Visual Pitch Estimation

A. Sophia Koepke
University of Oxford
koepke@robots.ox.ac.uk

Olivia Wiles
University of Oxford
ow@robots.ox.ac.uk

Andrew Zisserman
University of Oxford
az@robots.ox.ac.uk

ABSTRACT

In this work, we propose the task of automatically estimating pitch (fundamental frequency) from video frames of violin playing using *vision* alone. Here, we consider only monophonic violin playing (where only one note is being played at a time).
In order to investigate this task, we curate a new dataset of monophonic violin playing. We propose a Convolutional Neural Network (CNN) architecture that is trained using a student-teacher strategy to distil knowledge from the audio domain to the visual domain. At test time, our network takes video frames as input and directly regresses the pitch. We train and test this architecture on different subsets of our new dataset.
We show that this task (i.e. pitch prediction from vision) is actually possible. Furthermore, we verify that the network has indeed learnt to focus on salient parts of the image, e.g. the left hand of the violin player is used as a visual cue to estimate pitch.

1. INTRODUCTION

Humans can obtain some understanding of music simply by watching instruments being played, even without access to audio recordings of the music itself. Indeed, a trained musician might be able to transcribe an entire video purely from visual cues alone, although with great painstaking manual effort. The movement and position of the instrument and body (specifically the movement of the arms, hands and fingers) have a direct correlation with the sound produced. In this work, we investigate the following question: is it possible for a trained neural network to identify the pitch of played notes, simply from the frames of a silent video?

Our approach is a valuable first step towards the task of *complete* visual music transcription. While *audio* based music transcription is a widely studied and successful field, the task of *visual* music transcription has not been explored to a great extent. Performing this task from standard frame-rate visual information alone can be extremely useful in instances when the audio is of poor quality, missing, or mixed with information from other audio sources, e.g. in the case of polyphonic music. These scenarios are challenging for purely audio-based pitch estimation methods.

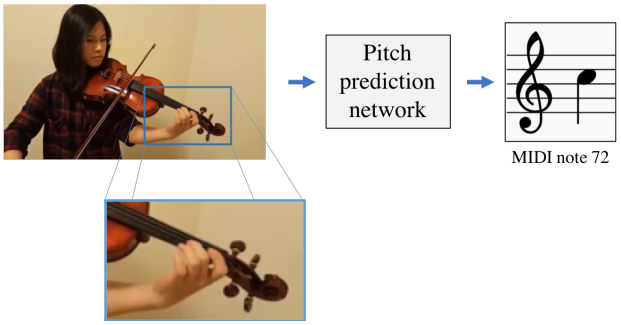


Figure 1. Pitch estimation from visual information. Given video frames, the network is tasked to predict pitch using *only* the visual information.

We investigate this by training a network to predict pitch information from video frames of monophonic solo violin recordings using only the visual image data (see Figure 1). Given a set of video frames, the network learns to regress the corresponding pitch. In order to perform this challenging task, our method makes use of two insights. First, using a teacher-student strategy (i.e. training one network using another network [1]) is important to enforce that the network learns the visual cues that are correlated with the corresponding sound. Second, using multiple frames as input (i.e. a short silent video clip) is preferable to using a single still frame. This is because the additional frames resolve ambiguities such as which string is vibrating (i.e. the string that is being played on with the bow). These insights inform our architecture choices, described in Section 3.

The models are trained and evaluated on a new dataset (Section 4) of violin playing. This dataset is divided into three subsets which vary in difficulty. The first two subsets are recordings of a single player photographed by a fixed mobile phone camera. The third subset consists of 'in-the-wild' videos downloaded from YouTube.

On all of these datasets, our method demonstrates that regressing pitch directly from video frames is indeed possible (Section 5). Finally, we verify that the method is making sensible predictions by investigating what regions of the image are most salient for the prediction. We find that our method focusses on the movement and location of the musician's arms, hands and fingers; this is similar to how a human would approach this task.

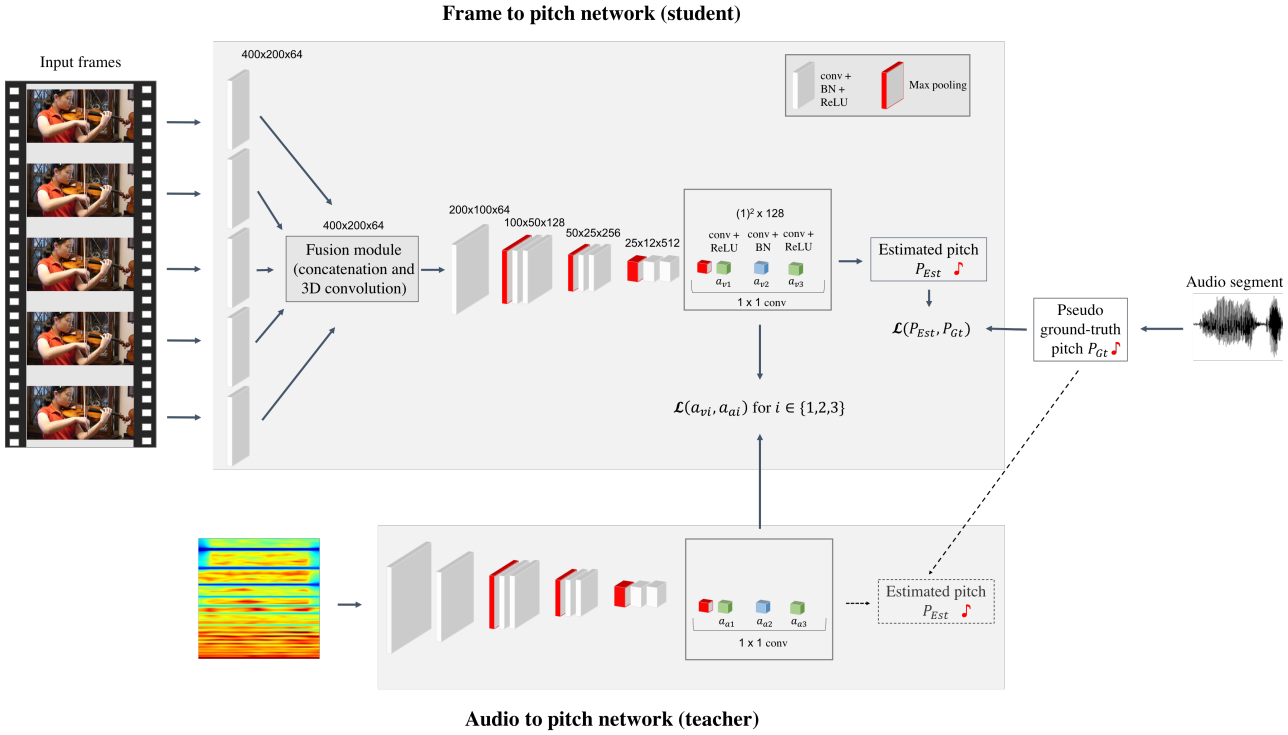


Figure 2. An overview of the visual pitch estimation method. We train our framework with a student-teacher strategy by distilling discriminative knowledge from a teacher network to a student network. The teacher network regresses pitch from audio whereas the student network is trained to regress pitch from visual information alone. Both networks are trained using pseudo ground-truth pitch information which is automatically extracted using an audio-based method and thereby does not require any manual annotation. The audio to pitch network is trained first and then used to train the frame to pitch network (student) by minimising the distance between the activations of the final three fully-connected network layers of the student and the teacher network. At test time, given one or multiple visual input frames, the student network is used by itself to regress pitch. In the case of multiple input frames, the outputs of the first convolutional layers of the student network are concatenated and fused through a 3D convolutional layer (Fusion module) before being input to the next convolutional layer.

2. RELATED WORK

Here, we only consider directly related work on cross-modal information transfer between audio and visual information.

Multi-modal audio-visual representations. Training strategies that encourage synchronisation between the audio and visual streams have been used successfully for speech synchronisation [2]. More generally the correspondence between the audio and visual streams (though not their strict synchronisation) has proven very useful for obtaining meaningful features for sound localisation and separation [3, 4]. In these works, the natural synchronisation in videos can be leveraged in a self-supervised manner to obtain useful image and audio representations. Aytar et al. [5] also exploit this natural synchronisation property in order to transfer knowledge from visual recognition networks into sound networks. However, we propose a framework that transfers knowledge the other way round, i.e. from audio to visual information.

Cross-modal audio-visual generation. Related to our framework are methods that generate audio from visual information, e.g. spectrograms or other sound features from visual information [6–10], or localise sound in video in order to separate different sounds [11, 12], or analyse vibrato-patterns for audio-visual association [13].

The URMP dataset by Li et al. [14] is targeted at cross-modal audio-visual generation. However, it poses two limitations for our task. Firstly, the image resolution of the released dataset is not very high which makes it difficult to actually recognise pitch from the visual information alone, as the key parts of the image (e.g. the fingers of the left hand) are too small. Secondly, the dataset was recorded in constrained settings with only a limited number of musicians which limits the generalisability of models trained on this data to other settings. Therefore, in addition to training and evaluating our models on the URMP dataset, we gathered a new dataset to train and test our framework that is of higher resolution and which also contains 'in-the-wild' videos.

Music transcription from silent video. More closely related to ours is the work by Gomez et al. [15] which proposes to leverage visual information to transcribe clarinet videos using the hand movements in recorded video sequences. However, unlike their method, we do not require any manual tracking or labelling (i.e. finger/hole positions) as supervision in order to train our network.

Zhang et al. [16] addressed a similar task to ours of visu-

ally obtaining pitch for violin by detecting the strings of a violin and by recognising finger events (such as their position and whether they are pressing on a string). However, their method is quite constrained; it involves tracking the fingers and the strings, and makes assumptions about the length of the fingerboard which requires the image data to always be perfectly aligned. In contrast, our method gives convincing results for different viewpoints and requires no manual labelling.

Another related method is the physics-based approach for recovering pitch from silent guitar video by Goldstein and Moses [17]. However, their method requires mounting a camera, that allows recording with high frame rates, on the guitar itself in order to use the actual string vibrations to predict pitch. Unlike their method, our set-up only requires the use of a normal camera and it learns to localise the left-hand position of the musician (relative to the instrument) in order to infer pitch. Our method can thus be applied retroactively to videos that have already been recorded.

3. MODEL

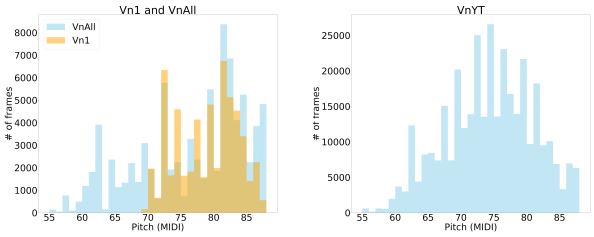
In this section, we describe the training and testing framework used to regress pitch from video frames. We treat this as a classification task. The network takes video frames as input and estimates the pitch as a MIDI number. An overview is given in Figure 2.

Teacher-student strategy. We found that directly regressing pitch from the video frames did not generalise at test time. This is presumably because the visual information relevant for the pitch prediction task occupies only a small part in the video frames.

As a result, we train two networks – a teacher and student network – such that the activations of the student network are similar to those of the teacher. The teacher network regresses pitch from audio and the student regresses pitch from video frames. The rationale for using this strategy is that, in order to contain relevant information about pitch, the high-level representation of the visual information (encoded in the student) should be close to that of the audio information (encoded in the teacher). This strategy proved crucial to obtain a network that generalises at test time.

The teacher network is first trained using STFT spectrograms as input to regress the pseudo ground-truth pitch (the method for obtain this pseudo ground-truth is described in section 5). The student network is then trained to regress the pitch with an additional loss that enforces that the activations of the higher level layers are similar to those in the teacher network. For this, we use an L1 loss which is weighed so that the contribution for each of the three fully-connected layers is as big as the pitch classification loss. Both networks are trained to predict pitch with a cross-entropy loss.

Neural Network Architecture. The teacher and student network architectures are loosely based on the VGG-M network architecture [18] and can be seen in more detail in Figure 2. For the student network, in the case of multiple input frames, the outputs of the first convolutional layers are concatenated and fused using a 3D convolutional layer to combine the information from the frames with spatial kernel size 3×3 followed by batch normalization and ReLU. The output serves as input to the second convolutional layer. For just a single input frame, the output of the first (2D) convolutional layer is directly input to the second convolutional layer. The first two convolutional layers consist of 7×7 convolutions, whereas the subsequent ones are 3×3 convolutions and the last three are 1×1 convolutions. All convolutional layers have stride 1 except for the second one, which has stride 2.



(a) *Vn1* and *VnAll* datasets. (b) *VnYT* dataset.

Figure 3. Pitch distribution over the number of frames in the three subsets of our dataset; the constrained single-string data *Vn1*, the data on all strings *VnAll*, and the in-the-wild data *VnYT*. Pitch is shown in MIDI numbers. The subsets cover the chosen full pitch ranges.

4. DATASETS

We curate a new violin playing dataset which consists of three subsets (*Vn1*, *VnAll* and *VnYT*) that differ in difficulty and size. The most challenging subset *VnYT* consists of in-the-wild violin solo videos downloaded from YouTube [1]. These are largely comprised of recordings of solo recitals, etudes, and orchestra auditions. Both *Vn1* and *VnAll* are recorded in simpler conditions: all videos are of a single violinist, have similar backgrounds and are taken from similar angles.

The datasets vary in terms of the range of pitches. Both *Vn1* and *VnAll* consist of recordings of a violinist playing in a practise-like set-up (but without the thousandfold repetitions of the same phrases). The easiest subset *Vn1* consists of videos that only contain violin playing on a single string, resulting in a range of 20 semitones (MIDI numbers between 68 and 88). Estimating the pitch is easier in this case, as there is no ambiguity concerning which string is being played. *VnAll* contains videos played in the full pitch range of the violin without being restricted to playing just on one string. For both *VnAll* and the most difficult subset *VnYT*, we consider a range of 33 semitones (MIDI numbers between 55 and 88).

All subsets are split into train/val/test sets. Disjoint parts of the same videos are used for training and validation. The test sets consist of frames that were not seen during training (from left-out unseen videos). The precise numbers of frames and videos are given in Table 1.

[1] Example videos: https://youtu.be/-ccYdhQAn10, https://youtu.be/YGCYelAHdaU

Dataset *Vn1*			
	Train	Val	Test
# of videos	5		1
# of frames	25308	2791	9875
Dataset *VnAll*			
	Train	Val	Test
# of videos	9		1
# of frames	54865	4107	6373
Dataset *VnYT*			
	Train	Val	Test
# of videos	122		10
# of frames	303391	33063	20067

Table 1. Dataset statistics. Details of the three different subsets, the controlled setting data on a single string *Vn1*, the controlled setting data on all strings *VnAll*, and the in-the-wild data on all strings *VnYT* and their respective train/val/test splits.

Network	RPA	RPA tol	PA	ACA	ACE
	Dataset *Vn1*				
Audio to pitch	98.30	99.14	96.74	86.03	0.06
Frame to pitch	89.98	91.57	62.41	51.64	0.45
5 fr. to pitch (3D conv)	93.8	94.91	66.7	58.75	0.43
	Dataset *VnAll*				
Audio to pitch	94.26	94.40	90.87	94.33	0.06
Frame to pitch	74.17	75.55	47.48	33.3	2.50
5 fr. to pitch (3D conv)	77.24	78.98	50.33	41.66	1.65
	Dataset *VnYT*				
Audio to pitch	98.30	99.14	96.74	86.03	0.06
Frame to pitch	44.3	51.37	33.18	45.2	2.50
5 fr. to pitch (3D conv)	62.5	67.89	48.44	51.77	2.34
	Dataset URMP				
Audio to pitch	98.28	98.5	96.73	98.88	0.07
Frame to pitch	53.11	58.3	42.71	39.86	2.73
5 fr. to pitch (3D conv)	57.3	62.04	45.26	41.79	2.43

Table 2. Evaluation of our models determining the accuracy in predicted pitch for the *Vn1*, *VnAll*, *VnYT* , and URMP test sets. Higher is better for Raw Pitch Accuracy (RPA), Raw Pitch Accuracy with a tolerance of one frame (RPA tol), Pitch Accuracy (PA), and Average Class Accuracy (ACA). Lower is better for Average Class Error (ACE). Using multiple input frames improves the performance.

Finally, for all three subsets, we extract frames and pseudo ground-truth pitch using the spectral domain YIN algorithm [19] using the implementation in the *aubio* library (yinfft) [20].

In addition to the above datasets, we consider the subset of the URMP dataset that contains videos of violin playing. We loosely crop the frames around the violinist and leave out 4 videos (single-instrument tracks) for testing and take the remaining violin videos for training and validation. This results in about 35000 frames for training and 12761 for testing. This dataset contains ground-truth pitch information. Therefore, we can train with actual ground-truth pitch. We consider a range of 33 semitones (MIDI numbers between 55 and 88). All models are trained and tested with the same train/val/test split on each dataset.

Furthermore, we generate STFT spectrograms for all mentioned datasets in order to train the audio to pitch network.

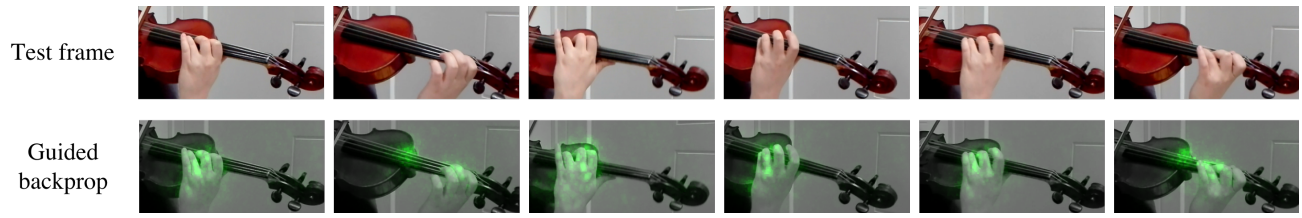
5. EXPERIMENTS

In this section, we evaluate both the audio to pitch (teacher), and the video frame to pitch (student) models. We consider using a single versus multiple input video frames. We first train the audio to pitch network to regress pitch from spectrograms. This network then serves as the teacher network when training the single frame to pitch network or the multi-frame to pitch network.

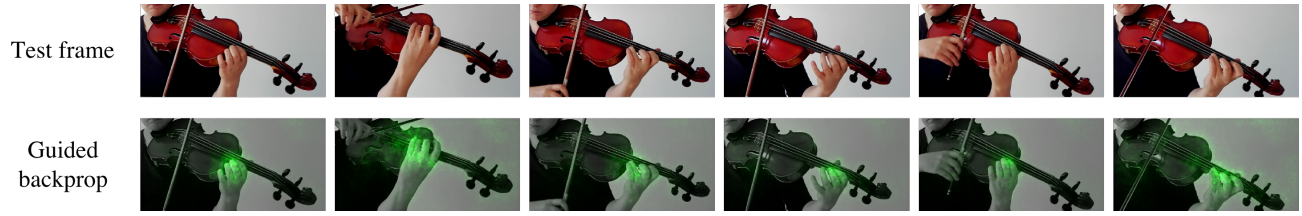
The models are trained in PyTorch [21] using the Adam optimiser [22] with an initial learning rate of 0.001. The learning rate is divided by a factor of 10 when the loss on the validation set plateaus. The batchsize is $N = 64$ for the single frame architecture and the audio to pitch network, and $N = 24$ for the architecture with 5 input frames. The frames are resized to 400×200. For the datasets *Vn1* and *VnAll*, the frames are consistently more tightly cropped around the instrument whereas there is much more variation of the location and relative size of the instrument in *VnYT*.

Evaluation measures. We report the performance of our models in Table 2. For Raw Pitch Accuracy (*RPA*), a predicted pitch is counted as correctly estimated if it lies within one semitone of the ground truth pitch. *RPA tol* additionally allows the prediction to be off by at most one frame. Furthermore, we report Pitch Accuracy (*PA*) and Average Class Accuracy (*ACA*). ACA gives the averaged per-pitch-class accuracy. The ACE describes the average error between the predicted and the ground truth pitch class (ACE of 1 corresponds to an average error of one semitone).

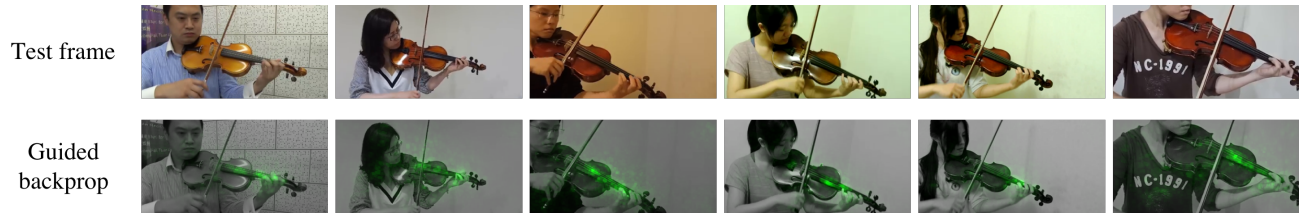
Video to pitch performance. The audio to pitch teacher networks reach an RPA of above 90% on the test sets. This serves as a very good starting point to train the student frame to pitch networks. It can be observed that our method performs best when trained and tested on the simpler dataset with minimal ambiguities *Vn1* and then *VnAll*. This corresponds with the intuition that this set-up is easier, as the fingers and therefore the pitch is more clearly visible at higher resolution as compared to *VnYT* or URMP. Nevertheless, a RPA of 62.5% for the frame to pitch network on the in-the-wild YouTube video dataset *VnYT* means that the pitch is estimated within a semitone of the ground-truth on average in 62.5% of the test cases; this verifies that our method generalises to unseen videos and people at test time on challenging 'in-the-wild' videos. When allowing for an offset of one frame in the predictions, we achieve an accuracy of 67.89% (RPA tol). This accounts for the case that the alignment between audio and visual information might not be perfect in the data which is the case for some of the downloaded videos. The reported lower performance on the URMP dataset may be due to the lower resolution size of the frames in the dataset and the limitations in terms of its dataset size which confirms the benefits of using our datasets to address this task. These results are impressive, given that our method estimates the pitch from visual information only and in unconstrained recording conditions. However, our method can only predict one pitch playing



(a) Gradient visualisation for the *Vn1* test set.



(b) Gradient visualisation for the *VnAll* test set.



(c) Gradient visualisation for the *VnYT* test set.

Figure 4. Discriminative information visualisations using guided backpropagation [23] for the test sets of the *Vn1* subset in (a), the *VnAll* subset in (b), and the *VnYT* subset in (c). Heat maps are overlaid in the second rows of (a), (b), and (c). As demonstrated, the networks focus on the left hand across all the test frames even though the hands are in different positions relative to the frame. In (c), it can be seen that the network also seems to be focussing on the strings implying that it may be using vibrations or the movement of the strings in order to estimate pitch. The location of the instrument and strings relative to the left hand might serve as a further cue for estimating pitch.

at a time and cannot identify chords as it has been trained only on monophonic data.

Another interesting point is that there is a consistent improvement when using multiple frames as opposed to a single one as input to the frame to pitch network. This can be seen very clearly for the *VnYT* dataset (RPA tol of 67.89% vs. 51.37%). This is presumably due to the fact that visually it can be hard to determine just from the fingers of the left hand which string a note is played on. To solve this problem, the network needs to determine which string is active (using for example information from the bowing hand / bow or from the vibration of the strings). While the placement of the hand should be visible from a single image, the vibration of the string is unlikely to be visible (unless there is significant motion blur) without taking into account more frames.

Visualizing what has been learnt. To gain an insight into what the networks have learnt and how they infer the pitch from a given frame, we apply guided backpropagation [23] to our trained networks to determine which parts of the images are most discriminative. As demonstrated in Figure 4, the networks have learnt that the fingers of the left hand and the left hand itself are most relevant for predicting the pitch given a still frame. Potentially the network also makes use of some information about the vibration of the played strings (e.g. by recognising motion blur around strings that are vibrating). This confirms that the networks do not simply memorise parts of a video, but instead learn to localise the left hands/fingers in the image in order to estimate pitch. However, the image regions which the networks focus on are actually quite small relative to the image size.

6. CONCLUSION

We have presented a method for addressing monophonic visual pitch estimation; given video frames of violin playing, our method can automatically estimate the pitch being played using *vision* alone. The presented task is extremely challenging, as it requires making use of subtle visual cues (such as the placement of the hand or string vibrations over the course of multiple frames), yet our network shows convincing results in three different scenarios: when only one string is played or all strings are played but the person and environment remains the same, and in unconstrained 'in-the-wild' videos. Moreover, our method is generalisable, as training the networks did not require any manual annotations; instead, the pseudo ground-truth pitch information was extracted automatically from the audio data. It will be interesting to use this framework to improve pitch prediction using both visual and audio information. This could prove useful when the audio is of poor quality. In addition to that, estimating pitch from vision might help the task

of sound source separation when similar instruments are played on. Furthermore, this method could be pushed further to estimating polyphonic violin music played on the same instrument.

Acknowledgments

This work is supported by the EPSRC programme grant Seebibyte EP/M013774/1: Visual Search for the Era of Big Data. We are very grateful to Yael Moses for insightful discussions. We thank Arsha Nagrani for feedback.

7. REFERENCES

[1] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *2th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[2] J. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[3] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. ECCV*, 2018.

[4] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. ECCV*, 2018.

[5] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.

[6] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017.

[7] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2014.

[8] W.-L. Hao, Z. Zhang, and H. Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," *CoRR*, 2018.

[9] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. CVPR*, 2016.

[10] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *ICCV workshop*, 2017.

[11] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[12] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *Proc. ACM SIGGRAPH*, 2018.

[13] 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.

[14] 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*, 2019.

[15] E. Gómez Gutiérrez, P. Arias Martínez, P. Zinemanas, and G. Haro Ortega, "Visual music transcription of clarinet video recordings trained with audio-based labelled data," in *ICCV workshop*, 2017.

[16] B. Zhang, J. Zhu, Y. Wang, and W. K. Leow, "Visual analysis of fingering for pedagogical violin transcription," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007.

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

[18] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC.*, 2014.

[19] P. Brossier, "Automatic annotation of musical audio for interactive systems," Ph.D. dissertation, Ph. D. thesis, Queen Mary University of London, London, UK, 2006.

[20] P. Brossier, M. Hermant, E. Müller, N. Philippsen, T. Seaver, H. Fritz, and S. Alexander, "aubio/aubio: 0.4.6," Oct 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1002162

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization." *Proc. ICLR*, 2015.

[23] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *ICLR workshop*, 2015.