion of the hand in the scene and the relative motion of the hand on the neck of the instrument constitutes another challenging aspect of the problem. Since these movements can be very subtle, the stability of the camera and its distance from the hand are also crucial parameters.

The standard process to infer onset locations is by working on the audio signal. In this paper, however, we present a method for visual onset detection. We deploy TCNs and CNNs, and we demonstrate that the visual modality can be a source of meaningful musical information that, handled correctly, can help to cope with this challenge. We focus on bowed string single-instrument performances where the hand and body movement can provide visual information for the onset locations.

The rest of the paper is organized as follows. Section 2 provides related work for the fields of onset detection, the multimodal or purely visual approaches for music analysis and generation. Section 3 describes the proposed method, and section 4 provides the experimental results. Section 5 concludes the paper with discussion and future directions.

2. RELATED WORK

2.1 Visual-based Music Analysis and Generation

Visual-based approaches on music analysis have been applied on several different tasks during the past years. The visual modality has played a critical role in tasks such as audio-visual source association, fingering analysis, playing/non-playing (P/NP) activity detection, vibrato analysis, automatic music transcription and onset detection [4]. Parekh et al. [5] have engaged in audio-visual source separation in polyphonic performances, focusing on motion tracking of bowed string performers, using unsupervised learning techniques like Non-negative Matrix Factorization (NMF). Several research teams have worked on fingering tracking focusing on instruments like guitar, based on computer vision and statistical tools [6–9], aiming as far as music transcription in symbolic forms like tablature [10, 11]. In [8], finger tracking on the guitar player was used to detect "key frames" (i.e. the time locations of chord changes), a notion very close to onset-corresponding frames. In [12] a method for guitar transcription is presented relying on video close-up recordings of the vibrating strings [12]. Fingering recognition and hand tracking systems have also been developed for piano performances [13, 14] and violin [15].

Other teams have undertaken audio-visual analysis using deep learning models. CNNs have been used in tasks like instrument recognition where the visual modality is prominent [16, 17]. CNNs have also been used for localization of specific regions in video frames that correspond to distinct music signal sources and thus also make it possible to separate the two signals [18]. Extracted skeleton poses have also been used to extract music information from the visual modality or to study audio-visual correspondence. Pedersoli and Goto [19] introduced the task of Dance Beat Tracking, where they employed TCNs to predict onset locations, having as input skeleton poses of the dancers while performing.

Apart from the context of music analysis of the visual content, body motion related recognition techniques have been deployed in the context of music generation, as for example using skeleton data from Kinect sensor for air-guitar playing [20] or using finger motion data for recognizing gestures in order to perform on virtual instruments [21]. GANs were recently been employed for visually enhanced audio inpainting based on live music performances [22].

2.2 Audio and Visual Onset Detection

Traditionally, the task of onset detection has been dealt with using information from the audio signal. The state-of-the-art uses a Convolutional Neural Network (CNN) architecture [23] in a similar way as CNNs have been used for edge detection in the field of computer vision. Three versions of 80-band log-mel features were used to represent the audio input with three different corresponding window sizes. Hence, the input was given in the form of three channels on which 2D convolutions were applied using two layers involving max-pooling. RNNs have also been employed successfully to tackle the problem [24].

Zhang and Wang [15] engaged in audio-visual music transcription for violin. Onset detection was a fundamental part of such a system. They presented both visual and audio methods for detecting onset times. The audio-based method relied on training Gaussian Mixture Models (GMMs) on Mel-Frequency Cepstral Coefficient (MFCC) features. The visual-based method was centered on the analysis of two different sources of visual information provided by two cameras, one recording the bowing motion

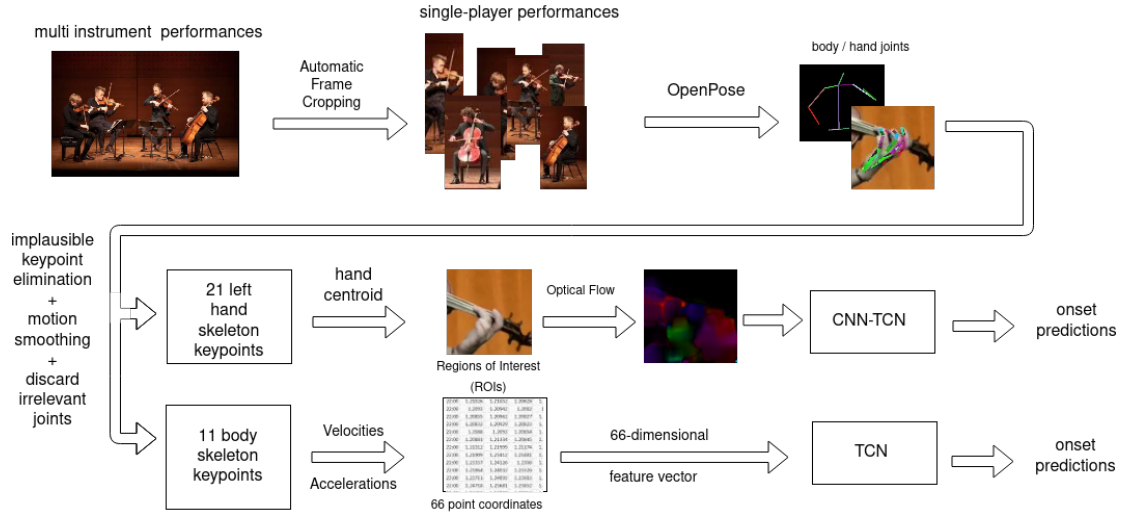

Figure 2. A flow diagram depicting our methodology for visual onset detection.

and the other recording the fingering transitions. Distinct prediction functions for the two video signals were employed, in the first case relying on tracking the hand holding the bow and in the second, concentrating on recognizing the violin strings and the relative position of the hands on them. The resulting prediction scores of each distinct source of information were finally combined using feature level and decision level fusion.

Audio-visual analysis focusing on onset detection for string ensembles has also been conducted by Li et al. [25], serving as a basis for score-informed audio-visual source association. The onset locations were estimated by focusing on bow stroke detection and more specifically, on the zero-crossings of the principal motion velocity, computed using the information by optical flow vectors. Audio-visual source association has been handled using vibrato analysis [26], too. In contrast with the aforementioned works, in [27], the visual information was reduced, using OpenPose [28], to keypoints representing body and finger joints. The vibrato and bow stroke approaches have been combined and the onset detection task has been aided by following the finger movements of the players' left-hand, thus also permitting the generalization of the analysis on woodwind and brass instruments. Recently, TCNs have been used to detect onsets from the audio and visual sources and for fusing these two modalities [29]. In our current work, we focus on the visual modality and present a significantly improved feature extraction procedure and a more robust TCN model architecture to process skeleton data. We compare this strategy to a pixel-based approach which aims at capturing left-hand motion.

3. METHOD DESCRIPTION

3.1 Method Overview

An overview of the proposed methods is illustrated in Fig. 2. First, the videos are cropped in order to work with single-player performances. For each single-player performance, we apply OpenPose in order to extract the performer's skeleton. The extracted skeleton is smoothed and we subsequently extract velocity and acceleration features, while from the centroid of the left-hand keypoints we extract Regions of Interest (ROIs) in order to isolate the hand from the rest of the image. For these ROIs, we compute the optical flow to capture the motion of the left hand. The extracted body skeleton features are used to train a TCN, while the left-hand features are used to train a network that comprises conventional CNN and TCN layers.

3.2 Video Processing

3.2.1 Preprocessing

As a first step, the videos were automatically cropped using ffmpeg command-line tool, in order to ensure that only one person would appear in each recording. For this task, we took advantage of the available information about the number of the players involved in every performance and we chose to segment the frames accordingly, in equal parts, with respect to the x-axis. Minor corrections were required to be done by hand in the case of only one recording. We used OpenPose [30] for 2D pose estimation by employing the BODY_25 output format with hand keypoint detection enabled using the officially suggested scale number and scale range values (6 and 0.4 respectively) for achieving best results. We thus extracted body skeletons comprised of 25 points, together with 21 extra keypoints for each hand.

3.2.2 Body Motion Information

Deriving onsets only from body movements does not require all of the predicted skeleton keypoints. In this setting, hand joint keypoints, ears and eyes, as well as those points corresponding to the knees and below were all ignored because they were considered redundant or in some cases they were prone to occlusions. We were left with 11 body

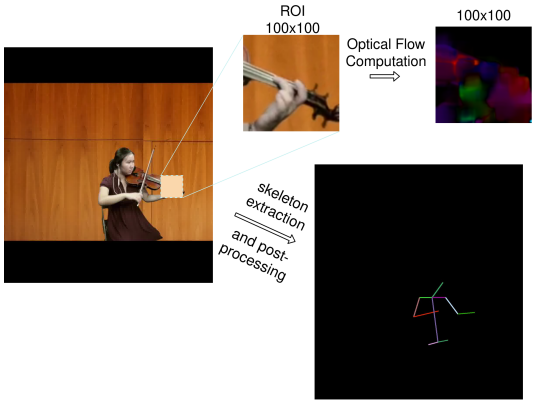

Figure 3. Skeletons and ROIs are extracted from a video frame involving a violonist from a performance in the URMP Dataset [28]. Optical flow is further computed for the isolated hand ROI.

keypoints (see Fig. 3) and hence 22 coordinate values. We followed the processing steps from [31] in order to create continuous skeletons: we eliminated joints with confidence score lower than 0.2 and, in order to rule out abrupt and unnatural keypoint shifts, we discarded keypoints if distance of a joint from one frame to the other was found larger than the 10 percent of the nose to hip distance. In frames where certain joints were occluded or eliminated (following the above criteria), or even in the rare cases where OpenPose failed to capture a skeleton, the keypoints were recreated using linear interpolation between valid frame instances. Centered moving average with window size 5 was used to further smooth the skeleton motion.

After having computed the centroid of the skeletons for each frame and the corresponding standard deviation of the keypoints, we have applied z-score normalization in 2D space using the mean values of the two magnitudes along time. This procedure was applied separately for each performance. In this manner, we have first of all imposed a new center of axis which enabled us to rule out the differences among the positioning of the performer in relation to the edges of the video frames for each distinct recording. Secondly, we eliminated the variation of skeleton sizes among the performances which is introduced especially due to the different distances of the performer from the camera. Finally, keypoint velocities and accelerations were used as additional features, leading to a 66-dimensional input.

3.2.3 Pixel Data

In order to get information from the hand responsible for the pitch changes, one should focus on the finger movements and positioning. The OpenPose keypoints corresponding to the performers' hands are highly inaccurate in this particular setting since they are often occluded by the neck of the instrument and involve jitter. The relatively long distance of the hand from the camera makes things even worse. This raises the need for extracting features directly from the images. For this task we chose to employ Convolutional Neural Networks (CNNs). In order to eliminate redundant visual information and focus on the (left) hand movements we took advantage of the predicted hand keypoints to isolate specific regions of interest (ROIs). To achieve this, we computed the centroids of the left hand joint for each consecutive frame as in [27], but we used only the average centroid across time. This point in 2D space serves as the center of a steady square ROI. We avoided the use of moving ROIs following the hand centroid as in [27] because we observed a strong effect of the changing background patterns to our recognition system. The size of the ROIs in the case of violin and viola performances was 100x100 pixels while for the violoncellos and the double basses 200x200 frames in order to capture all the range of vertical hand movement on the instrument's neck. These last larger ROIs were rescaled to 100x100 images in order to ensure same input sizes for the onset detector. When dealing with ROIs that extend out of the frame (on the right side to be specific), we employ zero-padding to ensure proper image sizes.

3.3 Temporal Convolutional Networks

Temporal Convolutional Neural Networks (TCNs) constitute a family of architectures designed to grasp dependencies among sequential data. Dilated one-dimensional convolutions enable control of the network's receptive field, that is the considered length of dependencies among temporal data. Every added layer increases the temporal scope of the network exponentially. Consequently, with few added trainable parameters one can ensure adequate source of information. Another hyper-parameter that is pertinent to the temporal dependencies that one intends to grasp is the size of the convolutional filters across the time axis. A good combination of these two values permits handling of complex data associations. As in the case of two-dimensional convolutions, the computations can be efficiently parallelized using GPUs. That is a limit that Recurrent Neural Networks (RNNs) are confronted with since each new forward pass shall wait for the output of the previous step to be produced. Various versions of TCNs have recently been introduced in [32–34], for generative or classification purposes.

Henceforth, TCNs definitely fit our needs for the problem of visual onset detection which requires handling sequential data. TCNs proved to lead to exceptional results with strong generalization abilities. Using a stack of TCNs, instead of only one, was found to lead to over-fitting in our setting which involves relatively few data. The architecture that was tested in this work is inspired by Wavenet [32] and the TCN proposed in [34]. We do not however stick to the proposed causal setting where each prediction relies on past observations. Our non-causal configuration visualized in Fig. 4 involves 9 layers ($l \in [1, 9]$) where two distinct sets of one-dimensional convolutional filters $W_{1,l}$ and $W_{2,l}$ are learned. Weight normalization is applied in both arrays. Hyperbolic tangent is used in the first case and sigmoid activation function in the other. The two outputs are then combined using element-wise multiplication.

This parallel configuration imposes non-linear filtering

$\tanh(W_{1,l} * x)$ and a learnable mask $\sigma(W_{2,l} * x)$ applied in each layer. In every layer, 256 convolutional filters comprise each of the two convolutional blocks. Residual connections appear in each layer involving 1x1 convolutions to upsample the input and fit it to the size of each layer's output when needed. The use of 9 layers with a variable dilation factor 2^l (assuming videos with standard frame rate (29.97 fps) as in our case) entails a receptive field of about 17 sec, centered around the timestamp for which a prediction is to be made. Small kernels of size 3 across the time axis are used and symmetrical zero-padding (increasing from layer to layer in accordance to the dilation factor) is applied at the beginning and at the end of each performance to ensure non-causality. Dropout with probability 0.25 is applied to avoid over-fitting, as well as gradient clipping with a corresponding coefficient of value 0.2. A linear fully connected layer followed by a softmax function is used to output 2-dimensional vectors for each input frame, with each coordinate representing the probability of an occurring and a non-occurring onset respectively.

3.4 Convolutional Neural Networks

However suitable TCNs may be when fed with features such as sequences of post-processed skeleton coordinates, they cannot efficiently extract information directly from pixel data as in the case of left-hand ROIs. Proper features should be extracted first in this case. An end-to-end configuration where latent features are learned has been proven an effective strategy since it can be directly adaptable to specific tasks. CNNs constitute a standard for image handling. In our case, a 3-layer CNN is employed with 5x5 kernel size as displayed in Fig. 4. The number of filters increases from one layer l to the next, by $4 \cdot l$. ReLU is used as activation function after batch normalization. Max pooling with kernel size 3x3 and stride 3 is applied in order to downsample each layer's output. In the network's output, 16 feature arrays of size 4x4 are produced and are then fed to the TCN after having been flattened, ending up with vectors of 256 dimensions.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

We trained and evaluated the proposed models on video recordings from the University of Rochester Multi-Modal Music Performance (URMP) Dataset [28]. This dataset provides videos of multi-instrument performances that were created by assembling audio-visual recordings of individual music players performing separately, yet coordinated. The audio recordings of individual instrumental performances are also provided in the dataset, thus enabling the matching of each separate track with the corresponding cropped video (i.e. individual performances). Among these videos only the string instrument performances were used in our experiments. So, in total 61 single-instrument performances comprise our data. The duration of these performances vary strongly from 35 sec to 3.8 min.

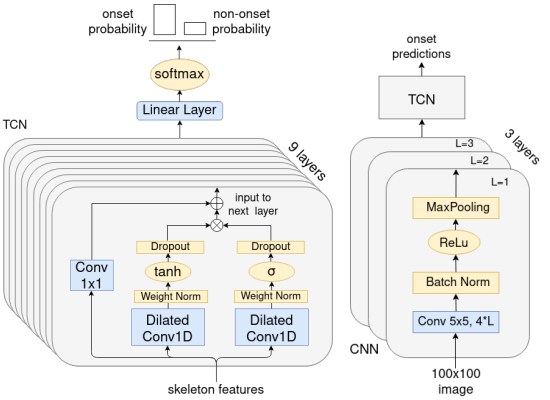

Figure 4. On the left side a 9-layer TCN is displayed, receiving as input skeleton features. A linear fully connected layer is applied on the output with a softmax activation function. On the right side a CNN-TCN model is depicted. The 3-layer CNN learns features from consecutive images and feeds them as input to a TCN.

4.2 Training and Evaluation Procedure

For the purpose of our models' evaluation we undertake 8-fold cross-validation using pytorch python library. The code can be found in [35]. The data were shuffled beforehand. All string instruments were involved in the training procedure. While training, a window of 3 frames around the annotated timestamps was considered to correspond to valid onset instances and, hence, we assigned as ground-truth for the 2D softmax output the probability vector $y = (1, 0)$ in the case of occurring onset and $y = (0, 1)$ when no onset occurs. Trainable parameters are considered to be all the 1D and 2D convolutional filters involved, plus the fully-connected linear layer in the output. Cross-entropy loss function was employed. The value 0.001 was opted for the initial learning rate, together with Adam optimizer. The average F measure is computed separately for the training, validation and test set. A maximum of 250 training epochs were run for each fold and the best model parameters were stored using as a criterion F measure values calculated for the validation set. After training, the best performing model versions in the validation set were then reloaded for the final testing. Two experiments were run:

- a TCN was applied on the extracted skeleton features
- a CNN-TCN was applied on the left-hand ROIs

An estimated note onset was considered to be correct if it was found ± 50 ms around the annotated timestamp. This is an adequately small range in the case of visual onset (also proposed in [4]), used instead of the $\pm$ 25 ms tight criterion, since the distance between subsequent video frames ($\sim$33 ms) exceeds this value. Local maxima of the activation function were computed. The peaks were found using a centered moving maximum with a window size of 3 consecutive frames. We used a threshold of 0.5 to derive predictions. For each individual recording, either audio or video, precision, recall and F measure were computed.

4.3 Results

4.3.1 Skeletons and Hand ROIs

As a first step, we investigated the ability to detect onsets by focusing on body movements. Sequences of post-processed 11-keypoint skeletons were given as input to the TCN model (TCN-Sk). Next, we applied the CNN-TCN (CNN-TCN-ROI) on the isolated ROIs capturing the left performers' hand. We compared the two results quantitatively and qualitatively. The model trained on the skeleton poses outperformed the pixel model by 17% in the overall accuracy as presented in Table 1. This fact leads us to the conclusion that the explicit movements of the body and especially the bowing motions of the right hand (see Fig. 1) can provide very clear information concerning the time of each note articulation, even on the relatively small training set we used. One can notice that Precision exceeds Recall for TCN-Sk. On the other hand, CNN-TCN-ROI yields slightly greater Recall than Precision. This signifies that TCN-Sk behaves in a more "reluctant" way as compared with CNN-TCN-ROI which takes more "risky" decisions, thus being prone to more false positive predictions. This can be interpreted by the fact that vibrato induces a lot of irrelevant motion which can be quite challenging for the corresponding classifier to discriminate from fingering transitions.

We also present the average results for each different instrument. Both models performed best in the case of violins where TCN-Sk reached the F measure value of 0.846. Both models exhibited their worst performance in the case of cellos. Each distinct model yields quite average results for double bass and viola. The relatively low results given by CNN-TCN-ROI in the case of cello performances can again be explained by the extensive use of vibrato by the cellist and the small area that the left hand occupies in the downscaled 100x100 frames. This in not the case for the double bass even if the same downscaling was forced on the extracted ROIs because, in all the three videos that this instrument appears, the position of the performer in each recording is closer to the camera than usual. Finally, the aforementioned pattern with regard to Precision and Recall holds true for both models in three out of four instruments.

4.3.2 Comparison with Previous Works

The above results can be compared with the performance of models presented in previous works. Bastas et al. [29] have deployed another TCN variant (with 6 layers) on post-processed upsampled skeleton data (93.75 fps). This frame rate was chosen to match the audio spectral representations which were themselves fed to a 4-layer TCN dedicated to the aural modality.

The current method outperforms the previous visual approach by 13.9% (see Table 2). There are various reasons for this. In the current setting we avoid standardization of the input vectors across time since it can lead to a loss of information with regard to the (relative) position of each keypoint. This is because the values of features corresponding to different coordinates of distinct body joints are forced to acquire a zero mean value across time. By avoiding standardization across time we also ensure intact ranges of motion of each keypoint, since we avoid the enforcement of a common standard deviation (i.e. $\sigma = 1$). What proved to be important is also the size of the receptive field which, in the previous method, is quite small (only 0.68 sec). Finally, we should also consider as a boosting factor the current configuration of the TCN-Sk which is more similar to the Wavenet architecture since it involves gated activation units. The results obtained by the pixel data are very promising as well. Even though the onset detection from fingers is more difficult, CNN-TCN-ROI yields results less than 4% worse than the results obtained from the skeletons with the previous method.

To obtain a more comprehensive view on the onset detection task, we compared our method to state-of-the-art methods applied on audio. The CNN model presented in [23] is trained on spectral representations from various audio excerpts. It is packaged in the madmom library and so it can be easily tested on the URMP Dataset. The TCN used in [29] is trained and tested using 8-fold cross-validation on the URMP Dataset. The results of TCN-Sk are less than 11% lower than the ones obtained by the CNN and 14.2% lower than the TCN that uses the aural modality as input. Finally, we notice a slightly greater standard deviation for the two newly introduced methods, which indicates better adaptation on a certain subset of the data, naturally on different instruments.

Instrument	Precision	Recall	F measure
Skeletons (TCN-Sk)			
Violin	0.867	0.833	0.846
Viola	0.785	0.722	0.746
Cello	0.655	0.596	0.620
Double Bass	0.730	0.734	0.732
Total	0.806	0.762	**0.779**
Hand ROIs (CNN-TCN-ROI)			
Violin	0.685	0.737	0.705
Viola	0.581	0.587	0.574
Cello	0.390	0.324	0.349
Double Bass	0.544	0.598	0.567
Total	0.604	0.622	**0.606**

Table 1. Precision, Recall and F measure results for the two proposed models (TCN-Sk and CNN-TCN-ROI). Separate average measurements for each instrument and total results after 8-fold cross-validation.

Model	F measure
TCN-Sk	0.779±0.079
CNN-TCN-ROI	0.606±0.092
TCN on Skeletons (Bastas et. al [29])	0.640±0.058
TCN on Audio (Bastas et. al [29])	0.921±0.018
CNN on Audio (Schlüter and Böck [23])	0.886±0.012

Table 2. Mean value and standard deviation of the F measure results for different tested methods on URMP [28].

Model	Recall
TCN-Sk	0.762
CNN-TCN-ROI	0.622
Combined True Positives	0.837

Table 3. Recall results of the separate models and of their combined yielded true positives.

4.3.3 Complementary Outputs

The question arises about whether these two methods capture complementary information or not. For this purpose, we assigned for each annotated onset a True label if it was indeed detected and False if it was not. We did this for all the string performances, for both the skeleton and the pixel model. The union of the Boolean labels outputed by the two models for each performance yields the combined true positive predictions. If the number of true positives or their proportionate occurrence (i.e. the Recall) would remain the same as the one of the best performing model, then there would not be any complementary information captured by the two approaches. However, the Recall value that results from the previous procedure is found to be higher by 7.5% than the Recall of TCN-Sk as depicted in Table 3. This finding is in agreement with the fact that note articulation might involve only fingering transitions (i.e. legatos) or only bow strokes (i.e. same note played). It also brings to light the distinct value and possibilities of each separate approach.

5. CONCLUSION

Both visual onset detection approaches (i.e. the one relying on skeleton features and the one relying on pixel data) proved successful in capturing information pertinent to this task. The subtlety of the fingering transitions on the instrument's neck is shown to pose greater difficulties for a model to grasp. However, the complementary information that can be captured with this method should be considered of great value for an enhancement of the overall performance on the task. Hence, as future work, one priority would be the development of a fusion method which would be able to efficiently combine the information captured from the two different information sources. Our preliminary experiments [29] on fusioning skeleton data with audio, pave the way for advancing new fusion methods. Multi-modal fusion involving trainable models was proved beneficial also in tasks like speech recognition [36]. One interesting direction would also be to study the possibility of an approach independent of a skeleton extraction phase. One reason for this is to reduce the time spent in the inference procedure by making predictions relying directly on the pixel data.

Acknowledgments

We would like to thank a lot our dear colleague Kosmas Kritsis for his help organizing some of the material and his valuable scientific comments.

6. REFERENCES

[1] F. Platz and R. Kopiez, "When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 1, pp. 71–83, 2012.

[2] C.-J. Tsay, "Sight over sound in the judgment of music performance," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 580–14 585, 2013.

[3] W. Thompson, P. Graham, and F. Russo, "Seeing music performance: Visual influences on perception and experience," *Semiotica*, vol. 2005, no. 156, pp. 203–227, 2005.

[4] Z. Duan, S. Essid, C. C. Liem, G. Richard, and G. Sharma, "Audiovisual analysis of music performances: Overview of an emerging field," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 63–73, 2018.

[5] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6–10.

[6] A.-M. Burns and M. M. Wanderley, "Visual methods for the retrieval of guitarist fingering," in *Proceedings of the 2006 conference on New interfaces for musical expression*. Citeseer, 2006, pp. 196–199.

[7] C. Kerdvibulvech and H. Saito, "Vision-based guitarist fingering tracking using a bayesian classifier and particle filters," in *Pacific-rim symposium on image and video technology*. Springer, 2007, pp. 625–638.

[8] Z. Wang and J. Ohya, "Tracking the guitarist's fingers as well as recognizing pressed chords from a video sequence," *Electronic Imaging*, vol. 2016, no. 15, pp. 1–6, 2016.

[9] J. Scarr and R. Green, "Retrieval of guitarist fingering information using computer vision," in *2010 25th International Conference of Image and Vision Computing New Zealand*. IEEE, 2010, pp. 1–7.

[10] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 93–96.

[11] B. Duke and A. Salgian, "Guitar tablature generation using computer vision," in *International Symposium on Visual Computing*. Springer, 2019, pp. 247–257.

[12] S. Goldstein and Y. Moses, "Guitar music transcription from silent video." in *BMVC*, 2018, p. 309.

[13] D. O. Gorodnichy and A. Yogeswaran, "Detection and tracking of pianist hands and fingers," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*. IEEE, 2006, pp. 63–63.

[14] A. Oka and M. Hashimoto, "Marker-less piano fingering recognition using sequential depth images," in *The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision*. IEEE, 2013, pp. 1–4.

[15] B. Zhang and Y. Wang, "Automatic music transcription using audio-visual fusion for violin practice in home environment," 2009.

[16] O. Slizovskaia, E. Gómez Gutiérrez, and G. Haro Ortega, "Correspondence between audio and visual deep models for musical instrument detection in video recordings," 2017.

[17] O. Slizovskaia, E. Gómez, and G. Haro, "Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017, pp. 226–232.

[18] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[19] F. Pedersoli and M. Goto, "Dance beat tracking from visual information alone."

[20] A. Zlatintsi, P. P. Filntisis, C. Garoufis, A. Tsiami, K. Kritsis, M. A. Kaliakatsos-Papakostas, A. Gkiokas, V. Katsouros, and P. Maragos, "A web-based real-time kinect application for gestural interaction with virtual musical instruments," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, 2018, pp. 1–6.

[21] K. Kritsis, A. Gkiokas, M. Kaliakatsos-Papakostas, V. Katsouros, and A. Pikrakis, "Deployment of lstms for real-time hand gesture interaction of 3d virtual music instruments with a leap motion sensor," in *Proceeding of the 15th Sound and Music Computing Conference (SMC2018)*, 2018, pp. 331–338.

[22] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 283–292.

[23] J. Schlüter and S. Böck, "Musical onset detection with convolutional neural networks," in *6th international workshop on machine learning and music (MML), Prague, Czech Republic*, 2013.

[24] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long-short term memory neural networks," in *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands*, 2010, pp. 589–594.

[25] B. Li, K. Dinesh, Z. Duan, and G. Sharma, "See and listen: Score-informed association of sound tracks to players in chamber music performance videos," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2906–2910.

[26] B. Li, C. Xu, and Z. Duan, "Audiovisual source association for string ensembles through multi-modal vibrato analysis," *Proc. Sound and Music Computing (SMC)*, 2017.

[27] B. Li, K. Dinesh, C. Xu, G. Sharma, and Z. Duan, "Online audio-visual source association for chamber music performances," *Transactions of the International Society for Music Information Retrieval*, vol. 2, no. 1, 2019.

[28] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.

[29] G. Bastas, A. Gkiokas, V. Katsouros, and P. Maragos, "Improving audio onset detection for string instruments by incorporating visual modality," *MML 2020*, p. 32.

[30] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[31] B. Li, A. Maezawa, and Z. Duan, "Skeleton plays piano: Online generation of pianist body movements from midi performance." in *ISMIR*, 2018, pp. 218–224.

[32] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[33] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[34] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[35] Visual onset detection from bowed strings. [Online]. Available: https://github.com/gbastas/avOnset.git

[36] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.