Musical Instrument Recognition in User-generated Videos using a Multimodal Convolutional Neural Network Architecture

Olga Slizovskaia
Universitat Pompeu Fabra
Barcelona, Spain
olga.slizovskaia@upf.edu

Emilia Gómez
Universitat Pompeu Fabra
Barcelona, Spain
emilia.gomez@upf.edu

Gloria Haro
Universitat Pompeu Fabra
Barcelona, Spain
gloria.haro@upf.edu

ABSTRACT

This paper presents a method for recognizing musical instruments in user-generated videos. Musical instrument recognition from music signals is a well-known task in the music information retrieval (MIR) field, where current approaches rely on the analysis of the good-quality audio material. This work addresses a real-world scenario with several research challenges, i.e. the analysis of user-generated videos that are varied in terms of recording conditions and quality and may contain multiple instruments sounding simultaneously and background noise. Our approach does not only focus on the analysis of audio information, but we exploit the multimodal information embedded in the audio and visual domains. In order to do so, we develop a Convolutional Neural Network (CNN) architecture which combines learned representations from both modalities at a late fusion stage. Our approach is trained and evaluated on two large-scale video datasets: YouTube-8M and FCVID. The proposed architectures demonstrate state-of-the-art results in audio and video object recognition, provide additional robustness to missing modalities, and remains computationally cheap to train.

CCS CONCEPTS

•Information systems → Multimedia and multimodal retrieval; *Music retrieval;* **•Computing methodologies → Neural networks;** *Visual content-based indexing and retrieval;*

KEYWORDS

multimodal musical instrument classification; convolutional neural networks; multimodal video analysis; feature fusion; multimedia information retrieval

ACM Reference format:
Olga Slizovskaia, Emilia Gómez, and Gloria Haro. 2017. Musical Instrument Recognition in User-generated Videos using a Multimodal Convolutional Neural Network Architecture. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 7 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079002

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079002

1 INTRODUCTION

Humans recognize a musical instrument by combining multiple perception modalities. For example, we can distinguish a violin from a cello by its timbre, size, bow movements and relative position of the instrument with respect to the performer's body. Although the task is fairly easy for humans to perform, combining multimodal information is not trivial for machine learning algorithms.

Musical instrument recognition is a well-known problem in the music information retrieval (MIR) field. State-of-the-art methods are based on the combination of audio feature extraction (representative of the time-frequency distribution of the signal), automatic classification methods and context information on the music material under analysis. Nowadays, these algorithms provide good accuracy in recognizing musical instruments from monophonic audio recordings (i.e. single instrument playing), although the performance depends on the number of instruments and size of the audio collection used for training [17]. This performance significantly drops in polyphonic music scenarios (i.e. more than one instrument playing), where it is easier to recognize instruments if they are predominant in the audio signal [6].

Nevertheless, current approaches are based on the analysis of good-quality audio material and fail for real-world scenarios such as the one addressed here. Moreover, it's typical to find the presence of the sound of the instruments. In contrast, in this paper, our problem is to recognize the physical presence of the instruments by either sound or visual component. With this method, we hope to advance the field of indexing music videos in large-scale collections.

User-generated videos are widely found on social networks to share own musical performances, and they may contain multiple instruments, different types of noise, blur, compression artifacts, and they are varied in terms of recording conditions and quality [38]. Yet, despite these downsides, collections of user-generated videos are a rich source of knowledge. While the most important information comes from audio, visual content also plays an important role in detecting musical instruments in videos. Thus, different taxonomies for musical instruments rely on audio characteristics as well as on visual characteristics (such as a keyboard, wood, brass, bowed string). In this work, we explore the relationship between audio and visual cues and take advantage of complementary information provided by the nature of the task.

In particular, we benefit from the information embedded in the audio and visual domains by means of a Convolutional Neural Network (CNN) architecture. We train and evaluate it on two large-scale video datasets: YouTube-8M [2] and FCVID [46] which contain, respectively, more than 60000 and 5000 musical performance videos with musical instruments. The proposed architectures

demonstrate state-of-the-art results in audio and video object recognition, provide additional robustness to missing modalities, and remains computationally cheap to train. In addition, our approach meets the standards of reproducible research.

Our contributions include: (1) a novel multimodal CNN architecture for audio-visual musical instruments recognition which outperforms unimodal state-of-the-art techniques with the largest musical performance videos datasets ever used in the literature; (2) evaluation on a few recent and popular audio-only and general-purpose CNN architectures in the context of user-generated musical performance videos; (3) both FCVID and YouTube-8M datasets have been constructed for visual concept recognition; this notwithstanding, we show in a set of experiments that audio information plays a crucial role in the categorization of musical videos and can significantly improve recognition performance over visual input.

2 RELATED WORK

The increasing popularity of deep neural networks in data analysis looks like a wildfire. The ideas of convolutional and recurrent neural networks smoldered for years until they were brought to the fore by enhanced computational abilities and the availability of massive labeled data collections. Nowadays, CNNs effectively deal with a huge number of unimodal tasks and become stronger for multimodal learning rapidly [22].

Video Recognition: The breakthrough in pattern recognition on static images was largely due to its impressive feature learning ability. The computer vision community has been struggled for decades to find a way to avoid handcrafted features for solving large-scale video analysis tasks in a unique non-specific way [23, 42].

Over the last years, most of the best solutions in action recognition [7, 15, 29, 37, 42], scene recognition [2, 23, 42] and general multi-label video classification [2, 23] tasks exploit either deep neural networks on raw spatio-temporal data [15, 23, 37, 42] or combine them with motion features such as improved Dense Trajectories (including HOG, HOF and MBH) [7, 42] and Optical Flow images [29]. The most straightforward way to incorporate temporal information in video CNNs is to switch from 2D convolutions to 3D convolutions [23, 42], although it leads to difficulties in the choice of parameters such as the optimal shape for the filters, the frame-rate for analysis or the clip size, to name a few.

Several alternative methods have been recently proposed, such as two-stream CNNs [15, 37], which use single-frame architecture for spatial modeling and precomputed multi-frame optical flow images for temporal modeling, while aggregating information at the prediction stage [37] or at several layers of the network [15]. The approach in [29] examines different feature-pooling methods on CNN architectures with up to 120 frames as well as the capability of Long Short-Term Memory networks to catch temporal information. Although this approach provides good results, its computational performance is far from satisfactory. A good compromise between accuracy and speed for large-scale video classification has been proposed by several teams of researchers [2, 23]. They build systems upon frame-level spatial features and exploit average pooling and Deep Bag of Frame (DBoF) pooling for clip-level and video-level predictions.

Audio Recognition: In recent years, there has been a strong interest in deep learning at the audio signal processing community. Apart from the tremendous attention received at speech recognition, important tasks in music information retrieval (MIR) have been addressed with deep learning methods. Among them, we find approaches for musical onset detection [35], musical instrument recognition [16, 26], automatic music transcription [36], acoustic event detection [14, 33], automatic tagging [9], audio source separation [8] and various classification tasks [18, 32].

Since research in this area is still very active, there is no generic architecture working well for all the mentioned problems. Although some end-to-end methods working with raw audio have been recently proposed [3, 43], they require huge data collections and a lot of time to train. The most common approaches first transform audio data into two-dimensional image-like representations (e.g. Short-Time Fourier Transform (STFT) spectrogram [8, 14], log mel-spectrogram [9, 16, 18, 35] or Constant-Q spectrogram [26, 36]) and then train various CNN architectures. Besides, most of the architectures are either shallow, consist of only straight layer connections, or exploit squared filter shapes, which came up directly from image processing. For our case, we explore few enhancements over traditional models, such as separable convolutions [11] and partially task-specific filter shapes [32].

Multimodal Learning: Different deep learning architectures have been proposed for audio-visual speech recognition [19, 20, 30], audio-visual emotion recognition [24, 31, 45, 47], cross-modal representation learning [3] or image classification and retrieval using images and text [39, 40].

Contrariwise, the majority of MIR-related multimodal research so far relies on handcrafted audio and visual descriptors and traditional machine learning algorithms. Among them, we would like to mention multimodal approaches on detecting the playing/non-playing activity [5], automatic music transcription [27], general-purpose audio/video classification [4, 28], and artist identification problem [34].

Depending on the task and method, there are a few approaches to aggregate information from different modalities. In multimodal deep learning [30], researchers distinguish three phases: feature learning, training, and testing. For multimodal fusion, both audio and visual information is available in all phases, while for cross-modality learning or shared representation learning [3], training and testing phases exploit either audio *or* video. Even for multimodal fusion, several aggregating strategies exist, namely, (1) early fusion [19, 30, 31], where the network learns hidden representation from concatenated multimodal input; (2) middle/slow fusion [15, 23], where the network may have multiple fusion layers and optimize several learning representation simultaneously; and (3) late fusion [30, 31, 45], where the networks for all data sources are optimized separately and the learned representations are then combined to model the joint distribution of multiple modalities. Although multimodal CNNs have been proposed before for video analysis tasks [3, 19, 20, 24, 30, 31, 45, 47], to the best of our knowledge, we present the first study in the context of multimodal musical instrument recognition in video recordings.

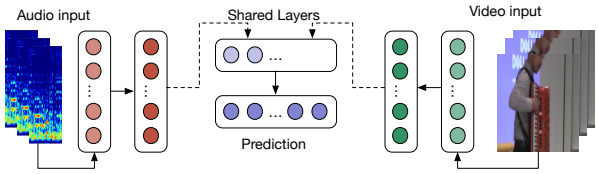

Figure 1: Schematic representation of our multimodal CNN architecture for musical instrument recognition.

3 PROPOSED METHOD

In this section, we describe our models for the task of multimodal musical instrument classification.

3.1 Visual-based recognition

Recent works [29, 44] report that spatio-temporal features can be better captured with long clips, while for short clips frame-level features have a greater impact on video object recognition performance [23]. Considering the fact that learning over long clips is a very time-consuming process, we follow the approach from [2] and extract frame-level features from videos.

For detecting instruments in static video frames, we experiment with Inception v3 architecture [41] since it's one of the most prominent and successful ones and it has been showed to provide a notable generalization ability in various tasks [18, 41]. We explore the influence of the total number of frames selected from the videos at the training phase. Moreover, we study the impact of fine-tuning the model over an independent set of images of musical instruments. The pretraining details are provided in Section 4.3.

3.2 Audio-based recognition

For audio feature representation learning, we have chosen the model from [16] (we refer to it later as *Han et al. 2016*) as a baseline. It's a classical deep CNN architecture with 8 convolutional layers stacked in a sequence, and followed by one fully connected layer. Max-pooling and dropout layers are placed after every second convolutional layer. All convolutional filters have shape of 3×3, which is similar to popular CNNs used in computer vision.

We also experiment with a modified model from [9] (we refer to it later as *Choi et al. 2016*) with a final classification softmax layer instead of gated recurrent unit layers. This architecture follows the idea of stacking convolutional layers as well, but has a larger receptive field and exploits more advanced activation function and batch normalization [21], the recent effective regularization technique.

In addition, we explore a recent Xception [11] architecture for audio-based instrument recognition. We modify the input layer so that the receptive field would be the same as in [9], and employ rectangular filters of size 48×3 at the first layer for better capturing the timbral characteristics of musical instruments. To reflect the changes of the input layer, we set the number of filters for separable convolutions equal to 768. The description of the network input is provided in Section 4.2.

3.3 Multimodal recognition

In this work, we investigate multimodal fusion strategy, so we use audio and video for both training and evaluation. Although the most direct approach for multimodal learning would be to train a model over concatenated audio-visual input (and thereby to fully integrate the modalities and learn a joint feature representation), earlier work in [30] demonstrates that there are almost no cross-modal connections in the resulting architecture. Moreover, such approach would limit us to a small number of hidden layers, which is not desirable. Thereby, following the literature [30, 45, 47], we individually train audio and video representation models and we then exploit learned features from the last layers of the networks to train and evaluate the joint model as shown in Figure 1. Since the specific parameters for the audio and visual networks change for each experiment, we comment on the architecture of the late fusion model. The input layer of the model takes a concatenated feature vector of size $(k + 1, n)$, where k is the number of video frames (plus one vector of the audio features), and n corresponds to the penultimate layer size in the audio and visual networks. The model consists of two fully-connected layers (each layer contains 1024 neurons and ReLU activation function) preceding the batch normalization, and a softmax prediction layer.

3.4 Implementation details

Our approach is implemented with Keras [10] and TensorFlow [1]. We found out that the best optimization strategy for video models consists of a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.0001, and a momentum of 0.9. We halved the learning rate every 5 epochs. We set up the batch size to 64 and an early stopping criterion to 5 epochs for our visual-based experiments. For audio architectures, we use the Adam [25] optimizer with various batch sizes and 10 epochs for an early stopping criterion. All experiments are conducted on a single NVIDIA Titan X 12GB GPU. The code, extracted features, pre-trained models, and experimental results are available online[1].

4 EXPERIMENTS AND RESULTS

4.1 Datasets

FCVID: Fudan-Columbia Video Dataset (FCVID) [46] contains videos, labels, several pre-computed descriptors and a category hierarchy. For our task, we consider a subcategory of FCVID dataset namely *Musical Performance With Instruments* containing 12 different classes including popular instruments, chamber music, rock band and orchestral performances. The subset contains 5154 videos with a total length of almost 260 hours. All videos in the dataset have been manually annotated by a team of 20 people (at least 3 annotations per video). Unluckily, we could not find any information about human performance rate and agreement rate for this dataset.

YouTube-8M: YouTube-8M Dataset [2] is a recently released large-scale video benchmark that consists of about 8 million YouTube video corresponding with 4800 visual entities. The vocabulary for the dataset has been created by humans, while the labels for individual videos have been automatically obtained. To evaluate our task we check the dataset entities and select those of them that match

[1] http://github.com/Veleslavia/ICMR2017

Property	FCVID [46]	YouTube-8M [2]
Total number of categories	12	13 (46)
Total number of videos	5,154	60,862 (235,260)
Total video duration	259.84 hr	4,152.09 hr
Mean video duration	3.03 min	4.09 min
Videos per category (mean/std)	429 / 101	4,677 / 6,445
Videos used in experiments	5,154	60,802

Table 1: Statistics of the musical instruments sub-datasets extracted from FCVID [46] and YouTube-8M [2] datasets. Numbers in parenthesis correspond to sub-dataset statistics before undersampling.

musical instruments. We gather a dataset containing 235k videos of 46 classes. At the same time, we found out that the resulting dataset contains a number of fine-grained categories representing by only a few videos while the top-3 categories form 75% of the dataset. To be able to compare results to FCVID dataset and to avoid problems related to dataset granularity and high imbalance, we reduce the number of categories to 13 and adjust the classes distribution by undersampling the top-3 classes. The final dataset then contains more than 60k videos with a total length of about 4k hours, that is the largest musical instrument recognition dataset for today. It's also worth mentioning that the original vocabulary for the dataset contains only visual entities and it has been built with an emphasis on the ease of visual object recognition. The average human performance reported in [2] is 78.8% in precision and 14.5% in recall. For our experiments we sample videos from both datasets to train, validation, and test splits with ratio 70%, 15%, and 15% respectively. The details about the datasets can be found in Table 1.

4.2 Data Preprocessing

First, we separate audio and visual data and preprocess them individually. For audio, we convert the stereo input to mono by averaging the left and right channels and downsample it. We then compute two different one-channel log-mel-spectrogram representations following the models proposed in [9, 16]. The model [16] (*Han et al. 2016*) has an input size of 128×43 (128 mel-frequency bins and 43 time frames) that corresponds to approximately 3 seconds of audio converted by Short Time Fourier Transform (STFT) with a Hann window of size 1024 samples and a hop size of 512 samples. The model [9] (*Choi et al. 2016*) has an input size of 96×1366 (96 mel-frequency bins and 1366 time frames, respectively) that corresponds to approximately 30 seconds of audio converted by using a STFT with window size of 512 samples and hop size of 256 samples. For all the experiments we select 30 seconds from each video: for model *Han et al. 2016* we select 10 segments by 3 seconds, uniformly distributed in original audio (and average predictions over 10 segments); for the model *Choi et al. 2016* we investigate two segmentation strategies: central cropping (30 seconds from the middle of the audio, *CC*) and uniformly cropped segments (10 segments by 3 seconds, *UC*). To obtain a proper input for our visual model we take frames from videos with 1 fps frame rate, then resize every frame to size 256×256×3, make a central crop with size 224×224×3 and apply random horizontal flipping. In the experiments where different numbers of frames are evaluated, we randomly select k frames for every video.

For both datasets, we only have one label per video, so we assign a video-level label to every selected frame.

Dataset	FMs	PT	Steps	Time	Hit@1	Hit@3	F1
FCVID	20	No	32K	19h	42.30	64.53	43.16
FCVID	30	No	16K	11h	65.39	81.75	67.29
FCVID	30	Yes	16K	11h	68.77	84.26	70.33
FCVID	50	No	24K	22h	67.47	83.21	69.38
FCVID	50	Yes	21K	19h	**69.39**	**84.32**	**71.23**
FCVID	100	No	43K	98h	68.56	83.97	70.42
FCVID	100	Yes	36K	84h	67.76	83.50	69.16
YT-8M	10	No	58K	82h	61.15	78.45	52.19
YT-8M	20	Yes	57K	92h	**70.07**	**84.20**	**71.09**

Table 2: Comparison of clip-level performance for visual instrument classification model trained on different number of frames (FMs) with or without pre-training (PT) on ImageNet musical instruments. All rows use the same Inception v3 architecture.

4.3 Experimental setup

Metrics: For experimental evaluation we use three standard information retrieval metrics: accuracy (Hit@1, the success rate at top-1 prediction), top-3 accuracy (Hit@3, the success rate at top-3 predictions), and F1-measure (the harmonic mean of precision and recall).

Pre-training of Inception v3 model: Since it has been proved that pre-training helps to improve generalization ability and reduce training time [13], we initialize Inception v3 model with the model weights trained from ImageNet [12] and fine-tune the model on a subset of musical instrument images as described in [38].

4.4 Results

Visual-only classification results: Table 2 provides a summary of the visual-based musical recognition experiments. We observe that using pre-training and increasing the number of frames for training from 20 to 50 provides a significant improvement to the performance of the classifier ($F1$ = 71.23, FCVID) vs the baseline method ($F1$ = 43.16, FCVID). However, further increase of the number of frames to 100 does not yields higher performance. Our experiments also demonstrate noticeable success in using a pre-trained model compared with one with random initialization. The combination of two aspects also demonstrates noticeable performance improvement on YouTube-8M dataset (from $F1$ = 61.15 to $F1$ = 70.07). At the same time, increasing the number of frames causes longer training process (from 22 to 84 hours for FCVID dataset and from 82 to 92 hours for YouTube-8M dataset), while the use of the pre-trained model decreases training time (from 22 to 19 hours for the 50-frames model on FCVID dataset).

Audio-only classification results: Results for audio-based music instrument recognition are presented in Table 3. We observe that the highest accuracy is obtained by *(Choi et al. 2016)* for

Method	#Params	Dataset	Hit@1	Hit@3	F1
Han et al. [16]	1.5M	FCVID	64.13	76.82	53.64
Choi et al. [9] + CC	2.4M	FCVID	77.73	92.05	77.18
Choi et al. [9] + UC	2.4M	FCVID	**79.81**	**96.09**	78.71
Xception [11] + UC	9.6M	FCVID	78.69	94.44	**79.35**
Han et al. [16]	1.5M	YT-8M	59.37	70.87	56.50
Choi et al. [9] + UC	2.4M	YT-8M	**83.58**	94.23	**84.26**
Xception [11] + UC	9.6M	YT-8M	83.53	**94.69**	84.16

Table 3: Clip-level performance of different audio architectures and frame selection methods trained and evaluated on FCVID (top) and YouTube-8M datasets (bottom).

Method	Dataset	Hit@1	Hit@3	F1
Xception [11] / 50 frames	FCVID	**88.28**	**97.00**	**88.27**
Choi et al. [9] / 50 frames	FCVID	86.97	96.09	87.25
Xception [11] / 20 frames	YT-8M	82.64	91.37	78.95
Choi et al. [9] / 20 frames	YT-8M	**84.01**	**93.41**	**84.69**

Table 4: Overall performance of the proposed multimodal neural network for Choi and Xception feature representations.

both datasets (79.81 for FCVID and 83.58 for YouTube-8M), although results are very close to the ones obtained using the Xception architecture (78.69 for FCVID and 83.53 for YouTube-8M). For FCVID dataset we experimented with central cropped (CC) and uniformly cropped (UC) segments for *(Choi et al. 2016)* architecture. Since we obtained that the UC segments provide additional robustness, we use them throughout all the remaining experiments.

In addition, we observe that classification results are significantly higher than the ones obtained using *(Han et al. 2016)* and that audio-based classification significantly outperforms video-based classification ($F1 = 79.35$ for audio vs $F1 = 71.23$ for video, FCVID, and $F1 = 84.26$ for audio $F1 = 71.09$ for video, YouTube-8M).

Multimodal classification results: Results for the combination of audio and video models are shown in Table 4. We observe that the highest accuracy of the audio-visual approach for FCVID is obtained using the Xception architecture ($Hit@1 = 88.28$), and the results are slightly lower for Choi ($Hit@1 = 86.97$). These results are noticeably better than the ones obtained by audio-only architectures for FCVID dataset and significantly higher than the ones obtained using a video-only architectures. For YouTube-8M dataset we observe that the classification performance of our multimodal method is 13% higher than the visual-only method. With comparison to the audio-only approach our combined method demonstrates similar results.

Confusion matrices: Figure 2 shows the confusion matrices obtained for FCVID dataset using the three proposed approaches: audio-only, video-only, and multimodal. As it can be seen the confusion is significantly reduced in the proposed multimodal approach vs the alternative methods, specially in the cases of harmonica (where the percentage of correct predictions increases from 75-76% to 99%), and violin, accordion, guitar, chamberMusic (with respectively, 11%, 10%, 10%, 10% of increase with respect to the audio alone and 10%, 14%, 13%, 21% of increase with respect to the video alone).

Figure 3 shows the confusion matrices obtained for YouTube-8M dataset. We notice several significant differences with comparison to FCVID dataset. The first one is that the classification performance varies drastically between both categories and approaches. We believe that this is related to high imbalance of the categories and substantial diversity in videos. We notice that the video sequence in our data often doesn't contain the target instrument while being annotated to the certain category.

To test this assumption we carry out a simple experiment on human recognition performance. Given a video from YouTube-8M dataset we ask non-expert humans to label it with one of the considered categories. In case of presence of multiple instruments we ask to choose the predominant one. The total amount of evaluated videos is 547, evaluated by 20 different people without specific musical training. We obtain the human performance rate for our task to be equal to 86.00 in precision, 85.00 in recall, and 85.00 in F1-measure. Those results are comparable to our multimodal results ($F1 = 85.00$ vs $F1 = 84.69$). That allows us to conclude that the task (and dataset) is difficult to solve even by humans.

The feedback from our participants also contains claims that the instruments are often not present in videos from YouTube-8M dataset. Despite it's much easier and faster to recognize the instrument by its shape, they say that if the instrument is not present on the frame it's still possible to recognize it from the audio.

5 CONCLUSION

To summarize, this paper makes several contributions. First, we introduce a multimodal method for musical instrument recognition in user-generated videos. Second, we show the case when visual object recognition can be enhanced by adding audio information. Third, we evaluate several baseline convolutional neural network architectures for audio classification. Fourth, we investigate the influence of amount of frames used for image-based object recognition in video and the influence of using a pre-trained model. We evaluate our method on a heterogeneous large-scale dataset of user-generated videos so that it can be used with different datasets and scenarios.

Our results demonstrate that both modalities are important to obtain better performance. In addition, we show that the audio-only network and our multimodal approach perform very close to the human performance rate for the musical instrument subset of automatically annotated YouTube-8M dataset. This illustrates the fact that people may not only determine the video concept based on visual cues but also on the auditory ones. Moreover, the considered audio-only models clearly outperform the video-only models and the multimodal network performs better than those based on a single modality in one of the considered datasets, illustrating the advantage of multiple modalities.

In future work, we will investigate the prospects of joint multimodal hidden representation learning, cross-correlations between

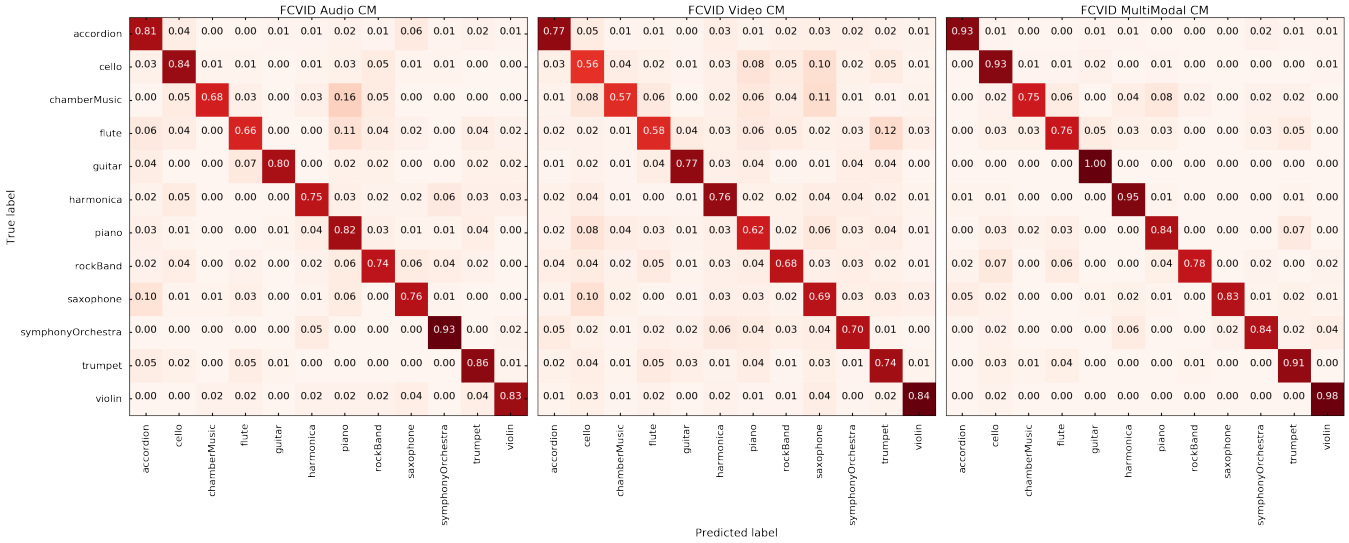

Figure 2: Comparison of confusion matrices for FCVID dataset. From left to right: audio-only recognition, video-only-recognition, multimodal recognition.

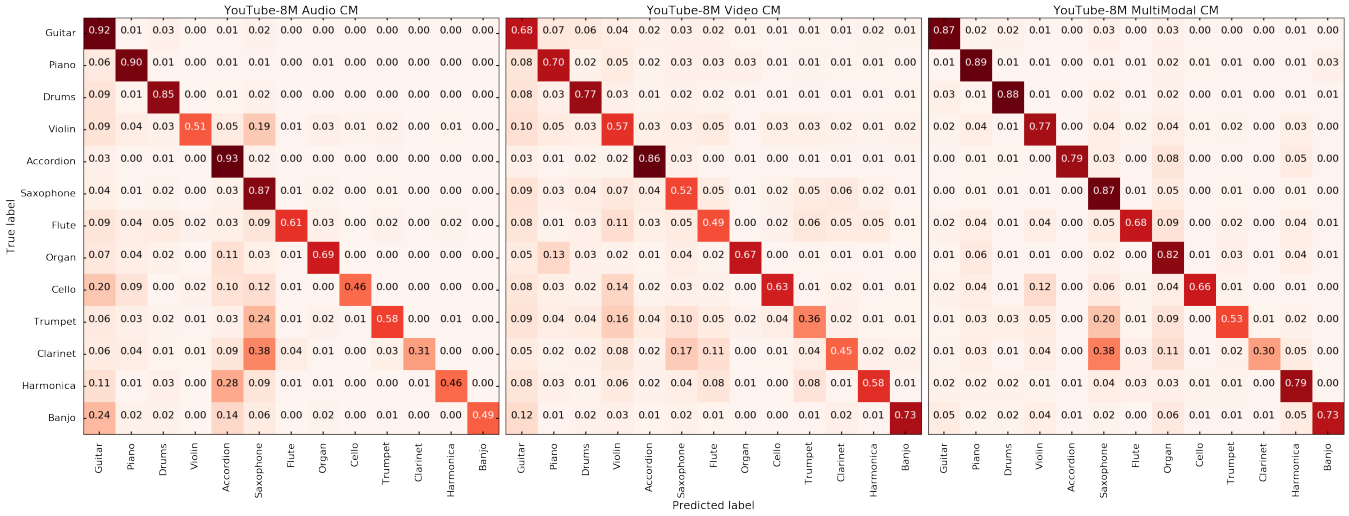

Figure 3: Comparison of confusion matrices for YouTube-8M dataset. From left to right: audio-only recognition, video-only-recognition, multimodal recognition.

audio and visual modalities, and the impact of dynamic video information.

ACKNOWLEDGMENTS

This work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), the CASAS Spanish research project (TIN2015-70816-R), and project TIN2015-70410-C2-1-R (MINECO/FEDER, UE). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *CoRR* abs/1609.08675 (2016). http://arxiv.org/abs/1609.08675

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. 892–900.

[4] Ana M. Barbancho, Lorenzo J. Tardón, Javier López-Carrasco, Jana Eggink, and Isabel Barbancho. 2015. Automatic Classification of Personal Video Recordings Based on Audiovisual Features. *Know.-Based Syst.* 89, C (Nov. 2015), 218–227. DOI : http://dx.doi.org/10.1016/j.knosys.2015.07.005

[5] Alessio Bazzica, Cynthia CS Liem, and Alan Hanjalic. 2016. On detecting the playing/non-playing activity of musicians in symphonic music videos. *Computer Vision and Image Understanding* 144 (2016), 188–204.

[6] J. Bosch, J. Janer, Ferdinand Fuhrmann, and Perfecto Herrera. 2012. 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)*. Porto, Portugal, 559–564.

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[8] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. 2017. Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. In *13th International Conference on Latent Variable Analysis and Signal Separation (LVA ICA2017)*.

[9] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic Tagging Using Deep Convolutional Neural Networks. In *International Society of Music Information Retrieval Conference, New York, USA*. ISMIR.

[10] François Chollet. 2015. Keras. https://github.com/fchollet/keras. (2015).

[11] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357* (2016).

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[13] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why Does Unsupervised Pre-training Help Deep Learning? *J. Mach. Learn. Res.* 11 (March 2010), 625–660.

[14] Miquel Espi, Masakiyo Fujimoto, Keisuke Kinoshita, and Tomohiro Nakatani. 2015. Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing* 1 (2015), 26.

[15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

[16] Yoonchang Han, Jaehun Kim, and Kyogu Lee. 2016. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *arXiv preprint arXiv:1605.09507* (2016).

[17] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov. 2003. Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research* 32, 1 (2003), 3–21. DOI: http://dx.doi.org/10.1076/jnmr.32.1.3.16798

[18] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2016. CNN Architectures for Large-Scale Audio Classification. In *arXiv*. https://arxiv.org/abs/1609.09430

[19] Di Hu, Xuelong Li, and others. 2016. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3574–3582.

[20] Jing Huang and Brian Kingsbury. 2013. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7596–7599.

[21] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 448–456.

[22] M. I. Jordan and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. DOI: http://dx.doi.org/10.1126/science.aaa8415

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[24] Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3687–3691.

[25] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Vincent Lostanlen and Carmine-Emanuele Cella. 2016. Deep convolutional networks on the pitch spiral for musical instrument recognition. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*. New York City, NY, USA.

[27] Bernardo Marenco, Magdalena Fuentes, Florencia Lanzaro, Martín Rocamora, and Alvaro Gómez. 2015. *A Multimodal Approach for Percussion Music Transcription from Audio and Video*. Springer International Publishing, Cham, 92–99. DOI: http://dx.doi.org/10.1007/978-3-319-25751-8_12

[28] L. Nanni, Y.M.G. Costa, D.R. Lucio, C.N. Silla Jr., and S. Brahnam. 2017. Combining visual and acoustic features for audio classification tasks. *Pattern Recognition Letters* 88 (2017), 49 – 56. DOI: http://dx.doi.org/10.1016/j.patrec.2017.01.013

[29] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4694–4702. DOI: http://dx.doi.org/10.1109/CVPR.2015.7299101

[30] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[31] Lei Pang and Chong-Wah Ngo. 2015. Mutlimodal Learning with Deep Boltzmann Machine for Emotion Prediction in User Generated Videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR '15)*. ACM, New York, NY, USA, 619–622. DOI: http://dx.doi.org/10.1145/2671188.2749400

[32] Jordi Pons and Xavier Serra. 2017. Designing Efficient Architectures for Modeling Temporal Features with Convolutional Neural Networks. In *42th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*. IEEE, New Orleans, USA.

[33] Justin Salamon and Juan Pablo Bello. 2016. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *CoRR* abs/1608.04363 (2016). http://arxiv.org/abs/1608.04363

[34] Alexander Schindler and Andreas Rauber. 2015. *An Audio-Visual Approach to Music Genre Classification through Affective Color Features*. Springer International Publishing, Cham, 61–67. DOI: http://dx.doi.org/10.1007/978-3-319-16354-3_8

[35] Jan Schluter and Sebastian Bock. 2014. Improved musical onset detection with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 6979–6983.

[36] S. Sigtia, E. Benetos, and S. Dixon. 2016. An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 5 (May 2016), 927–939. DOI: http://dx.doi.org/10.1109/TASLP.2016.2533858

[37] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568–576.

[38] Olga Slizovskaia, Emilia Gómez, and Gloria Haro. 2016. Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies. In *13th Sound and Music Computing Conference (SMC 2016)*. Hamburg, Germany.

[39] Kihyuk Sohn, Wenling Shang, and Honglak Lee. 2014. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*. 2141–2149.

[40] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 4489–4497. DOI: http://dx.doi.org/10.1109/ICCV.2015.510

[43] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR* abs/1609.03499 (2016). http://arxiv.org/abs/1609.03499

[44] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2016. Long-term Temporal Convolutions for Action Recognition. *arXiv preprint arXiv:1604.04494* (2016).

[45] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. 2016. Video Emotion Recognition with Transferred Deep Feature Encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. ACM, New York, NY, USA, 15–22.

[46] Jun Wang Xiangyang Xue Shih-Fu Chang Yu-Gang Jiang, Zuxuan Wu. 2015. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *arXiv preprint arXiv:1502.07209* (2015).

[47] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2016. Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 281–284.