Universidade Federal de Minas Gerais

Computer Science Department
Projeto Orientado em Computação II

Project Report

Auditory Annoyance Source Localization in Images via Sound Propagation Estimation

Scientific Research

Student:
Edson Roteia Araujo Junior <edsonroteia@dcc.ufmg.br>

Advisor:
Erickson Rangel do Nascimento <erickson@dcc.ufmg.br>

December 2019

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

UFMG

Abstract

In this work, we aim to perform auditory source localization in images. We want not only to estimate which objects in the scene are making the annoying sound but also to infer how their sounds interact with the environment. To achieve that, we propose a methodology that uses a classification task and Class Activation Mapping for doing the PA localization, and a pipeline composed of a depth estimation neural network and a sound propagation library for estimating the PA propagation on the environment. Our experiments show that our method achieves satisfactory auditory source localization, and it can generate a map that represents how the annoying sound is propagating in the scene.

1 Introduction

When perceiving the world, human beings often rely on multiple modalities to make sense of their experiences. With the growth of video sharing social networks (*e.g.* Youtube), an enormous amount of audio-visual data is available. Thus, many methods are exploring these two modalities from videos to accomplish tasks such as audio separation [1, 2, 3, 4, 5, 6], audio source localization [1, 7, 6, 8] and cross-modal retrieval [8, 9].

Auditory annoyance is linked to inciting negative psychological and physiological effects on people [10, 11]. Knowing where this feature in an image comes from can help intelligent systems take actions to reduce the annoyance for humans in several real-world applications. For instance, intelligent car systems can take control of windows position or vehicle speed to make the ride more pleasant to the driver or passenger if it detects annoying signals coming from outside. The auditory annoyance can be calculated by a metric proposed by *Zwicker et al.* [12], called Psychoacoustic Annoyance (PA). In our work, we aim to combine sound and visual data to localize

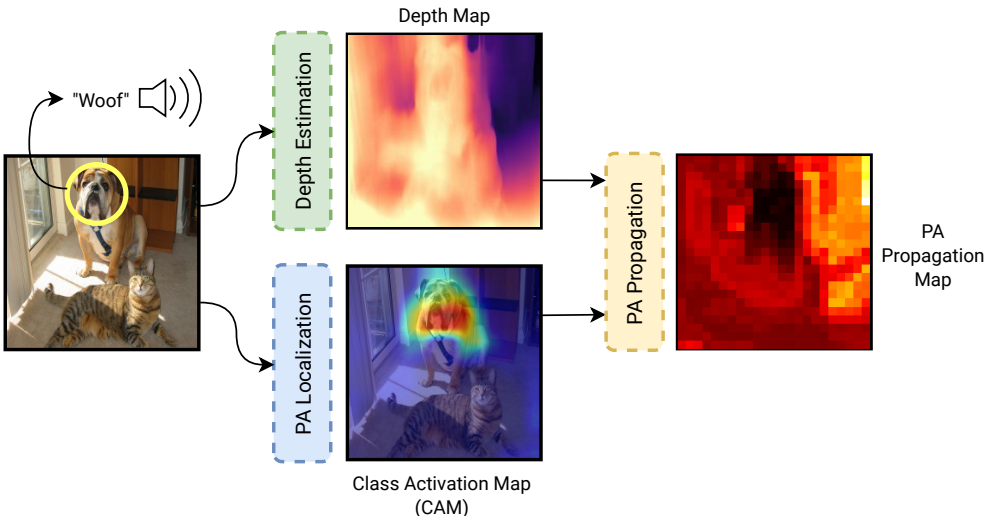

Figure 1: Our ultimate goal is to generate a heatmap per frame, indicating PA levels of the sounds emitted by the regions of the image, taking into consideration the sound propagation dynamics estimation.

annoying sound sources in images. Not only that, we also want to consider the way sound propagates in the environment to create a better understanding of the PA distribution in the scene (see Fig. 1).

Our task is, given an input image and audio pair, to localize where the annoying sounds are coming from. We want to not only to estimate which objects in the scene are making the annoying sound but also to infer how their audio waves propagate in the environment. With that, we can also apply our method in a frame-based manner to creating more pleasant first-person videos, emphasizing the least annoying sound sources.

In POC I, we extended the method of *Zhao et al.* [6] and it led us to a better understanding of the process of localizing sounds in videos. Although we couldn't evaluate the sound propagation, the results have shown that the methodology we extended was still able to separate sounds well with our modification. Even though we believe the approach we used can be improved to get us closer to solving the problem, we invested our efforts in creating a novel PA Propagation Map extraction pipeline.

This document is organized as follows. In Section 2, we review related work in both sound localization and audio-visual source separation tasks. In Section 3, we present our methodology that is able to accomplish this project's goal. In Section 4, we present our experimental setup, as well as both quantitative and qualitative results. Finally, in section 5 we summarize our method and contributions. We also propose new research directions that can use this work as a starting point.

2 Related Work

2.1 Self-supervised Learning

Our method will be based on the idea of learning features by solving a task that can use the structure of the input data as its training labels. In other words, the training process occurs in a supervised manner, but without any human interference on the labels themselves.

2.2 Sound Localization

Recently, there has been much progress in using sound and visual signals to extract information from videos. One of the main tasks in this area is sound localization. This task consists of highlighting regions of an image or set of images that correspond to a given sound signal. Accomplishing this localization can be challenging for many reasons (for instance, the sound source is not present in the image, occlusion, noise, etc.).

Arandjelović and Zisserman [8] performed sound localization in images by using an architecture that takes an image and 1-second audio as input and outputs whether they are correspondents or not. They extract local region-level image descriptors and compute a similarity score between the audio and each descriptor. The authors'

approach can retrieve the sound localization image map by evaluating which region contributed the most to the correspondence score.

The work of *Owens et al.* [1] accomplishes sound localization by using a method called class activation map (CAM). The CAM method assigns a probability value to a space-time video patch, representing how likely it is to be a sound source.

In a different approach, *Zhao et al.* [6] produce a pixel-level sound localization, in which each pixel in the image has a unique sound that is computed by an audio synthesizer network. Their method generates a mask for a pixel to be applied to the input sound wave so that the resulting spectrogram corresponds to the audio of that specific pixel. They achieve sound localization by calculating the sound intensity of the pixels, creating a map that highlights the audio sources.

2.3 Audio-Visual Source Separation

Methods using visual information to help with the sound separation task have become more popular, especially in the speech separation task.

The work proposed by *Ephrat et al.* [3] presents a model for enhancing the speech of desired speakers in a video. The speech enhancement is accomplished by extracting audio and visual features from the input video and producing audio masks that, when applied to the original audio, produce a clean audio signal for each speaker. Their method relies on having all faces in the frame, with a limited position spectrum, without occlusion in order to extract their speech audio signal.

Gao et al. [2] propose a framework to learn object sounds from unlabeled videos. They are able to separate object-level sounds by learning audio bases from an unsupervised step, and then a basis dictionary is built. When a new input video is given to their method, it visually detects the objects and retrieves their respective bases in the dictionary. These bases are used to guide the network to factorize the audio input and produce the separate audio signals for each object.

In *Zhao et al.* [13], the authors improved their previous results by also considering motion of objects in the scene. The method exploits the coherence of signals from both audio and video from a large quantity of unlabeled videos. The results were an overall improvement over the last object localization and sound separation method.

Gao et al. [14] propose a new training paradigm for separating audio sources from unlabeled videos. Their method is able to disentangle sounds in realistic test videos.

Although these methods achieve satisfying results in the sound separation task, they all fail in estimating the sound propagation in the environment. In this work, we aim to generate a auditory annoyance heatmap considering the sound interaction in the environment.

3 Methodology

3.1 Psychoacoustical Annoyance Metric (PA)

One of the first steps of our methodology is to calculate the PA value for sounds. The acoustic annoyance of a sound is related to some psychoacoustical indices described by *Zwicker et al.* [12]. These indices can be briefly described as:

- *Fluctuation and Roughness*: these indices measure the modulation of a signal over the time. A modulated signal with higher values for these indices tends to be more unpleasant.
- *Loudness*: this property is based in perceived loudness and it is based on human subject studies. It measures how loud people with normal hearing perceive a sound.
- *Sharpness*: it is calculated by a weighted sum of specific loudness levels in different bands. A sound with higher sharpness is more unpleasant.

The PA value can be calculated by a funcion of these characteristics as follows:

$$PA = N_5 \left(1 + \sqrt{\omega_S^2 + \omega_{FS}^2} \right), \tag{1}$$

$$\omega_S = \mathbb{1}[S > 0] \times (1.75 - S) log(N_5 + 10), \tag{2}$$

$$\omega_{FS} = 2.78 \times N_5^{-0.4} \times (0.4F + 0.6R), \tag{3}$$

where N is the loudness, N_5 is the 95th percentile of loudness, S is the sharpness, F and R are fluctuation and roughness respectively, and $\mathbb{1}[X]$ is the indicator function, that evaluates the predicate X, returning 1 if it is true and 0 otherwise.

3.2 PA Propagation Map Extraction Pipeline

Our method is based on finding the PA sound source in the image plane and projecting it into a 3-D scene so we can estimate the audio propagation in the real world. The localization is done by training a VGG19 architecture for classifying images into levels of auditory annoyance, using the corresponding audio and its PA value to supervise such training. The 3-D estimation is done by using a Convolutional Neural Network architecture to estimate the depth information from the scene. A sound simulation framework is then used to get the propagated sounds for each 3-D point of the environment that corresponds to a pixel in the image and calculate the PA for them. Figure 2 illustrates the main steps of our approach.

3.2.1 PA Localization

Since one of the essential parts of our methodology is to estimate where the annoying sound source is coming from in the image plane, we propose a localization estimation method based on Class Activation Mapping (CAM). We use a VGG19 [15]

architecture to create a classifier that has as inputs RGB images and predicts how annoying the acoustic environment of the scene portrayed in the image is. The localization is achieved by training the classifier using the PA value for the corresponding audio from each image as supervision. After our model is trained, we extract the CAM from the last convolutional layer using the method from [16]. The Grad-CAM method uses the gradients backpropagated into the network, corresponding to the class it is predicting, to estimate the most activated features in the image. Our premise is that if the model learns well enough, there should be a correspondence between objects in the scene with the acoustical annoyance classification. These objects or regions of the image are a good estimation of where the sound source is coming from. The activation map extracted gives a rough estimation of where the object emitting the annoying sound is.

PA Estimation Module. In order to supervise our training, we must extract the PA values from each sound in the training dataset. We do that by creating a module that receives audio signals of variable length and outputs a PA value corresponding to each audio signal. The PA is calculated using equations 1-3 for each second of the audio. The final value is the average the annoyance levels from every second of the input sound signal.

Peak Finding Module. After the CAM is extracted, we use a peak finding algorithm that takes a two-dimensional array and finds all local maxima by simple comparison of neighboring values. In this work, we are limiting the number of peaks to one, and thus, by taking the highest peak, we always suppose that there is only one annoying sound source in the scene.

3.2.2 Depth Estimation

In order to move towards understanding the 3-D scene, we used the depth estimation model proposed by *Godart et. al.*, *monodepth2* [17]. The used network takes a single color input I_t and produces a depth map D_t using a U-Net as the main component. Given that the model was trained using a stereo camera setup as supervision, and given the parameters of the cameras used for training, we are able to generate our depth map in meters. In this way, for each pixel (x, y) in the input image, we have its value Z for the estimated distance from the image plane.

3.2.3 3-D Peak Localization

Given that in this step of the pipeline, both peak localization in the image plane and the depth estimation are calculated, we now want to estimate where the peak is in the real world scene. It is important to note here that these are all rough estimates of the scene understanding.

From camera to world coordinates. Since we have a depth estimation from each pixel in the image, we can estimate their relative positions in the world if we determine a focal length for the camera capturing that scene. By using an arbitrary

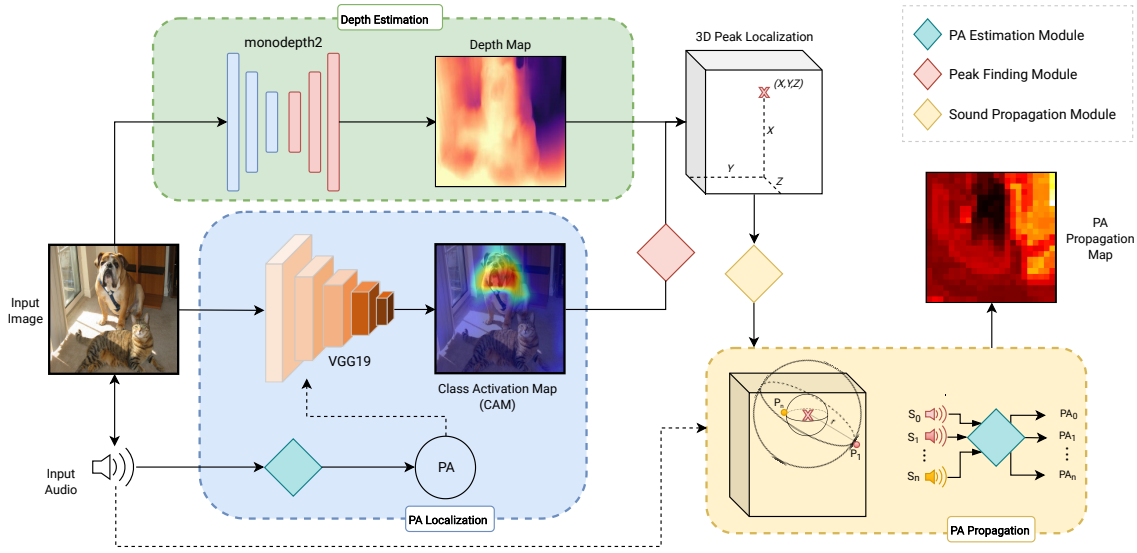

Figure 2: Our methodology is composed of three main stages: We first localize the sound source by extracting the class activation map from the last convolutional layer of a VGG19 that was trained to classify images into levels of Psychoacoustic Annoyance (PA). Then, we extract depth estimation values for each pixel of the input image and create a 3-D map of all scene points. Lastly, we use the 3-D information to propagate the sound using a sound simulation environment and calculate the PA from the sounds generated for each pixel to compose our final PA Propagation Map.

focal length of the camera, we can use triangle similarity calculation to estimate the final (X, Y, Z) coordinates of each pixel in the real world.

Sound Propagation Module. We use the *pyroomacoustics* library [18] to do the sound propagation simulation. The method used receives as input the sound source position and microphone positions in 3-D space, as well as the original sound wave. We create an anechoic room that is large enough to fit the points estimated from the input image. We do that by creating a cubic room of initial size $1m^3$, and we double the size of its dimensions until all the points are contained in it. A simulation is then made to generate the propagated audio in each microphone, resulting in an audio wave per 3-D point.

3.2.4 PA Propagation

Finally, we calculate the PA values from each of the pixels' audios, and we back-project the values into a heatmap. This final heatmap is what we call our PA Propagation Estimation Map.

4 Experiments

In this section, we investigate the performance of our method evaluating it both qualitatively and quantitatively on different image-audio pairs.

4.1 Experimental Setup

Dataset We compose our testing dataset for the framework from images extracted from the Sound Localization dataset proposed by [19]. The dataset is composed of 5,000 image-sound pairs with bounding box annotations from 3 different annotators each. Each annotation is a bounding box and a label that defines it as "object" or "environment", the human perception of whether the sound is coming from an object present in the scene projection in the image or from an outside and more abstract source, such as the sound of rain hitting the floor or traffic noises.

Metrics To evaluate our localization method, we compute how many times our estimated annoying sound source is inside the annotated bounding box from the image. Since we only estimate one annoying sound source in the scene, and it is represented by a single pixel, we check if the pixel coordinates in the image is inside the annotated bounding box, consisting of a hit.

Implementation Details In our experiments, we use an extension of the UFMG Dataset [20], proposed in POC I, to train our localization network. The dataset is composed of 11 egocentric videos with approximately 20,000 frames each. To compose our training data, we randomly selected 5,000 frames from each video and calculated the PA for its corresponding audio segment. We ended up with a training set of 55,000 image-PA pairs. To simplify our learning process, we categorize frames based on their annoyance value range. We use the pre-trained weights from the VGG19 and fine-tune the classification model by changing the last fully connected layer into having as output a 10-class probability array.

4.2 Results

Quantitative Results Table 1 shows the results in terms of the hit rate in our test set. For our method to get a hit, the predicted localization, i.e., the pixel with the highest value from the CAM, has to be inside the annotated bounding box. Although our method is straightforward to train and requires almost no modification from existing architectures, we were able to fairly localize sound sources in the scene with an average higher than 50% for all subsets. We emphasize that our method is not concerned with every sound source themselves but only the annoying ones. That means that we can look for better ways to analyze these localization results based on the PA feature.

Qualitative Results Figure 3 depicts our qualitative results. For each input image, we show the CAM, Depth Estimation, and PA Propagation Map estimated by our method. The results here show that we can create good quality sound source localization estimation and, given a satisfactory depth estimation, we are able to create a map that estimates how high the PA is for each scene point. We emphasize that the final map is heavily dependent on both peak localization and depth estimation processes. If one of these two methods fail, either the sound

	Regions Hit	Total Regions	Hit Rate (%)
S1	231	500	46.2
S2	247	500	49.4
S3	289	500	57.8
S4	284	500	56.7

Table 1: Quantitative results of our localization method. We selected 4 sets from our testing dataset to compose our localization evaluation set. For each image on the set, there is exactly one annotated ground-truth region. Our method has an average of 52.53% in hit rates for all sets.

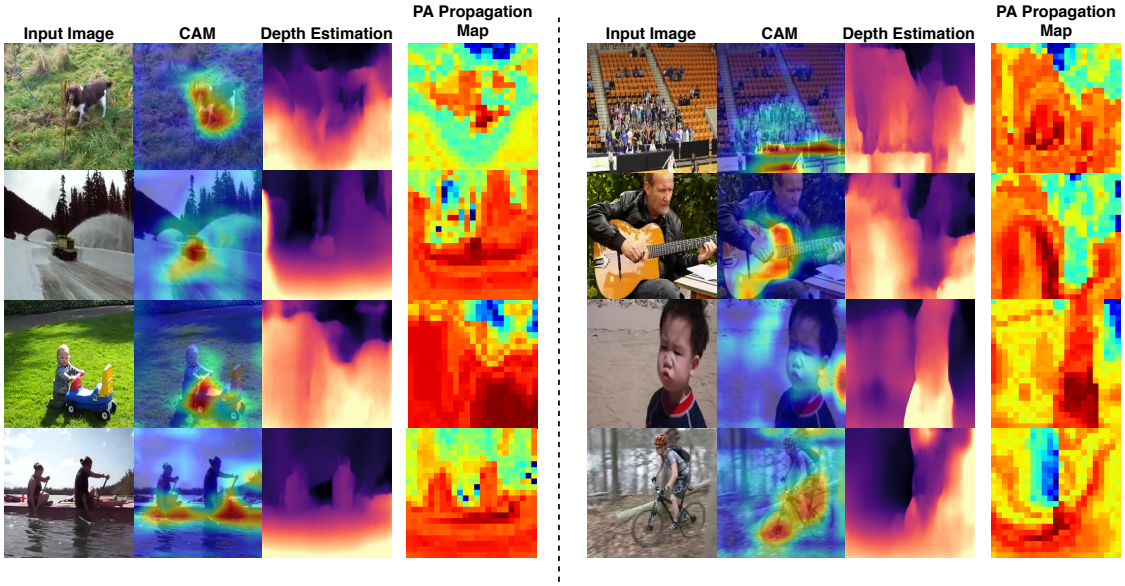

Figure 3: Qualitative results for our pipeline steps in 8 image-sound pairs from the testing dataset. The source localization is shown in the CAM column, and the Depth Estimation is used to recreate a 3-D environment. The ultimate goal of our methodology is to generate a PA Propagation Map of the scene that can enable the understanding of the acoustic scenario.

source will be in the wrong location for the sound propagation estimation, or the environment will be poorly recreated.

5 Conclusions

In this work, we proposed a novel methodology to estimate the Psychoacoustic Annoyance localization and propagation in the environment using a single input image and audio pair. We used ideas from sound localization methods to propose our PA localization method using a classification task and Class Activation Mapping. We also present a method of using depth estimation information to estimate sound propagation in an environment given a 2-D image. Our results show that the CAM extraction can be used to localize annoying sound sources in input images by training a CNN classifier based on the scene annoyance levels. We also show the benefits of

using the 3-D estimation to understand the scene better and propagate the audio and its features accordingly.

Future Work We understand that to build a more robust method, we can consider multiple sound sources in our estimation pipeline since our sound propagation module can capture the interaction between multiple audios. As a future research direction, we expect that the time consumption can be reduced by utilizing our method as a supervisor for an end-to-end architecture that is able to estimate the PA propagation directly.

References

[1] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *CoRR*, abs/1804.03641, 2018.

[2] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[3] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, abs/1804.03619, 2018.

[4] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement using noise-invariant training. *CoRR*, abs/1711.08789, 2017.

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *CoRR*, abs/1804.04121, 2018.

[6] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh H. McDermott, and Antonio Torralba. The sound of pixels. *CoRR*, abs/1804.03160, 2018.

[7] Arda Senocak, Tae-Hyun Oh, Jun-Sik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. *CoRR*, abs/1803.03849, 2018.

[8] Relja Arandjelović and Andrew Zisserman. Objects that sound. *Lecture Notes in Computer Science*, page 451–466, 2018.

[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017.

[10] R. Rylander. Physiological aspects of noise-induced stress and annoyance. *Journal of Sound and Vibration*, 277(3):471 – 478, 2004. Fifth Japanese-Swedish Noise Symposium on Medical Effects.

[11] W. Babisch et al. Health status as a potential effect modifier of the relation between noise annoyance and incidence of ischaemic heart disease. *Occupational and Environmental Medicine*, 60(10):739–745, 2003.

[12] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.

[13] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. *CoRR*, abs/1904.05979, 2019.

[14] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. *CoRR*, abs/1904.07750, 2019.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.

[17] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[18] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. 10 2017.

[19] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] Michel Melo Silva, Washington Luis Souza Ramos, Joao Pedro Klock Ferreira, Mario Fernando Montenegro Campos, and Erickson Rangel Nascimento. *Towards Semantic Fast-Forward and Stabilized Egocentric Videos*, pages 557–571. Amsterdam, NL, Oct. 2016.