

# A Bimodal Learning Approach to Assist Multi-sensory Effects Synchronization

Raphael Abreu  
CEFET/RJ

Rio de Janeiro, Brazil  
raphael.abreu@eic.cefet-rj.br

Joel dos Santos  
CEFET/RJ

Rio de Janeiro, Brazil  
jsantos@eic.cefet-rj.br

Eduardo Bezerra  
CEFET/RJ

Rio de Janeiro, Brazil  
ebezerra@cefet-rj.br

**Abstract**—In mulsemmedia applications, traditional media content (text, image, audio, video, etc.) can be related to media objects that target other human senses (e.g., smell, haptics, taste). Such applications aim at bridging the virtual and real worlds through sensors and actuators. Actuators are responsible for the execution of sensory effects (e.g., wind, heat, light), which produce sensory stimulations on the users. In these applications sensory stimulation must happen in a timely manner regarding the other traditional media content being presented. For example, at the moment in which an explosion is presented in the audiovisual content, it may be adequate to activate actuators that produce heat and light. It is common to use some declarative multimedia authoring language to relate the timestamp in which each media object is to be presented to the execution of some sensory effect. One problem in this setting is that the synchronization of media objects and sensory effects is done manually by the author(s) of the application, a process which is time-consuming and error prone. In this paper, we present a bimodal neural network architecture to assist the synchronization task in mulsemmedia applications. Our approach is based on the idea that audio and video signals can be used simultaneously to identify the timestamps in which some sensory effect should be executed. Our learning architecture combines audio and video signals for the prediction of scene components. For evaluation purposes, we construct a dataset based on Google’s AudioSet. We provide experiments to validate our bimodal architecture. Our results show that the bimodal approach produces better results when compared to several variants of unimodal architectures.

## I. INTRODUCTION

Multimedia applications involve the presentation of different audiovisual objects organized in time and space [1], [2]. Given the nature of the content they present, the majority of current multimedia applications stimulate only two human senses: sight and hearing. As discussed in [3], aiming at increasing the user quality of experience (QoE) and immersion with multimedia applications, the literature present works [4]–[6] that propose the use of other sensory effects in multimedia applications in order to provide users with new sensations during a multimedia presentation. In [3], the term mulsemmedia (*MULTiple Sensorial MEDIA*) is put forward to denote multimedia applications in which traditional media content (text, image, audio, video, etc.) can be related to media objects that target other human senses (e.g., smell, haptics, etc.).

To clarify, let us describe a simple yet powerful example<sup>1</sup>.

Consider a non-linear show, i.e., a show whose narrative line is not known *a priori* and is constructed based on user interaction. In this show the user actively participates in the construction of the narrative line by choosing the next sight in a city tour from a list of available options provided by the application. At each sight, a video and complementary information about it are presented to the user. At the beginning of the show, the user may choose whether he/she wants to interact with the application. If not, a default tour is presented. Let us now consider an evolution of such application which includes sensory effects. In this new application, several sensory effects are presented along the narrative line in a synchronized way, with the purpose of increasing the immersion of the user in the audiovisual content. In this scenario, if the user chooses to visit a particular sight, e.g., the beach at Rio de Janeiro, the sensory effects to be presented by the application would mimic the sight’s environmental conditions (e.g., hot wind blowing, smell of the sea breeze, etc).

Synchronization plays an important role for multimedia applications. Multimedia authoring languages are domain specific languages that provide constructions for defining how media objects shall be presented during the execution of a multimedia application, i.e., their temporal synchronization. Moreover, those languages also manage user interaction as a special case of temporal synchronization. Examples of such languages are SMIL (*Synchronized Multimedia Integration Language*) [7] and NCL (*Nested Context Language*) [8].

When considering the use of sensory effects in multimedia applications, the usual approach is to use audiovisual media objects as the base for synchronization, such that the timestamps in which sensory effects are to be executed are defined in relation to certain media objects. For example, light and heat effects may be presented when an explosion occurs in the main video. One should notice, however, that although multimedia languages ease the specification of the synchronization aspect in an application, the task of defining all the moments in which a given sensory effect shall be executed is still carried manually by the author. In general, authors relate sensory effects to the content being presented by the application.

In [9], we define a *scene component* as, a given element (rock, tree, dog, person, etc.) or concept (happy, crowded, dark, etc.) that appears in the content of a media object. In the application example we presented earlier, scene components

<sup>1</sup>The application we describe here is inspired by the Day’s Route application available at <http://club.ncl.org.br/node/69>

may refer to the sun, the beach, trees, flowers, and other elements that may appear in the Rio de Janeiro's sights presented in the touristic program.

In previous work [9], [10], we tackle the problem of automatically recognizing scene components in audiovisual objects, in order to assist the realization of the synchronization task in mulsemmedia applications. We proposed an architecture capable of identifying the presence of scene components in video and audio objects and defining, in a semi-supervised manner, the synchronization among sensory effects and an application main video and/or audio.

In this paper, we focus on the recognition of scene components specifically related to sensory effects. Examples of such kind of scene components are wind, explosion, rain, lightning, etc. Some of these components may be predominantly found either in the video (e.g., lightning) or in the audio (e.g., wind). Other components (e.g., explosion) can be found in both modalities of the audiovisual object. In particular, we propose the use of a bimodal neural network architecture for increasing the accuracy of recognition of scene components specifically for the task of synchronizing sensory effects in mulsemmedia applications. Our premise is that trying to combine both modalities in the identification can produce a better accuracy when compared to the separate identification. We perform computational experiments with different network architectures, both unimodal (either audio or video) and bimodal (audio and video). Through experimental verification, we show that the proposed bimodal fusion approach is superior in accuracy when compared with any single modal recognition system.

The contributions of this paper are twofold: (i) we propose a bimodal learning architecture for predicting scene components in audiovisual content, and (ii) we present the first version of a dataset tailored for the prediction of scene components related to sensory effects.

The remaining of this paper is structured as follows. Section II presents related work regarding the use of bimodal neural networks to combine audio and video information in several application domains. Section III provides a brief description of AudioSet, the collection of video clips we use on our learning architecture. This Section also describes how we built our own training set based on AudioSet. Section IV discusses the network architectures used in this work and how both modalities can be used to improve accuracy. Section V presents experimental results along with a corresponding analysis. Section VI concludes and discusses future work.

## II. RELATED WORK

Over the last few years, several neural network architectures have been proposed as a fusion model for audio and visual data, in a plethora of application domains, such as speech recognition [11]–[13], speaker recognition [14]–[16], video classification [17], emotion recognition [18], natural sound recognition [19], [20], and person recognition [21].

One of the first attempts of using neural networks to explore the correlation between audio and visual information

was made by Ngiam et al [11]. They use an extension of Restricted Boltzmann Machines with sparsity to learn better single modality representations given unlabeled data from multiple modalities. They apply their model to improve the accuracy of speech recognition by using not only the audio signal, but also the image frames of lips of the speaker. The authors show that better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are used at feature learning time. Also aiming at leveraging images of people speaking, in [12] the authors use 3D convolutional neural networks for audio-visual matching in which a bridge between spatio-temporal features is established to build a common feature space between audio-visual modalities. Another architecture to learn a model for audio visual speech recognition is proposed in [13]. This model comprises a convolutional neural network (CNN) followed by a Long Short-Term Memory (LSTM) neural network to handle visual modality, another LSTM RNN to handle audio modality, and a multimodal layer to fuse the outputs of both modalities.

In [14], an audio-visual speaker recognition method is presented. The method works by fusing face and audio information via multi-modal correlated neural networks. The facial features are learned by convolutional neural networks. In [15], the authors use a CNN as a face feature extractor from face imagery data, which are latter stacked with mel frequency cepstrum coefficients. In [16], the authors present a model to learn bimodally informative structures from audio-visual signals. They approach the problem of multimodal data processing by representing each signal as a sparse sum of audio-visual kernels. The authors also propose an unsupervised learning algorithm to form dictionaries of bimodal kernels from audio-visual material.

In [17] a multilayer and multimodal fusion framework of deep neural networks for video classification is proposed. The authors use four modalities to extract complementary information across multiple temporal scales. For each single modality, discriminative representations are computed for convolutional and fully connected layers. For the fusion of multiple layers and modalities, they propose an adaptive boosting model to learn the optimal combination of them. In contrast to our current work, the authors do not use audio modality.

In [18], the authors propose a multimodal Deep Convolution Neural Network (DCNN) to combine audio and visual cues for emotion recognition. They first fine-tune two DCNN models to perform audio and visual emotion recognition tasks respectively on the corresponding labeled speech and face data. Then, the outputs of these two DCNNs are integrated in a fully-connected neural network, which is trained to obtain a joint audio-visual feature representation for emotion recognition.

In [19], SoundNet is presented, a deep convolutional architecture for natural sound recognition. SoundNet learns audio representations directly on raw audio signals. The audio recognition model is trained by transferring knowledge from pre-trained visual representations and large amounts of unlabelled video. The authors achieve state-of-the-art accuracy on three standard acoustic scene classification datasets. Another ap-

proach to recognize natural (environmental) sound is presented in [20], in which the author’s convolutional neural networks are designed specifically for object recognition in images, and can be successfully trained to classify spectral images of environmental sounds. Similar to our current work, both [19] and [20] try to classify non-human sounds. However, they use only the sound modality, not exploiting information from other modalities to help identifying the sound. Moreover, they do not aim at assisting synchronization task in mulsemmedia applications.

In [21], the authors propose an audio-visual bimodal person recognition system. This system uses CNNs as a primary model architecture. First, two separate Deep CNN models are trained with the help of audio and facial features, respectively. The outputs of these CNN models are then combined/fused to predict the identity of the subject (person).

### III. DATASET CONSTRUCTION

AudioSet [22] is an extensive collection of 10-second segments (clips) of sound that belongs to YouTube videos. Each segment comes annotated with audio events found in it. This dataset contains 632 audio event classes and over 2 million sound clips. The dataset is divided in three disjoint sets: a balanced training set, an evaluation set and a unbalanced training set. The first two sets are balanced so that every label is associated to roughly 60 examples. The unbalanced training set contains the remainder examples in the collection. AudioSet does not actually come with the video contents. Instead, each entry in this catalog makes a reference to the corresponding video ID, whose content can in turn be accessed from YouTube.

Although AudioSet does not provide video features, it provides a CSV file that ties the YouTube video ID to the audio events found in the segment. For example, the entry `7ZDX0YRZHVK,20.000,30.000,“/M/02_41,/M/0838F”` in that CSV file means the YouTube video whose ID is `7ZDX0YRZHVK`, for the 10 seconds timeframe from 20s to 30s, presents sounds of “Fire” (`/M/02_41`) and “Water” (`/M/0838F`). With this information, we could select the related videos and download them (from YouTube), for every label in our chosen subset of AudioSet. In the remaining of this section, we describe how we constructed our training and validation datasets taking AudioSet as starting point.

The training dataset we constructed for use in our experiments (see Section V) is a subset of the clips referenced by AudioSet. In particular, we used the unbalanced training set already provided by AudioSet for building our own training set. This subset was built by applying a two-step selection procedure, as described in the following paragraphs.

In the first step, we selected the examples associated to a small subset of the 632 audio event classes contained in AudioSet. Since in this work we are interested in sound events that can be associated to sensory effects, we selected only examples coming from the following 7 labels for event classes from AudioSet: *Wind*, *Thunder*, *Rain*, *Ocean*, *Fire*, *Explosion* and *Gunshot, gunfire*. The overall distribution of these 7 labels

in the original dataset is shown in Fig. 1. Note that the distribution is not uniform as this dataset is unbalanced.

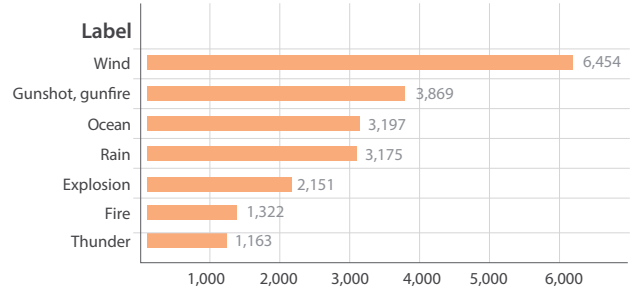


Fig. 1. Distribution of labels in the unbalanced training dataset of AudioSet for the chosen labels (*Wind*, *Thunder*, *Rain*, *Ocean*, *Fire*, *Explosion* and *Gunshot, gunfire*).

In the second step, we took as input the examples resulting from the first step. In order to alleviate the existing imbalance in the AudioSet subset associated to the above selected labels, we fixed an upper limit to download at most 2000 examples for each of the 7 labels that we used. However, due to label co-occurrence in the data, the actual number of downloaded examples for some labels (namely, *Rain* and *Wind*) ended up to be higher than the upper limit we established. In total, we downloaded 11,518 distinct segments. A total of 568 videos failed to download (which accounts for a degradation of  $\approx 5\%$ ). The cause for this failure may be that the videos were not accessible from our location, or that these videos have been removed. Thus the resulting size of our training set is 10,950 distinct segments with a total of 12,829 labels. Fig. 2 shows the resulting distribution of retrieved segments for our training dataset.

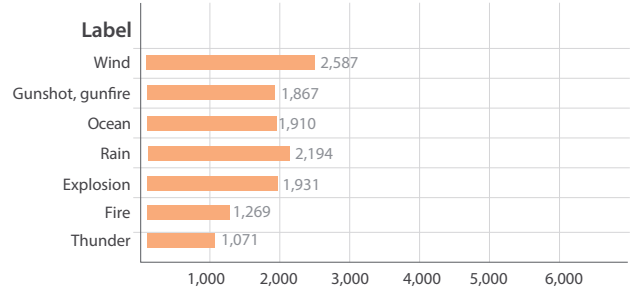


Fig. 2. Distribution of segments in our training dataset, for the chosen labels: *Wind* (20.17%), *Thunder* (8.35%), *Rain* (17.1%), *Ocean* (14.89%), *Fire* (9.89%), *Explosion* (15.05%) and *Gunshot, gunfire* (14.55%).

In order to build our validation dataset, we extracted segments from all the videos referenced in the evaluation set of AudioSet. For this validation dataset, we applied only the first step of the procedure we used to build the training set. The second step was not necessary given that no label had more than 2000 segments. In total 582 segments were selected to download. However due to unavailability to download some segments, the total set became 532 segments with 657 total labels. The resulting evaluation set provides at least 53 examples for each label. The least represented label is *Thunder*

with 53 examples while the most represented is *Wind* with 173 examples. Fig. 3 shows the resulting distribution of retrieved segments for our validation dataset.

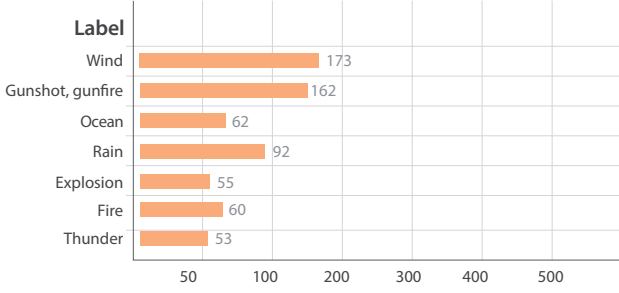


Fig. 3. Distribution of segments in our validation dataset, for the chosen labels: *Wind* (26.33%), *Thunder* (8.07%), *Rain* (14.0%), *Ocean* (9.44%), *Fire* (9.13%), *Explosion* (8.37%) and *Gunshot, gunfire* (24.66%).

#### IV. LEARNING ARCHITECTURE

In this section, we describe our proposed neural network architecture to learn a model for the task of predicting scene components within an audiovisual content. Fig. 4 presents an overview of our learning architecture.

As input to our learning architecture, we consider a collection of clips, each one labelled with at least one event class (see Section III). Besides, we assume that the audio and visual input signals for each clip used to learn the model are contiguous audio (spectrogram) and video frames, respectively. Each training example is labelled with one or more labels of our select subset of AudioSet.

Our overall bimodal learning architecture comprises three modules (dashed rectangles in Fig. 4). Two of them are responsible for the bimodal feature extraction, that is, for extracting features from audio and visual signals contained in each training example (clip). The two extracted feature vectors are concatenated and the resulting vector is used as input to the fusion module. We describe the audio and video feature extraction modules in Section IV-A and Section IV-B, respectively. In Section IV-C, we detail the fusion module. Note that, in all three modules, all hidden layers are equipped with ReLU non-linearities [23].

##### A. Audio Module

The architecture used in this network module is somewhat similar to the one employed in [24]. The primary architectural difference from the aforementioned implementation is that we used the Mel-spectrograms corresponding to an 1-channel input in our network, while the author of [24] added their deltas, forming a 2-channel input.

The first convolutional layer ( $Conv_A^1$ ) consists of 80 filters with  $57 \times 6 \times 1$  receptive fields. Then a max-pooling operation is applied with a pooling shape of  $4 \times 3$  and stride of  $1 \times 3$ . In order to avoid overfitting, we apply 50% dropout rate [25] after this layer. The second convolutional layer ( $Conv_A^2$ ) consists of 80 filters with  $1 \times 3$  receptive fields. Another max-pooling is applied with a pooling shape of  $1 \times 3$  and stride of  $1 \times 3$ .

Batch Normalization [26] was applied after each convolutional layer.

The activation volume resulting from  $Conv_A^2$  is flattened to a  $3600 \times 1$  vector and further fed into a fully connected stage with 2 layers ( $FC_A^1$  and  $FC_A^2$ ), the first with 5000 units, and the second with 1000 units, to standardize the feature-length dimensionality and feed to the fusion network. A 50% dropout rate is also applied after the first fully connected layer ( $FC_A^1$ ).

##### B. Video Module

This module extracts features from a sequence of frames by first applying three convolutional stages to it (see parts labelled  $Conv_V^i$ ,  $1 \leq i \leq 6$ , in Fig. 4). Each stage is composed of two identical convolutional layers followed by a batch normalization and max-pooling operation with shape of  $2 \times 2$  and stride of  $2 \times 2$ . All the convolutional filters have  $3 \times 3$  receptive fields and pad the output with zeros to keep the same input size in all of the stages. The amount of filters in the first, second and third convolutional stages are 32, 64 and 128, respectively. A 50% dropout rate is applied after the first convolutional stage.

The goal of this module is to learn a representation for video clips that will enhance the prediction based on audio. For this matter, every convolutional layer is applied in the time dimension as well (i.e., to every frame in the sequence) in order to learn spatio-temporal features.

The convolutional layer activations are flattened to a  $512 \times 1$  shape. This process happens for every frame of the input sequence. The resulting shape is joined with the subsequent frames. The resulting shape is further fed to a Long-Short Term Memory (LSTM) layer [27] with 256 cells to learn the long-term temporal structure of frames. Lastly, this is followed by a fully connected layer ( $FC_V$ ) with 1000 units. A 50% dropout rate is applied after the first convolutional stage and after the LSTM layer ( $LSTM_V$ ).

##### C. Fusion Module

The fusion network is modelled to do the prediction task. Thus we concatenate the two vectors resulting from the Audio and Video modules. This concatenation leaves us with a 2000-dimensional feature vector, which is used as input to the fusion module. The fusion module is a fully connected feed-forward network, which does the prediction. It comprises 2 hidden layers, each of them with 500 units. A 50% dropout rate is applied after each hidden layer.

It is possible that more than one label are associated to a single training example in our selected subset of AudioSet (see Section III), which results in a classification setting that is both multi-class and multi-label. Hence, the output layer of the fusion module comprises 7 units, one for each label in our dataset. We map the predictions to the labels using a sigmoid output layer. The reason for this choice is that, for our multi-label classification problem, it does not make sense to use softmax in the output layer, since we wanted to produce an independent output for each label. For this reason, we decided to use sigmoid non-linearity in the output layer,

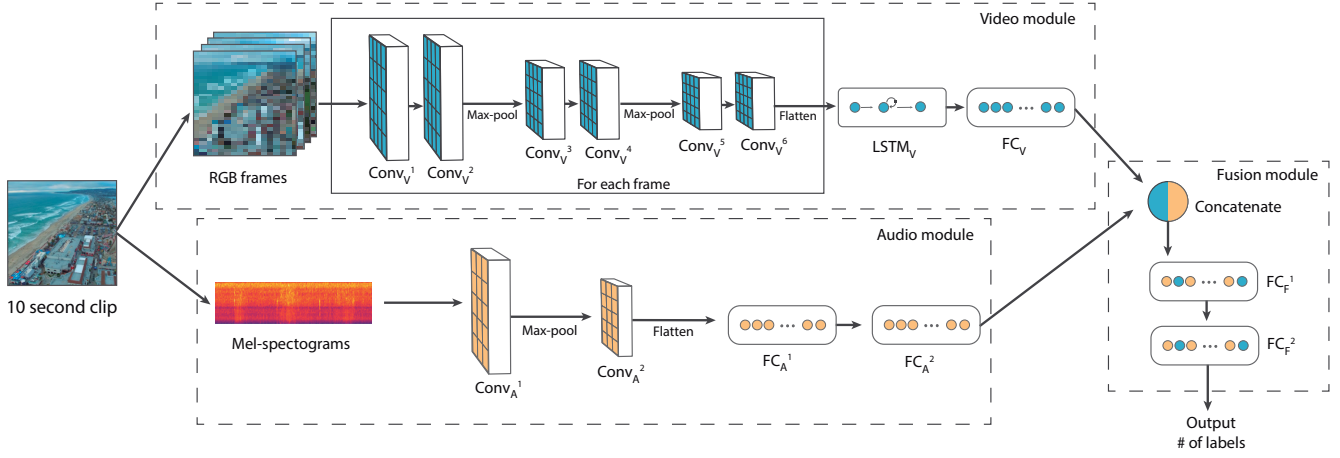


Fig. 4. An illustration of our learning architecture, which comprises two components: the bimodal feature extraction net and the fusion net. The bimodal net extracts features from audio and visual signals contained in each training example (clip). The computed feature vectors are concatenated and the resulting vector used as input to a 2-layer fully connected feed-forward network, which does the prediction.

since for this function the predicted outputs for each label is in the continuous range  $[0, 1]$ , where a value near one means the presence of the corresponding event label, and a value near zero means its absence.

We employ a multi-label training loss in the fusion module. Precisely, we use the binary cross entropy loss function, which is adequate for such classification setting.

## V. EXPERIMENTS AND RESULTS

This Section presents the experiments we performed in order to validate our learning architecture.<sup>2</sup> Section V-A describes the data preprocessing activities. Section V-B presents the metrics we used to evaluate our models. Finally, Section V-C describes the experiments we conduct to validate our learning architecture, along with a corresponding analysis of results.

### A. Data preparation

In Section III, we describe the procedure we followed to build our training and validation datasets. The preparation of segments in these datasets was performed as follows:

- **Video:** In AudioSet, every segment has a fixed sized of 10 seconds. To reduce the problem space, we subsampled every video-clip down to 32 frames. We also downsized each frame to  $32 \times 32$  pixels and kept RGB values, leaving us with  $32 \times 32 \times 32 \times 3$  tensor for the video signal of each training example. After that, we normalized the value of each pixel to keep it within the range  $[0, 1]$ .
- **Audio:** The task of learning temporal information from sound events can be accomplished by feeding the raw signals directly into a neural network, as shown in [28]. But as have been verified in [24], some raw audio data can be transformed into lesser complex acoustic

characteristics that represent the sound, such as Log-scaled mel-spectrograms. Thus we employ the same technique presented by that author to transform the audio signal of each training example. Firstly, all audio clips were resampled at 22050 Hz and min-max normalized between  $-1$  and  $1$ . Log-scaled Mel spectrograms were extracted from all recordings with window size of 1024, hop length of 512 and 60 mel-bands. Different from the implementation in [24], we did not split the spectrograms into frames.

The reason for this design decision is to maintain the same temporal scale for the both inputs in our bimodal architecture. This decision enables the network to learn on the whole 10s clips along with video representation.

### B. Evaluation Metrics

To measure accuracy of the trained models, we utilized several evaluation metrics: Micro-averaged F1-score, Exact Match Ratio, and Hamming Loss. Bellow, we briefly describe each one of them in this section. For the equations presented in this section, consider that  $|L|$  is the total number of labels, and  $|D|$  is the number of examples in the validation set.

As the datasets utilized in this work are unbalanced we need performance metrics that are independent of the class distribution. Thus we opted to estimate the Micro-averaged F1-score [29] (Micro-F1) which can be seen as the weighted average of F1 scores over all the labels. Following [30], the Micro-F1 can be defined as presented in Eq. 1. In this equation,  $x_i$  is the predicted value for a given example, and  $y_i$  is the corresponding ground truth.

$$\text{Micro-F1} = \frac{2 \sum_{l=1}^{|L|} \sum_{i=1}^{|D|} x_i^l y_i^l}{\sum_{l=1}^{|L|} \sum_{i=1}^{|D|} x_i^l + \sum_{l=1}^{|L|} \sum_{i=1}^{|D|} y_i^l} \quad (1)$$

Exact Match Ratio (MR) [31] considers one instance as correct if and only if all associated labels are correctly predicted.

<sup>2</sup>Source code for assembling the dataset and for training the networks can be downloaded from [https://github.com/MLRG-CEFET-RJ/bimodal\\_audioset](https://github.com/MLRG-CEFET-RJ/bimodal_audioset)



Exact Match Ratio formula is given by Eq. 2, where  $|D|$  is defined as above,  $y_i$  is the ground truth for the  $i$ -th example,  $x_i$  is the prediction, and  $I$  is the indicator function (equals 1 if the statement  $x_i = y_i$  is true, and equals 0 otherwise).

$$\text{MR}(x_i, y_i) = \frac{1}{|D|} \sum_{i=1}^{|D|} I[x_i = y_i] \quad (2)$$

A disadvantage of MR criterion is that it does not take partial matches into account. In order to account for partially correctness we also employ Hamming Loss, that represents how many times on average, a label is incorrectly predicted [32]. The Hamming Loss formula is presented in Eq. 3, where  $|D|$  and  $|L|$  are defined as above,  $x_i$  is the predicted value for a given example, and  $y_i$  is the corresponding ground truth. The exclusive disjunction (xor operation) is used to compute the symmetric difference of the values for each label. It returns 0 if the label is equal in both  $x_i$  and  $y_i$  and 1 otherwise.

$$\text{HammingLoss}(x_i, y_i) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{\text{xor}(x_i, y_i)}{|L|} \quad (3)$$

### C. Results and Analysis

Network optimization hyper-parameters for all modalities were similar to the ones used in [24]. We optimized our models using Stochastic Gradient Descent with Nesterov momentum of 0.9 and learning rate of 0.01. Both Video Module and Audio Module were trained for 40 epochs with mini-batch implementation, with batch size of 32. We implemented our learning architecture using Keras [33] with TensorFlow [34] as backend.

In order to validate our proposed learning architecture (see Section IV), we first investigate three different experimental settings: we train models using audio signal only, video signal only, and both signals (i.e., bimodal).

The first and second experimental settings correspond to training the Audio and Video modules, respectively, as separated networks. Results for these experimental settings are presented in the first three lines of Table I. We call the network trained only on audio features as *Audio Only*, the network trained only on video as *Video Only* and the bimodal network architecture as *Bimodal*.

In order to keep the output of our separate networks analogous to the Bimodal network output, a 50% dropout was also applied after the last hidden layer of our separate modalities (respectively after  $FC_V$  for Video Only and after  $FC_A^2$  for Audio Only). After the dropout we map the predictions to the labels using a sigmoid output layer in the same manner as described in Section IV-C.

One could argue that our bimodal architecture produces better accuracy only because of the concatenation and addition of two fully connected layers ( $FC_F^1$  and  $FC_F^2$  of Fig. 4). In order to verify this assumption, we investigate two other experimental settings. For the first additional setting, we duplicate the 1000-dimensional feature vector produced by layer  $FC_A^2$ , which results in two 1000-dimensional vectors, say,  $v_1$  and

$v_2$ . After that, we provide vectors  $v_1$  and  $v_2$  as input to the Fusion Module. For the second additional setting, we repeat the procedure for the output of layer  $FC_V$ . Results for these experimental settings are presented in the forth and fifth lines of Table I. We call ‘Fused Audio’ the audio network trained together with the fusion module; and, correspondingly, for the video network, we use ‘Fused Video’.

TABLE I  
EVALUATION RESULTS FOR UNIMODAL AND BIMODAL SETTINGS.

Model type	Micro-F1	Hamming Loss	MR
Audio Only	0.496756	0.145811	0.33%
Video Only	0.509954	0.174544	0.31%
Bimodal	0.639498	0.123523	0.48%
Fused Audio	0.441629	0.165682	0.32%
Fused Video	0.466116	0.178303	0.29%
Audio Only + Video Only	0.593649	0.168367	0.30%

Fig. 5 shows the learning curves for all settings (we stopped at 40 training epochs because models were starting to overfit after this limit). It can be seen that training on our bimodal architecture results in faster convergence.

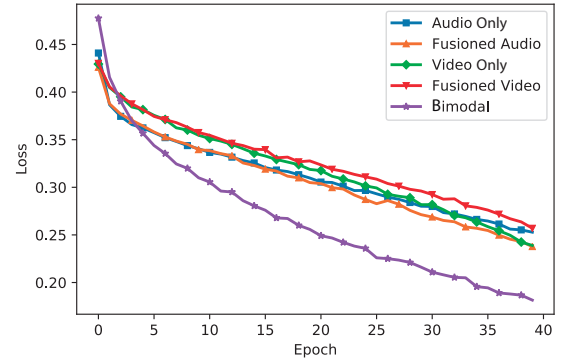


Fig. 5. Learning curves for unimodal and bimodal settings. Our bimodal learning architecture shows a faster convergence than unimodal networks.

In a last experiment, we wanted to verify whether the separate modalities with their predictions are combined produce a better accuracy than the bimodal network. For this experimental setting we extracted the prediction vectors for audio and video networks and applied a element-wise sum for every prediction vector. We call such setting *Audio Only + Video Only*. The results for this experimental setting is presented in the sixth line of Table I.

To assess the results per label, we also evaluated the amount of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for each label of the main network modalities. Audio only is presented on Table II, Video only is presented on Table III and the bimodal architecture results is presented on Table IV. We can observe that the label *Fire* was better recognized by the video network, while *Rain* was better recognized by the audio network. Our bimodal

network surpasses the predictions of the individual networks for both of these labels.

We can also observe that, for the label “Ocean”, the amount of FP were considerably higher in the Video Only network than in the Audio Only network, while the TP rate for the same label was also higher on the Video Only network than on the Audio Only network. On the other hand, the Bimodal network has a much lower ratio between the FP rate and TP high. That indicates that the Bimodal network can correctly predict more labels and thus outperform both separate networks.

TABLE II  
COUNTS FOR THE AUDIO ONLY SETTING.

Label	TP	FP	TN	FN
Wind	80	47	312	93
Thunder	15	2	477	38
Rain	44	24	416	48
Ocean	17	35	435	45
Fire	14	17	455	46
Explosion	18	20	457	37
Gunshot, gunfire	80	9	361	82
Total	268	154	2913	389

TABLE III  
COUNTS FOR THE VIDEO ONLY SETTING.

Label	TP	FP	TN	FN
Wind	112	80	279	61
Thunder	12	16	463	41
Rain	36	33	407	56
Ocean	49	112	358	13
Fire	28	12	460	32
Explosion	22	30	447	33
Gunshot, gunfire	74	33	337	88
Total	333	316	2751	324

TABLE IV  
COUNTS FOR THE BIMODAL SETTING.

Label	TP	FP	TN	FN
Wind	111	53	306	62
Thunder	28	14	465	25
Rain	70	44	396	22
Ocean	41	49	421	21
Fire	42	24	448	18
Explosion	21	14	463	34
Gunshot, gunfire	95	13	357	67
Total	408	211	2856	249

In order to visualize how the network is behaving we formulate images which will best represent each label using Class Model Visualization [35]. In Fig. 6 we present audio and video inputs that try to maximize the output of the network for each label. The left (Fig. 6(a)) subplots represent the inputs for video on frames 6, 12 and 24<sup>3</sup>. In Fig. 6(b) we present the formulated log-scaled mel-spectrogram feature of the recording that maximizes each label. We can see from these visualizations that the network is actually learning the sound events from a relevant time period in the log-scaled

mel-spectrogram feature. As for video, some labels seem to have distinction between frames as time progresses (e.g., *Fire*, *Gunshot*, *gunfire*), which indicates that during training the network has captured an understanding of the label presence over time.

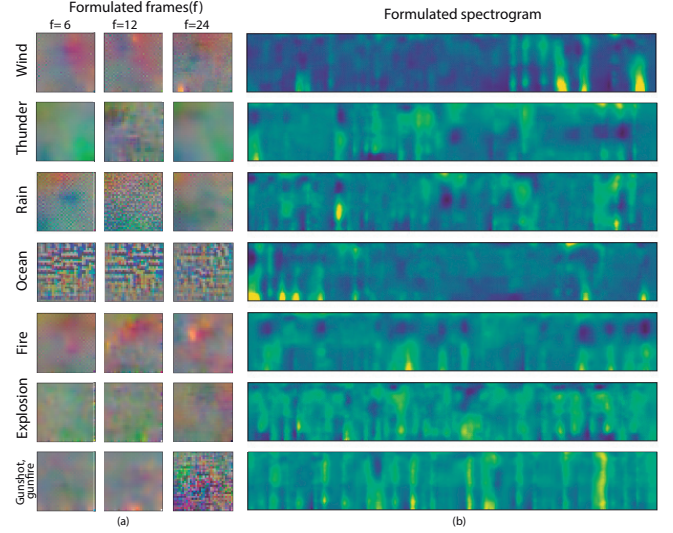


Fig. 6. Class Model Visualization of Audio and video inputs for each label.

## VI. CONCLUSION

Manually identifying the timestamps to start the execution of sensory effects in mulsemmedia applications is often time consuming and error-prone. In this work, we showed how a bimodal deep learning architecture can be trained in order to solve this task.

Our proposed bimodal architecture comprises three modules, two of them for extracting audio and video features from input videos, and a third one, called fusion module, that combines both results to do the prediction of scene components. In order to validate our architecture, we employed the architecture on a subset of the AudioSet dataset targeting the prediction of scene components. The proposed architecture was validated using six different experimental settings. In all cases, our experiments showed that our bimodal architecture performed better.

We also employed a method to extract a dataset tailored for training predictive models for scene components identification. We extracted a subset derived from the AudioSet dataset to train our models. The main challenge of this first version is the lack of relationship between video and audio in some Youtube clips. An example of this is the existence of clips labelled as *Rain* that encompasses video occurrences from inside houses, games, real life scenarios and static images with only sound of rain effects. Another challenge was the training of our networks with multi-labeled data, which means that maybe multiple labels co-exist in the same clip. It is known that the multi-label training method has the potential to find correlations among labels [36]. If a real correlation between the labels is present and desired it can help the learning

<sup>3</sup>We also invite the reader to consult the accompanying video containing the full animation: <https://youtu.be/dTVbsootmiA>

process. On the other hand, if no correlation is desired it could endanger the learning process.

Given the challenges described above, a venue for future work is to apply data cleaning and data augmentation techniques to extend and improve the quality of the dataset presented in this work. As another future work, other learning techniques, such as binary relevance [36], might be explored to seek improvements in the learning process.

We also consider to incorporate deep CCA [37] in the Fusion Module of our learning architecture, in order to take into account correlations between audio and video modalities.

Finally we also plan to publish a scene recognition web service built upon the bimodal learning architecture proposed in this paper.

#### ACKNOWLEDGMENT

The authors would like to thank CNPq, CAPES, and FAPERJ for partially funding this research.

#### REFERENCES

- [1] G. Blakowski and R. Steinmetz, "A media synchronization survey: Reference model, specification and case studies," *Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 5–35, January 1996.
- [2] H. L. Hardman, "Modeling and authoring hypermedia documents," Ph.D. dissertation, Universit t Amsterdam, 1998.
- [3] G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver, "Mulsemedia: State of the art, perspectives, and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 1s, p. 17, 2014.
- [4] M. Waltl, C. Timmerer, and H. Hellwagner, "Improving the quality of multimedia experience through sensory effects," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*. IEEE, 2010, pp. 124–129.
- [5] B. Rainer, M. Waltl, E. Cheng, M. Shujau, C. Timmerer, S. Davis, I. Burnett, C. Ritz, and H. Hellwagner, "Investigating the impact of sensory effects on the quality of experience and emotional response in web videos," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 278–283.
- [6] Z. Yuan, G. Ghinea, and G.-M. Muntean, "Beyond multimedia adaptation: Quality of experience-aware multi-sensorial media delivery," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 104–117, 2015.
- [7] W3C, "Synchronized multimedia integration language - smil 3.0 specification," <http://www.w3c.org/TR/SMIL3>, 2008.
- [8] ITU, "Nested context language (ncl) and ginga-ncl for iptv services," <http://www.itu.int/rec/T-REC-H.761-200904-S>, 2009.
- [9] R. Abreu and J. A. F. dos Santos, "Using abstract anchors to aid the development of multimedia applications with sensory effects," in *Proceedings of the 2017 ACM Symposium on Document Engineering*. ACM, 2017, pp. 211–218.
- [10] —, "Using abstract anchors for automatic authoring of sensory effects based on ambient sound recognition," in *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web*. ACM, 2017, pp. 437–440.
- [11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML '11. USA: Omnipress, 2011, pp. 689–696.
- [12] A. Torfi, S. M. Iranmanesh, N. M. Nasrabadi, and J. M. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.
- [13] W. Feng, N. Guan, Y. Li, X. Zhang, and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 681–688.
- [14] J. Geng, X. Liu, and Y. m. Cheung, "Audio-visual speaker recognition via multi-modal correlated neural networks," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, Oct 2016, pp. 123–128.
- [15] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multi-modal speaker naming," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1107–1110.
- [16] G. Monaci, P. Vanderghenst, and F. T. Sommer, "Learning bimodal structure in audio-visual data," *IEEE Transactions on Neural Networks*, vol. 20, no. 12, pp. 1898–1910, Dec 2009.
- [17] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM '16. New York, NY, USA: ACM, 2016, pp. 978–987.
- [18] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16. New York, NY, USA: ACM, 2016, pp. 281–284.
- [19] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 892–900.
- [20] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Computer Science*, vol. 112, pp. 2048 – 2056, 2017, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, September 2017, Marseille, France.
- [21] H. Z. Sagar Vegad, Harsh Patel and M. Naik, "Audio-visual person recognition using deep convolutional neural networks," *Journal of Biometrics & Biostatistics*, vol. 8, no. 5, pp. 1–7, 2017.
- [22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
- [23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.
- [24] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.
- [25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 448–456.
- [27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [28] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 421–425.
- [29] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.
- [30] L. Tang, S. Rajan, and V. K. Narayanan, "Large scale multi-label classification via metalabeler," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 211–220.
- [31] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda, "Maximal margin labeling for multi-topic text categorization," in *Advances in neural information processing systems*, 2005, pp. 649–656.
- [32] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *European conference on machine learning*. Springer, 2007, pp. 406–417.
- [33] F. Chollet et al., "Keras," 2015.
- [34] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., "Tensorflow: A system for large-scale machine learning," in *OSDI*, vol. 16, 2016, pp. 265–283.
- [35] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
- [36] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [37] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.