Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies

Olga Slizovskaia
Universitat Pompeu Fabra
olga.slizovskaia@upf.edu

Emilia Gómez
Universitat Pompeu Fabra
emilia.gomez@upf.edu

Gloria Haro
Universitat Pompeu Fabra
gloria.haro@upf.edu

ABSTRACT

The goal of this work is to incorporate the visual modality into a musical instrument recognition system. For that, we first evaluate state-of-the-art image recognition techniques in the context of music instrument recognition, using a database of about 20000 images and 12 instrument classes. We then reproduce the results of state-of-the-art methods for audio-based musical instrument recognition, considering standard datasets including more than 9000 sound excerpts and 45 instrument classes. We finally compare the accuracy and confusions in both modalities and we showcase how they can be integrated for audio-visual instrument recognition in music videos. We obtain around 0.75 F1-measure for audio and 0.77 for images and similar confusions between instruments. This study confirms that visual (shape) and acoustic (timbre) properties of music instruments are related to each other and reveals the potential of audiovisual music description systems.

1. INTRODUCTION

Human perception of music is based on integrating stimuli from various modalities, mostly from the auditory and visual domains. Nevertheless, research in music description has traditionally focused on the analysis of audio recordings, without taking account of visual information [1]. The increasing availability of music videos on the internet (ex: Youtube contains a huge amount of user-generated music performances) opens the path to incorporating visual description in several music information retrieval tasks. One of the most well-established ones is musical instrument recognition, which has been researched for decades [2].

Several of the few existing works that include the analysis of both the visual and aural components are focused on the analysis and transcription movements of performer playing on percussion instruments. Marenco et al. present a method for stroke classification in audio and video recordings of Candombe drumming [3]. They employ a feature level fusion approach on edge and color filtering for drumhead, stick and hand detection from video frames and spectral features from audio. A correlation study on visual novelty and onset intensity in video recordings of drum, guitar and bass guitar performances [4] also provides some useful insights into the image processing returns for music analysis. Perhaps, the most out of ordinary idea proposed in [5] is the use of thermal imaging to detect musical instruments, while more general ones for multimodal music content analysis (including musical instrument detection) can be found in [6].

The goal of this work is to incorporate the visual modality into a musical instrument recognition system. First, we evaluate state-of-the-art image recognition techniques in the context of music instruments. Second, we reproduce the results of a state-of-the-art method for audio-based musical instrument recognition. Third, we illustrate how both approaches are integrated for musical instrument detection.

2. AUDIO AND VISUAL METHODS

2.1 Image-based musical instrument recognition

2.1.1 Selected approach

During the last years there has been a growing interest in the use of neural networks for pattern recognition. This popularity is due to several different factors. First, methods for training very deep neural networks (even with hundreds of layers) on massive datasets using the GPU for calculations have been proposed. The second reason is the success of the convolutional neural network model overcame the ImageNet 2012 image classification contest. In the last four years deep convolutional neural networks have become a standard method for image recognition, and a variety of architectures and techniques have been developed to improve the recognition accuracy. In this work we take as a basis the VGG-16 model developed by Simonyan and Zisserman [7] that demonstrated the first and the second places in the localization and classification tasks of the ImageNet ILSVRC2014 competition.

The network receives an input RGB image of size $224 \times 224 \times 3$ that is first preprocessed by substracting, for each pixel, the mean RGB value calculated from all images in the dataset. The VGG-16 network contains 16 layers of the following types: convolutional layer (CL), pooling layer (PL), fully connected layer (FC) and rectified linear units (ReLU). Furthermore, the first two fully connected layers use dropout regularization (DL) with dropout ratio set to 50%. The convolutional layers have a kernel of size 3x3 pixels and compute a dot product between the kernel and an input layer; the rectified linear units apply an elementwise activation function which is simple $\max(0, x +$

$N(0,1)$; the pooling layers perform a downsampling operation; the fully connected layers compute probability score, and the dropout layers thin the network to reduce overfitting. Finally, the VGG-16 model has the following architecture:

$$\begin{gathered}[Input] \rightarrow \\ \{[CR] \rightarrow [CRP]\} * 2 \rightarrow \\ \{[CR] \rightarrow [CR] \rightarrow [CRP]\} * 3 \rightarrow \\ \{[FRD]\} * 2 \rightarrow [FC] \rightarrow \\ [Probability],\end{gathered}$$

where $[CR]$ denotes the $[CL3] \rightarrow [ReLU]$ layer, $[CRP]$ denotes the $[CL3] \rightarrow [ReLU] \rightarrow [PL]$ layer, and $[FRD]$ denotes the $[FC] \rightarrow [ReLU] \rightarrow [DL]$ layer respectively.

For our problem we use the weights of the original ImageNet pretrained VGG-16 model to initialize our network. We treat the model as a general-purpose feature extractor and retrain the last fully connected layer of the network.

2.1.2 Image dataset

For visual instrument recognition we employ a subset of the large hand-labeled ImageNet ILSVRC collection [8]. The collection originally bears 1000 classes and is intended for evaluation of image classification methods. The chosen synsets [1] are the following: accordion, banjo, cello, drum, flute, guitar, oboe, piano, saxophone, trombone, trumpet, violin.

There is a total of 19593 images of 12 classes, including images with a single instrument or with other objects around.

2.2 Audio-based musical instrument recognition

2.2.1 Selected approach

For audio classification we use a standard bag-of-features pipeline. As a baseline we select the approach from [9]. Following the steps in [9] we split audio files with a fixed framesize of 46 ms and hopsize of 24 ms using a Blackman-Harris windowing function, extract a big amount of spectral, cepstral and tonal descriptors (such as spectral centroid, spectral spread, spectral energy, pitch confidence, pitch salience etc.) and compute commonly used statistical measures (e.g., mean, variance and standard deviation) from both the actual and the delta values as described in the previous work [10]. We utilize the Essentia [11] library for feature extraction. Then we normalize all attributes using $L2$ normalization and perform χ^2 feature selection as preprocessing techniques. The original method proposes Support Vector Machine (SVM) algorithm for the final classification. We were also interested in an evaluation of scalable boosted decision trees (XGBoost) implemented in [12] due to its high performance as the winning solutions from Kaggle [2] and KDDCup [3] challenges.

	IRMAS	RWC	ImageNet
Classes	11	45	12
Samples p/class, median	626	43	1675
Samples p/class, std	125	64	125

Table 1: Summary of datasets used for instrument recognition evaluation.

2.2.2 Audio dataset

For evaluating musical instrument recognition in audio, we use two standard music collections, as detailed below. We also considered the University of Iowa Musical Instrument Samples (UIOWA MIS) [13], composed of 2182 samples of 20 instruments and referenced in the literature as a baseline dataset, obtaining 100% precision and recall.

IRMAS. This dataset includes musical audio excerpts from more than 2000 recordings in various styles and genres with annotations of the predominant instrument present. It was used for the evaluation in [9] and originally compiled for [10]. We use the training part of the collection that contains 6705 audio files in 16 bit stereo wav format sampled at 44.1kHz. They are excerpts of 3 seconds for 11 pitched instruments such as cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and additionally human singing voice. The median number of samples per class is 626 with standard deviation 125.

RWC Musical Instrument Sound. In our evaluation we also use the Real World Computing (RWC) Music Database: Musical Instrument Sound [14]. It contains 3544 audio excerpts labeled in 50 pitched and percussion instruments, and human voice. We take only classes that contain more than 20 objects since the original frequency distribution of the data makes difficult to perform the standard cross-validation procedure. Eventually, 45 instrument and voice classes are selected for the evaluation including piano, electric piano, glockenspiel, marimba, accordion, harmonica, classic guitar, ukulele, acoustic guitar, mandolin, electric guitar, electric bass, violin, viola, cello, contrabass, harp, timpani, trumpet, trombone, horn, soprano sax, alto sax, tenor sax, baritone sax, English horn, bassoon, clarinet, piccolo, flute, recorder, shakuhachi, shamisen, Japanese percussion, koto, concert drums, rock drums, jazz drums, percussion, soprano voice, alto voice, tenor voice, baritone voice, bass voice, R&B vocal).

Summary statistics on the datasets with the number of samples per class can be obtained in Table 1.

2.3 Multimodal fusion techniques

There are two main strategies used to integrate information from several sources into a joint multimodal system: early fusion, also known as feature level fusion, and late fusion, also known as decision level fusion. In the first case, all features from different data modalities are incorporated into a large single vector for further training. In the sec-

[1] Corresponding wnids: n02672831, n02787622, n02992211, n03110669, n03249569, n03372029, n03467517, n03838899, n03928116, n04141076, n04487394, n04536866

[2] https://www.kaggle.com/competitions

[3] http://www.kdd.org/kdd-cup

ond case, data from different sources is used for training independently and the integration is performed on the final prediction stage. Compared with early fusion, late fusion is easier to implement, has lower computational complexity and has been shown effective in practice [15]; while early fusion looks more natural from a perceptual point of view. Furthermore, early fusion requires to use a general classifier, while late fusion let us use classification methods which are more tailored to each modality.

In our case studies we follow late fusion and consider audio and visual sources independently combining them on a single frame decision level. The code of experiments, audio-based pretrained models and features are available online [4]. The finetuned VGG-16 network is available upon request.

2.4 Evaluation strategy

We first evaluate the performance of individual audio/image classifiers using standard metrics such as precision, recall and F1-score.

For the evaluation of the image-based recognition system we use stratified 5-fold cross-validation to get the average overall accuracy. Additionally, we split each train subset into the indeed training subset and the validation subset in the proportion 3:1. For each fold we select the model with the best classification accuracy on the validation subset and then evaluate on the test subset. Finally, we use a total of 11756, 2939 and 3919 images for train, validation and test sets for each fold respectively.

In order to compare the performance of the two audio-based classifiers, SVM and XGBoost, on the same dataset we follow the approach described below:

- we divide the dataset into 10 subsets for stratified 10-fold cross-validation;
- we perform multi-dimentional grid search to find the best performing combination of hyperparameters;
- once parameters are optimized, we apply the classification method and evaluate the accuracy on each subset;
- overall accuracy is averaged across all partitions; we also use these values to measure the statistical difference between classifiers;
- to compare algorithms, we use the McNemar's test as described in [16]. For each sound excerpt in the test subset, we record how it was classified by classifiers f_A and f_B and construct the following contingency table:

n_{00}– number of examples misclassied by both f_A and f_B	n_{01}– number of examples misclassied by f_A but not by f_B
n_{10}– number of examples misclassied by f_B but not by f_A	n_{11}– number of examples misclassied by neither f_A nor f_B

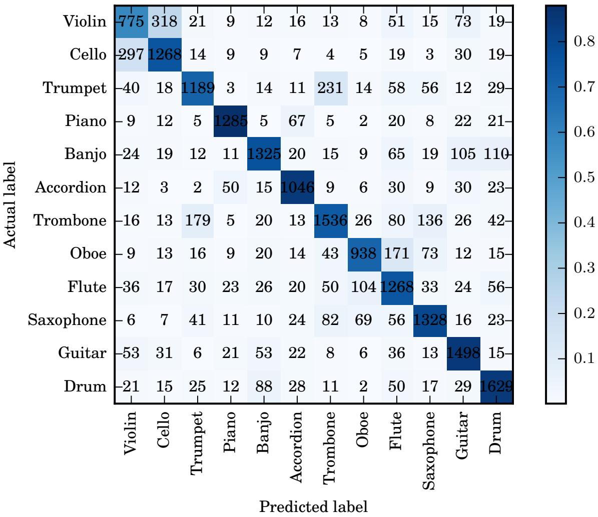


Figure 1: Confusion matrix for ImageNet musical instrument subset.

where $n = n_{00}+n_{01}+n_{10}+n_{11}$ is the total number of excerpts in the test subset.

Under the null hypothesis, the two algorithms should have the same error rate, which means that $n_{01} = n_{10}$. The McNemar's test is based on the χ^2_{MN} test statistic:

$$\chi^2_{MN} = \frac{(| n_{01} - n_{10} |) - 1^2}{n_{01} + n_{10}}$$

Next, χ^2_{MN} is compared by to the χ^2 statistics. If χ^2_{MN} exceeds $\chi^2_{1,1-\alpha}$ statistic, then we reject the null hypothesis (in our case, SVM classifier and XGBoost classifier perform equivalently on the same dataset) with $1 - \alpha$ confidence.

3. RESULTS AND DISCUSSION

3.1 Experimental results

3.1.1 Image classification

We observe in Table 2 that the overall performance is 0.77 F1 for both validation and test sets. Piano is the best classified instrument in both validation (0.88 F1) and test (0.88 F1) sets, followed by banjo, guitar, accordion and drum. Violin and flute yield the poorer performances, around 0.6 and 0.71 respectively. Figure 1 shows that the most relevant confusions correspond to instruments from the same family such as trumpet vs trombone, flute vs oboe, saxophone vs trombone or guitar vs banjo. This result is not surprising as they share similar shapes.

3.1.2 Audio classification

We observe in Table 3 that XGBoost outperforms SVM approach for IRMAS dataset, with an accuracy of 0.67 (F1). With this approach and in this dataset, Voice is the best classified instrument (0.79 F1), followed by Piano (0.75 F1), which was also the best classified instrument in the

[4] https://github.com/Veleslavia/SMC2016

Instrument	Val Prec	Val Rec	Val F1	Test Prec	Test Rec	Test F1
Violin	*0.62*	*0.59*	*0.60*	*0.60*	*0.58*	*0.59*
Cello	0.76	0.79	0.77	0.73	0.75	0.74
Trumpet	0.78	0.73	0.75	0.77	0.71	0.74
Piano	**0.88**	**0.88**	**0.88**	**0.89**	**0.88**	**0.88**
Banjo	**0.82**	0.75	0.78	**0.83**	0.76	0.80
Accordion	0.78	0.85	0.82	0.81	**0.85**	**0.83**
Trombone	0.75	0.73	0.74	0.77	0.73	0.75
Oboe	0.78	*0.67*	0.72	0.79	*0.70*	0.74
Flute	*0.67*	0.75	*0.71*	*0.67*	0.75	*0.71*
Saxophone	0.78	0.77	0.78	0.78	0.79	0.79
Guitar	0.81	**0.87**	**0.83**	0.80	**0.85**	0.82
Drum	0.81	0.84	0.82	0.81	**0.85**	**0.83**
Overall	0.77	0.77	0.77	0.77	0.77	0.77

Table 2: Validation and test performances of finetuned VGG-16 CNN method on ImageNet musical instrument subset.

image-based approach (there was no voice class in the image dataset). Violin and flute are some of the instruments with lower accuracy (0.58 F1), as it happened for the image dataset. In this approach, the saxophone has also a low accuracy, which contrasts with the image results. Figure 2 shows that the most relevant confusions correspond to instruments from the same family such as violin vs cello.

For RWC (see Table 4), XGBoost also outperforms SVM approach, with an overall accuracy of 0.83 (F1). With this approach and in this dataset, clarinet is the best classified instrument (0.79 F1). Drums is the worst classified instrument (0.58 F1). Figure 3 reveals a high confusion rate between the three classes of drums.

Although it is difficult to directly compare the results obtained from heterogeneous sources from different databases, the results are competitive with [9] and [7]. We significantly improved the classification performance with XGBoost algorithm for both audio datasets (the null hypothesis is rejected at the 0.01 significance level). We found similar confusions concerning musical instruments from the same family but there seem to be differences and similarities in the way the different instruments are distinguished through audio and visual descriptors.

3.2 Case study for combined audio and image classification

To identify musical instruments in a video we use a single-frame model from [17]. We extract image frames from video and synchronized audio excerpts of 3 seconds from corresponding audio signal. We employ the finetuned VGG-16 model to classify image frames and the IRMAS-trained XGBoost model to classify audio frames. Figure 4 illustrates an example of the results obtained for a selected set of video frames [5], [6], [7], [8].

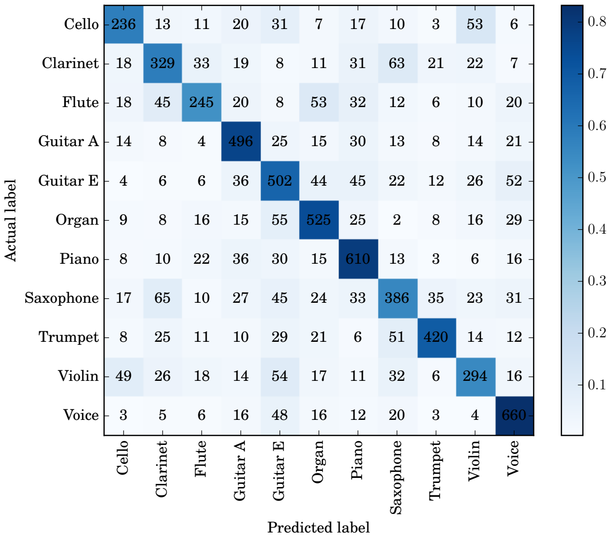


Figure 2: Confusion matrix for IRMAS dataset.

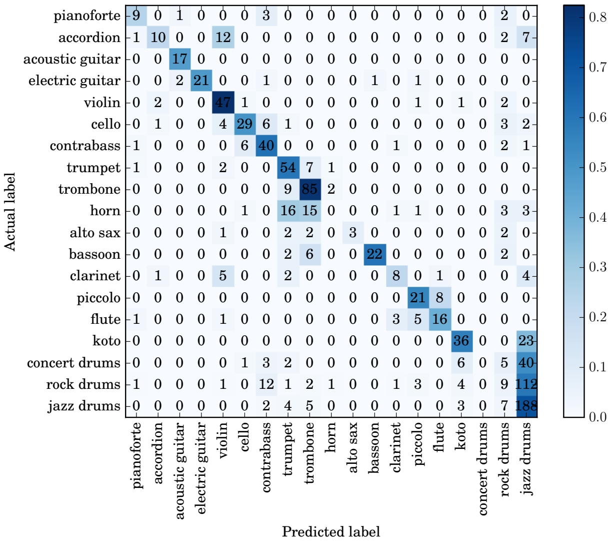


Figure 3: Confusion matrix for RWC dataset (SVM classifier).

We now provide detailed comments to each video frame. Figure 4a shows an example of the best prediction for both audio and video modalities. We observe close results in figure 4b with the high probability of the visual detection and the same confusions in the top-2 prediction as those found in the complementary confusion matrix 1. The audio has lower quality and less satisfying, although we have background voice in the first audio frame and low classification confidence in the second audio frame. We consider the next accordion example 4c as a visual-only problem since the audio classifier does not have a suitable class label. The image quality and confusions seem appropriate, and may be related to the fact that they have almost the

[5] https://youtu.be/mMl_P7zVrQw?t=23
[6] https://youtu.be/eeri7gE3ZJ0?t=17
[7] https://youtu.be/jjj0Ju3mDFk?t=31
[8] https://youtu.be/J2URcUQSpv4?t=24

Instrument	SVM Prec	SVM Rec	SVM F1	XGB Prec	XGB Rec	XGB F1
Cello	0.51	*0.22*	*0.31*	*0.61*	0.58	0.60
Clarinet	*0.43*	0.48	0.45	*0.61*	0.59	0.60
Flute	**0.77**	*0.22*	0.34	0.64	*0.52*	*0.58*
Guitar ac.	0.51	0.58	0.54	0.70	0.77	0.73
Guitar el.	0.44	0.61	0.51	0.60	0.66	0.63
Organ	0.53	0.64	0.58	0.70	0.74	0.72
Piano	*0.43*	0.70	0.53	0.72	0.79	0.75
Saxophone	0.47	0.27	0.34	0.62	0.55	*0.58*
Trumpet	0.72	0.48	0.58	**0.80**	0.69	0.74
Violin	0.57	0.42	0.48	*0.61*	0.55	*0.58*
Voice	0.63	**0.75**	**0.69**	0.76	**0.83**	**0.79**
Overall	0.54	0.52	0.50	**0.68**	**0.68**	**0.67**

Table 3: Performance of the state-of-the-art SVM method compared to the XGBoost algorithm on IRMAS dataset.

same appearance of keyboard. ImageNet confusion matrix 1 also confirms this assumption. The recognition performance on the latest example 4d seems worse than expected. Nevertheless, each frame contains the correct label in the top-2 prediction of the classifiers.

Additionally, it is worthy to mention that the pattern recognition with convolutional neural networks can be challenging even for two very similar frames as confirmed in [18].

In the presented examples, we obtained a worse generalization ability for audio than for images. It can be partially explained by the high quality of the training image dataset, while real-world audio excerpts contain a lot of background noise and low-level features have been found not to be robust even to small modifications [19].

4. CONCLUSIONS

In this article, we have studied the quality of image classification and audio classification in musical instrument recognition for several datasets. Despite the difficulties associated with direct comparison of the performance obtained from heterogeneous datasets we have shown state of the art results in both modalities. Moreover, we evaluated and compared the performance of two audio classifiers and outperformed state of the art. In addition we have demonstrated the integrated single-frame method applied for real-world video recording of a musical performance.

In future work we intend to create an annotated video dataset for musical instrument detection, investigate convolutional neural networks approach for spatio-temporal feature learning in both sound and video components and explore techniques for generating audio-visual description of performance recordings.

Acknowledgments

This research was partially supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502).

Instrument	SVM Prec	SVM Rec	SVM F1	XGB Prec	XGB Rec	XGB F1
Piano	0.60	0.25	0.35	0.94	0.92	0.93
Accordion	0.50	0.22	0.31	0.91	0.93	0.92
Guitar ac.	0.68	0.47	0.56	0.88	**0.97**	0.92
Guitar el.	**0.95**	0.49	0.65	**0.98**	0.93	0.95
Violin	0.37	**0.82**	0.51	0.83	0.93	0.88
Cello	0.49	0.51	0.50	0.90	0.93	0.91
Contrabass	0.51	0.62	0.56	0.85	0.88	0.86
Trumpet	0.35	0.60	0.44	0.85	0.81	0.83
Trombone	0.64	0.81	**0.71**	0.87	0.91	0.89
Horn	*0.00*	*0.00*	*0.00*	0.78	0.78	0.78
Alto sax	0.38	0.08	0.13	0.92	0.63	0.75
Bassoon	0.73	0.61	0.67	0.86	0.89	0.88
Clarinet	0.50	0.22	0.31	0.95	**0.97**	**0.96**
Piccolo	0.36	0.55	0.43	0.92	0.95	0.94
Flute	0.55	0.41	0.47	0.88	0.95	0.91
Koto	0.51	0.57	0.54	0.91	**0.97**	0.94
Drums c.	*0.00*	*0.00*	*0.00*	*0.56*	*0.32*	*0.40*
Drums r.	0.10	0.05	0.07	0.59	0.67	0.63
Drums j.	0.38	0.72	0.50	0.75	0.79	0.77
45 classes	0.42	0.47	0.40	**0.83**	**0.84**	**0.83**

Table 4: Performance of the state-of-the-art SVM method compared to the XGBoost algorithm on selected instruments in RWC dataset.

We thank Juanjo Bosch for assistance with IRMAS dataset and providing the code for feature extraction, and Marius Miron for useful comments about convolutional neural networks.

5. REFERENCES

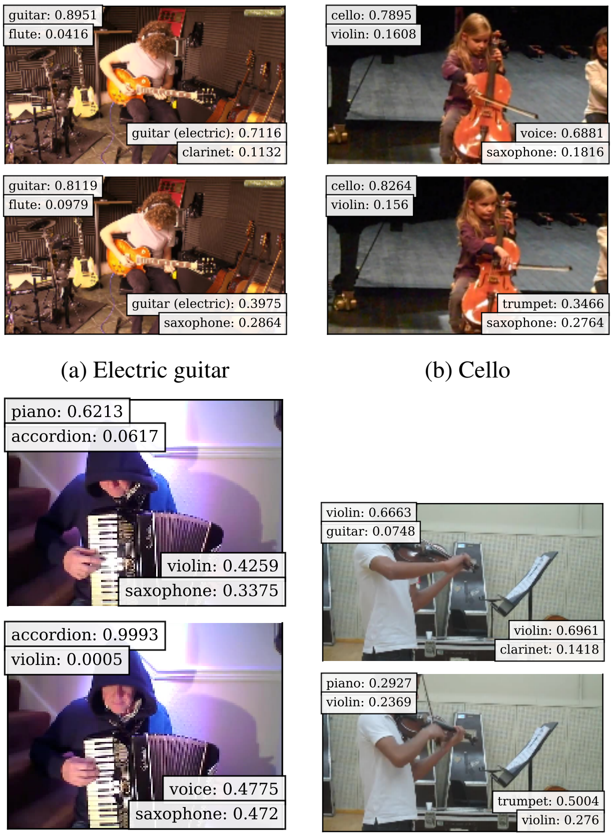
[1] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2–3, pp. 127–261, 2014.

[2] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal processing methods for music transcription*. Springer, 2006, pp. 163–200.

[3] B. Marenco, M. Fuentes, F. Lanzaro, M. Rocamora, and A. Gómez, "A multimodal approach for percussion music transcription from audio and video," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2015, pp. 92–99.

[4] C. Liem, A. Bazzica, and A. Hanjalic, "Looking beyond sound: Unsupervised analysis of musician videos," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, July 2013, pp. 1–4.

[5] A. Lim, K. Nakamura, K. Nakadai, T. Ogata, and H. G. Okuno, "Audio-visual musical instrument recognition," *73* , vol. 5, p. 9, 2011.



(a) Electric guitar

(b) Cello

(c) Accordion

(d) Violin

Figure 4: Illustrative examples of guitar, cello, accordion and violin video frames with the top-2 prediction from audio and image sources. The best two prediction of the image-based model is located in the top left corner. The best two prediction of the audio-based model is located in the bottom right corner.

[6] S. Essid and G. Richard, "Fusion of Multimodal Information in Music Content Analysis," in *Multimodal Music Processing*, ser. Dagstuhl Follow-Ups, M. Müller, M. Goto, and M. Schedl, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, pp. 37–52. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2012/3465

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[9] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 2012.

[10] F. Fuhrmann and P. Herrera, "Polyphonic instrument recognition for exploring semantic similarities in music," in *International Conference on Digital Audio Effects (DAFx)*, 2010.

[11] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra, "Essentia: An audio analysis library for music information retrieval." Citeseer.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[13] "The university of iowa musical instrument samples (uiowa mis)," theremin.music.uiowa.edu, accessed: July 19, 2016.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Music genre database and musical instrument sound database." in *ISMIR*, vol. 3, 2003, pp. 229–230.

[15] G. Ye, D. Liu, I.-H. Jhuo, S.-F. Chang *et al.*, "Robust late fusion with rank minimization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3021–3028.

[16] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classication with convolutional neural networks," in *Proceedings of International Computer Vision and Pattern Recognition (CVPR 2014)*, 2014.

[18] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 427–436.

[19] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra, "What is the Effect of Audio Quality on the Robustness of MFCCs and Chroma Features?" in *International Society for Music Information Retrieval Conference*, 2014, pp. 573–578.