Vision-based Detection of Acoustic Timed Events: a Case Study on Clarinet Note Onsets

A. Bazzica*[1], J.C. van Gemert[2], C.C.S. Liem[1], and A. Hanjalic[1]

[1]Multimedia Computing Group - Delft University of Technology, The Netherlands
[2]Vision Lab - Delft University of Technology, The Netherlands

Acoustic events often have a visual counterpart. Knowledge of visual information can aid the understanding of complex auditory scenes, even when only a stereo mixdown is available in the audio domain, e.g., identifying which musicians are playing in large musical ensembles. In this paper, we consider a vision-based approach to note onset detection. As a case study we focus on challenging, real-world clarinetist videos and carry out preliminary experiments on a 3D convolutional neural network based on multiple streams and purposely avoiding temporal pooling. We release an audiovisual dataset with 4.5 hours of clarinetist videos together with cleaned annotations which include about 36,000 onsets and the coordinates for a number of salient points and regions of interest. By performing several training trials on our dataset, we learned that the problem is challenging. We found that the CNN model is highly sensitive to the optimization algorithm and hyper-parameters, and that treating the problem as binary classification may prevent the joint optimization of precision and recall. To encourage further research, we publicly share our dataset, annotations and all models and detail which issues we came across during our preliminary experiments.
Keywords: computer vision, cross-modal, audio onset detection, multiple-stream, event detection

1 Introduction

Acoustic timed events take place when persons or objects make sound, e.g., when someone speaks or a musician plays a note. Frequently, such events also are visible: a speaker's lips move, and a guitar cord is plucked. Using visual information we can link sounds to items or people and can distinguish between sources when multiple acoustic events have different origins. We then can also interpret our environment in smarter ways: e.g., identifying the current speaker, and indicating which instruments are playing in an ensemble performance.

Understanding scenes through sound and vision has both a *multimodal* and a *cross-modal* nature. The former allows us to recognize events using auditory and visual stimuli jointly. But when e.g., observing a door bell button being pushed, we can cross-modally infer that a bell should ring. In this paper, we focus on the cross-modal case to detect acoustic timed events from video. Through visual segmentation, we can spatially isolate and analyze sound-making sources at the individual player level, which is much harder in the audio domain [2].

*alessio.bazzica@gmail.com (now at Google)

As a case study, we tackle the musical note onset detection problem by analyzing clarinetist videos. Our interest in this problem is motivated by the difficulty of detecting onsets in audio recordings of large (symphonic) ensembles. Even for multi-track recordings, microphones will also capture sound from nearby instruments, making it hard to correctly link onsets to the correct instrumental part using audio alone. Knowing where note onsets are and to which part they belong is useful for solving several real-world applications, like audio-to-score alignment, informed source separation, and automatic music transcription.

Recent work on cross-modal lip reading recognition [5] shows the benefit of exploiting video for a task that has traditionally been solved only using audio. In [11], note onset matches between a synchronized score and a video are used to automatically link audio tracks and musicians appearing in a video. The authors show a strong correlation between visual and audio onsets for bow strokes. However, while this type of visual onset is suitable for strings, it does not correlate well to wind instruments. In our work we make an important step towards visual onset detection in realistic multi-instrument settings focusing on visual information from clarinets, which has sound producing interactions (blowing, triggering valves, opening/closing holes) representative for wind instruments in general.

Our contributions are as follows: (i) defining the visual onset detection problem, (ii) building a novel 3D convolutional neural network (CNN) [14] without temporal pooling and with dedicated streams for several regions of interest (ROIs), (iii) introducing a novel audiovisual dataset of 4.5 hours with about 36k annotated events, and (iv) assessing the current gap between vision-based and audio-based onset detection performance.

2 Related work

When a single instrument is recorded in isolation, audio onset detectors can be used. A popular choice is [13], which is based on learning time-frequency filters through a CNN applied to the spectrogram of a single-instrument recording. While state-of-the-art performance is near-perfect, audio-only onset detectors are not trained to handle multiple-instrument cases. To the best of our knowledge, such cases also have not been tackled so far.

A multimodal approach [1] spots independent audio sources, isolates their sounds and is validated on four audiovisual sequences with two independent sources. As the authors state [1], their multimodal strategy is not applicable in crowded scenes with frequent audio onsets. Therefore, it is not suitable when multiple instruments mix down into a single audio track.

A cross-modal approach [4] uses vision to retrieve guitarist fingering gestures. An audiovisual dataset for drum track transcription is presented in [9] and [6] addresses audiovisual multi-pitch analysis for string ensembles. All works devise specific visual analysis methods for each type of instrument, but do not consider transcription or onset detection for clarinets.

Action recognition aims to understand events. Solutions based on 3D convolutions [14] use frame sequences to learn spatio-temporal filters, whereas two-streams networks [8] add a temporal optical flow stream. A recurrent network [7] uses LSTM units on top of 2D convolutional networks. While action recognition is similar to visual-based acoustic timed events detection, there is a fundamental difference: action recognition aims to detect the presence or absence of an action in a video. Instead, we are interested in the exact temporal location of the onset.

In action localization [12] the task is to find what, when, and where an action happens. This is modeled with a "spatio-temporal tube": a list of bounding-boxes over frames. Instead, we are not interested in the spatial location; we aim for the temporal location only, which due to the high-speed nature of onsets reverts to the extreme case of a single temporal point.

3 Proposed baseline method

Together with our dataset, we offer a baseline model for onset detection. The input for our model is a set of sequences generated by tracking a number of oriented ROIs from a video of a single clarinetist (see Figure 1). For now, as a baseline, we assume that in case of a multi-player ensemble, segmentation of individual players already took place. The ROIs consider those areas in which the sound producing interactions take place: mouth, left/right hands, and clarinet tip, since they are related to blowing, fingering, and lever movements respectively.

Figure 1: Raw video frames example.

Each sequence is labeled by determining if a note has started during the time span of the *reference frame*. A sequence consists of 5 preceding frames, the reference frame, and 3 succeeding frames, forming a sequence of 9 consecutive frames per ROI. We use a shorter future temporal context because the detector may otherwise get confused by *anticipation* (getting ready for the next note). Examples of onset and not-an-onset inputs are shown in Figure 2.

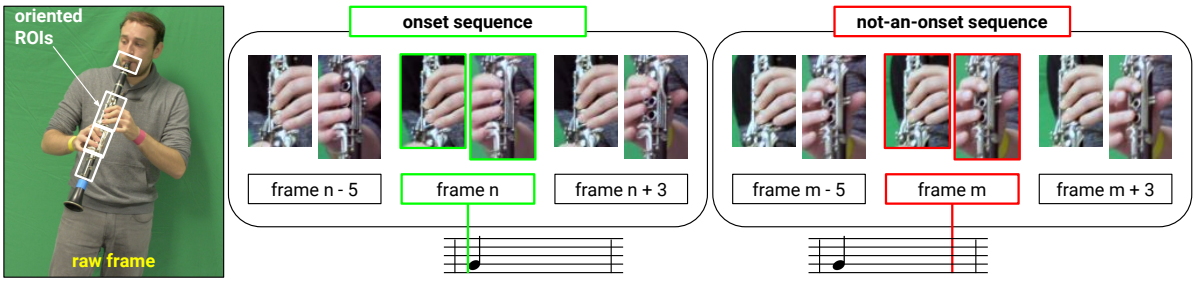

Figure 2: Onset and not-an-onset input sequence examples with 2 ROIs from 3 frames.

Our model relies on multiple *streams*, one for each ROI. Each stream consists of 5 convolutional layers (CONV1-5), with a fully-connected layer on top (FC1). All the FC1 layers are concatenated and linked to a global fully-connected layer (FC2). All the layers use ReLU units. The output consists of two units ("not-an-onset" and "onset"). Figure 3 illustrates our model and, for simplicity, it only shows one stream for the left hand and one for the right one.

To achieve the highest possible temporal resolution, we do not use temporal pooling. We use spatial pooling and padding parameters to achieve slow fusion throughout the 5 convolutional layers. We aim to improve convergence and achieve regularization using batch normalization (BN) [10], L2 regularization and dropout. Since we use BN, we omit the bias terms in every layer including the output layer.

We use weighted cross-entropy as loss function to deal with the unbalanced labels (on average, one onset every 15 samples). The loss is minimized using the RMSprop algorithm. While training, we shuffle and balance the mini-batches. Each mini-batch has 24 samples, half of which are not-an-onset ones, 25% onsets and 25% *near-onsets*, where a near-onset is a sample adjacent to an onset. Near-onset targets are set to $(0.75, 0.25)$, i.e., the non-onset probability is 0.75. In this way, a near-onset predicted as onset is penalized less than a false positive. We also use data augmentation (DA) by randomly cropping each ROI from each sequence. By combining DA and balancing, we obtain epochs with about 450,000 samples. Finally, we manually use early-stopping to select the check-point to be evaluated (max. 15 epochs).

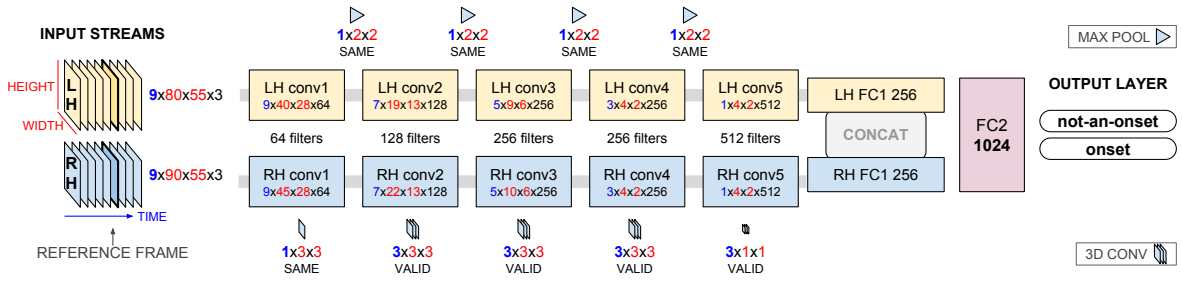

Figure 3: Proposed model based on 3D CNNs, slow fusion, and multiple streams (one for each ROI). LH and RH indicate the left and right hand streams respectively.

4 Experimental testbed: Clarinetists for Science dataset

We acquired and annotated the new *Clarinetists for Science* (C4S) dataset, released with this paper[1]. C4S consists of 54 videos from 9 distinct clarinetists, each performing 3 different classical music pieces twice (4.5h in total). The videos have been recorder at 30 fps, about 36,000 events have been semi-automatically annotated and thoroughly checked. We used a colored marker on the clarinet to facilitate visual annotation, and a green screen to allow for background augmentation in future work. Besides ground-truth onsets, we include coordinates for face landmarks and 4 ROIs: mouth, left hand, right hand, and clarinet tip.

In our experiments, we use leave-one-subject-out cross validation to validate the generalization power across different musicians (9 splits in total). From each split, we derive the training, validation and test sets from 7, 1, and 1 musicians respectively. Hyper-parameters, like decaying learning rate and L2 regularization factors, are manually adjusted looking at f-scores and loss for train and validation sets. We compute the f-scores using 50 ms as temporal tolerance to accept a predicted onset as true positive. We compare to a ground-truth informed random baseline (correct number of onsets known) and to two state-of-the-art audio-only onset detectors (namely, SuperFlux [3] and CNN-based [13]).

5 Results and discussion

During our preliminary experiments, most of the training trials were used to select optimization algorithm and suitable hyper-parameters. Initially, gradients were vanishing, most of the neurons were inactive, and networks were only learning bias terms. After finding hyper-parameters overcoming the aforementioned issues, we trained our model on 2 splits.

method	Split 1	Split 2	Average
informed random baseline	27.4	19.6	23.5
audio-only SuperFlux [3]	82.8	81.3	82.1
audio-only CNN [13]	94.3	92.1	93.2
visual-based (proposed)	26.3	25.0	25.7

Table 1: F-scores with a temporal tolerance of 50 ms.

By inspecting the f-scores in Table 1, we see that our method only performs slightly better than the baseline, and that the gap between audio-only and visual-based methods is large (60% on average). We investigated why and found that throughout the training, precision and recall

[1]For details, examples, and downloading see `http://mmc.tudelft.nl/users/alessio-bazzica#C4S-dataset`

often oscillate with a negative correlation. This means that our model struggles with jointly optimizing those scores. This issue could be alleviated by different near-onsets options or by formulating a regression problem instead of a binary classification one.

When we train on other splits, we observe initial f-scores not changing throughout the epochs. We also observe different speeds at which the loss function converges. The different behaviors across the splits may indicate that alternative initialization strategies should be considered and that the hyper-parameters are split-dependent.

6 Conclusions

We have presented a novel cross-modal way to solve note onset detection visually. In our preliminary experiments, we faced several challenges and learned that our model is highly sensitive to initialization, optimization algorithm and hyper-parameters. Also, using a binary classification approach may prevent the joint optimization of precision and recall. To allow further research, we release our novel fully-annotated C4S dataset. Beyond visual onset detection, C4S data will also be useful for clarinet tracking, body pose estimation, and ancillary movement analysis.

Acknowledgments

We thank the C4S clarinetists, Bochen Li, Sara Cazzanelli, Marijke Schaap, Ruud de Jong, dr. Michael Riegler and the SURFsara Dutch National Cluster team for their support in enabling the experiments of this paper.

References

[1] Z. Barzelay and Y. Y. Schechner. Onsets coincidence for cross-modal analysis. *IEEE TMM*, 12(2), 2010.

[2] A. Bazzica, C.C.S. Liem, and A. Hanjalic. On detecting the playing/non-playing activity of musicians in symphonic music videos. *CVIU*, 144, March 2016.

[3] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *DAFx*, 2013.

[4] A.M. Burns and M.M. Wanderley. Visual methods for the retrieval of guitarist fingering. In *NIME*, 2006.

[5] J.S. Chung, A.W. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. *arXiv preprint arXiv:1611.05358*, 2016.

[6] K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma. Visually informed multi-pitch analysis of string ensembles. In *ICASSP*, 2017.

[7] J. Donahue, L.A., S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[9] O. Gillet and G. Richard. ENST-Drums: an extensive audio-visual database for drum signals processing. In *ISMIR*, 2006.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[11] 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 *ICASSP*, 2017.

[12] P. Mettes, J.C van Gemert, and C.G.M. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016.

[13] J. Schlüter and S. Böck. Improved musical onset detection with convolutional neural networks. In *ICASSP*, 2014.

[14] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015.