A Multimodal Approach to Mapping Soundscapes

Tawfiq Salem salem@cs.uky.edu
Menghua Zhai ted@cs.uky.edu
Scott Workman scott@cs.uky.edu
Nathan Jacobs jacobs@cs.uky.edu

Department of Computer Science
University of Kentucky

Abstract

We explore the problem of mapping soundscapes, that is, predicting the types of sounds that are likely to be heard at a given geographic location. Using a novel dataset, which includes geo-tagged audio and overhead imagery, we develop an approach for constructing an aural atlas, which captures the geospatial distribution of soundscapes. We build on previous work relating sound to ground-level imagery but incorporate overhead imagery to overcome the limitations of sparsely distributed geo-tagged audio. In the end, all that we require to construct an aural atlas is overhead imagery of the region of interest. We show examples of aural atlases at multiple spatial scales, from block-level to country.

1. Introduction

The visual appearance of a place and its soundscape, the totality of sounds one hears in a location, are inextricably linked. For example, in an urban environment, such as on a busy street corner, you can expect to hear honking, people talking, and, potentially, a siren. In contrast, in a rural environment, such as a forest, you could expect to hear animals chattering, leaves rustling, and perhaps the sound of rushing water. Given a photograph, humans have the ability to imagine the sounds they might hear in that moment.

Studies have shown that environmental noise affects social behavior [8], among other things. Basner et al. [4] summarize research related to noise exposure, including auditory and non-auditory health effects such as reduced cognitive performance and sleep disturbance. Models capturing the relationship between sound and specific locations could be used, for example, to help people decide where to live, or where to place sound barriers.

The objective of our work is to develop methods for understanding the types of sounds that could be heard at a specific geographic location. Several recent works have taken advantage of the synergy between sound and visual appearance to learn better representations. Aytar et al. [3] leverage two million unlabeled videos to learn a state-of-the-art sound representation for acoustic classification. Owens et al. [6] incorporate ambient sounds as a supervisory signal in order to learn visual representations. Most similar to our work, Aiello et al. [2] proposed a method for constructing sound maps by using sound-related image tags on a large set of geo-referenced ground-level imagery. This method requires high-quality image tags, which aren't always available, and performs poorly when ground-level imagery is sparsely distributed, such as away from major tourist landmarks.

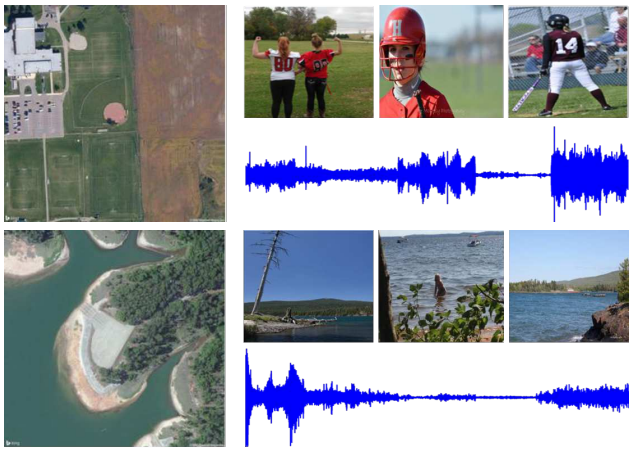

Figure 1. We propose a multimodal approach for relating overhead image appearance with sounds in order to map soundscapes. (left) Overhead image; (right) Similar ground-level images and sounds output by our method.

We take a different approach and explore the problem of generating a location-dependent sound model. Our approach builds upon recent advances in both ground-level and overhead image understanding. A key element of our approach is that we learn a joint feature representation between sound, ground-level, and overhead image appearance (Fig. 1). A unique advantage of our approach is that it enables us to generate a location-dependent sound map (or an aural atlas) using only overhead imagery, which is available at most locations.

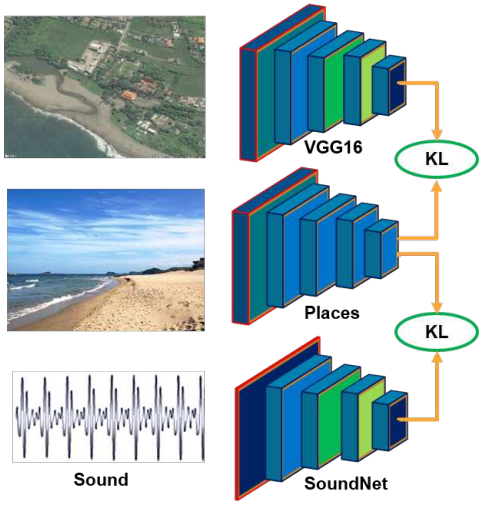

Figure 2. An overview of our network architecture.

2. Cross-View Aural Mapping

The objective of our work is to construct a map of the aural environment, which we represent as a conditional probability distribution, $P(s|l)$, where l is the geographic location and s represents the sound. In this work, we explore a novel approach, conditioning our aural map on the overhead imagery of a location, $P(s|I(l))$, where $I(l)$ is an overhead image of location l. This is a promising approach because many visual features that relate sound to location are visible from above. Furthermore, high-resolution overhead imagery is available across the globe and is updated frequently. We further factorize the distribution as:

$$P(s|I(l)) = \sum_c P(s|c)P(c|I(l)),$$

where c represents a cluster of related sounds.

Our approach consists of three phases: learning a suitable feature space, clustering sounds, and learning to predict a distribution over sound clusters from the overhead imagery using a convolutional neural network.

2.1. Cross-View Sound Dataset

To support our work, we constructed a dataset of geo-tagged sounds and co-located overhead images, which we refer to as the Cross-View Sound (CVS) dataset. We collected 23,308 geo-tagged audio files from FreeSound[1], a popular crowd-sourced repository. For each audio file, we downloaded the corresponding overhead image from Bing Maps (scale 0.60 m/pixel). Analysis of the geolocation associated with the sound files reveals that the sounds are recorded from around the world, with more sounds recorded in Europe and U.S than other parts of the world. Further, examining the tags associated with the sound files shows that the sounds cover a wide range of human and natural aspects. For our experiments, we filtered out sounds that were shorter than 2 seconds and sounds for which there was no overhead imagery available at the selected scale. This results in 15,773 sounds and their corresponding overhead images.

[1] https://freesound.org

2.2. Learning a Shared Feature Space

In this phase, our goal is to learn a shared feature representation that is suitable for our task. Specifically, we want a feature representation that can jointly describe audio and overhead imagery. To do this, we propose a convolutional neural network (CNN) architecture that relates sounds with co-located overhead images. Our approach builds on recent work that targets these two subproblems individually. An overview of our architecture is shown in Fig. 2.

To extract audio features, we use SoundNet [3], a deep convolutional architecture for sound recognition, trained by transferring knowledge from existing visual recognition networks. To train SoundNet, images from unlabeled videos are passed through the *Places* network [10] (while different *Places* models are available, we always refer to the one used to train SoundNet), and the output distributions are used as the target label for a network that takes as input the corresponding audio file. Given an audio file, the output of SoundNet is a distribution over 401 visual scene categories. The resulting network performs remarkably well, despite being trained without any manually annotated audio files.

To learn the overhead image feature representation, we use a multimodal training approach similar to SoundNet and Workman et al. [9]. Specifically, we learn to predict a distribution over ground-level scene categories from overhead imagery. Each ground-level image is labeled using the *Places* network [10], generating a distribution over 401 scene categories. We then train a VGG-16 [7] network to predict these distributions using only the overhead image, minimizing the KL-divergence. We trained the network on the CVUSA dataset [9], which contains approximately 1.5 million geo-tagged pairs of overhead and ground-level images. The network is initialized to the weights of the *Places* network, and optimized using *Adam* with learning rate of 0.001 for 5 epochs.

This process results in a shared feature representation that allows the direct comparison of three different modalities: audio, ground-level imagery, and overhead imagery. We could, for example, use an image retrieval approach to identify sounds related to an overhead image. The problem with this approach is that the sounds close to the overhead image in the feature space will all be similar, and therefore potentially not representative of the diversity of sounds one could hear in a particular area. To overcome this, we introduce a clustering approach to group sounds, which we then use to map soundscapes.

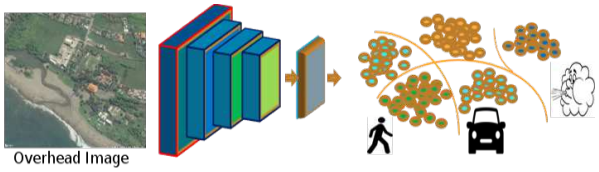

Figure 3. The model architecture for predicting a distribution over sound clusters from an overhead image.

2.3. Clustering Sounds

We group the sounds into a discrete set of clusters using hierarchical clustering [5]. For a given image-sound pair, we extract the predicted distributions over the 401 scene categories for each modality and concatenate them to form an 802-dimensional vector. Then, this concatenated representations is used as input for clustering. The result of this process is a set of clusters $C = c_1, \ldots, c_k$. Finally, we filter out small clusters (less than 500 sounds), leaving 10 clusters. In the following section, we describe our process for estimating the conditional distribution over sound clusters for a given location.

2.4. Predicting Sound Clusters from Overhead Imagery

We assign each sound to a unique cluster and treat the cluster assignment, c_i, as the label of a given location, l_i. For each location, we obtain the co-located overhead image, $I(l_i)$, and train a CNN to predict the sound cluster, c_i, from the image. We fine-tune the network described in Section 2.2, adding a fully connected layer at the end with ten outputs (Fig. 3). We minimize the cross-entropy loss using *Adam* with a learning rate of 0.001 for 20 epochs. We now have all of the components of our model and can use $P(c|I(l))$ to visualize soundscapes.

3. Experiments

We evaluated our approach both quantitatively and qualitatively using a TensorFlow [1] implementation. We begin with an analysis of the shared feature space.

3.1. How good is our feature space?

For a given overhead image, we extract the output distribution over scene categories and identify the closest sounds in CVS and the closest ground-level images in CVUSA, using KL-divergence. Several qualitative examples are shown in Fig. 1. The leftmost column shows the overhead image and the right columns show the top three ground-level images above the top three sounds. For example, in the bottom row, the overhead image is of a lake, and the three closest ground-level images appear to be captured on or near a lake. The results are similar when listening to the closest sounds; in Fig. 1 (top) the most similar sound contains people cheering. The predicted sounds, dataset, and more results will be made available online at http://cs.uky.edu/~salem/audio-mapping/.

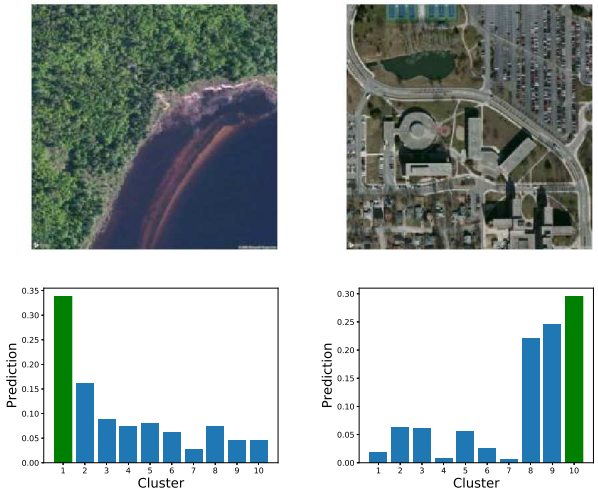

Figure 4. Given an overhead image (top), our model outputs a distribution over sound clusters (bottom).

Table 1. Quantitative performance of different networks.

Network	Precision	Recall	F_1-score
sound	0.24	0.19	0.19
joint	0.51	0.34	0.36

3.2. What is the best way to cluster?

We compare our approach for clustering the sounds against a baseline approach using only sound features. As described in Section 2.4, we train two models on the two different clustering approaches. For evaluation, we split the CVS dataset into 90% training and 10% testing. The resulting test set contains 1,578 sounds and corresponding overhead images.

For a given overhead image, each model outputs a probability distribution over the sound clusters. The precision, recall and F_1-score for these two models are shown in Table 1. The model that was trained on clusters generated from joint features achieved better performance compared to the model trained on sound features. The superior performance can be attributed to the fact that the clustering approach based on joint features takes into account the semantic relation between overhead images and the corresponding sounds. Fig. 4 shows the output distributions over the 10 clusters for two test images.

3.3. Visualizing An Aural Atlas

Using the trained CNN model to predict a distribution over sound clusters from an overhead image enables us to construct sound maps at various spatial scales: block level, city level, and country level.

Block level: Consider the overhead image on the left of Fig. 5 which contains beach, water, roads, and buildings. Clearly the sounds at these places would be different. For

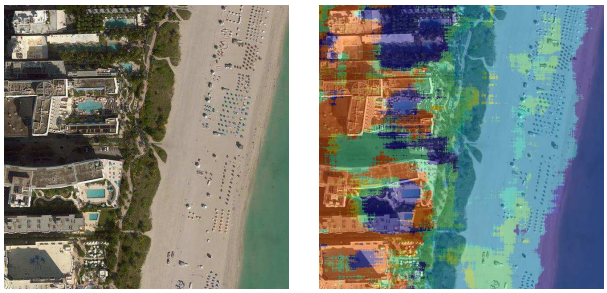

Figure 5. Block-level audio mapping: (left) An overhead image of a small geographical region on Miami beach. (right) A per-pixel labeling of sound clusters.

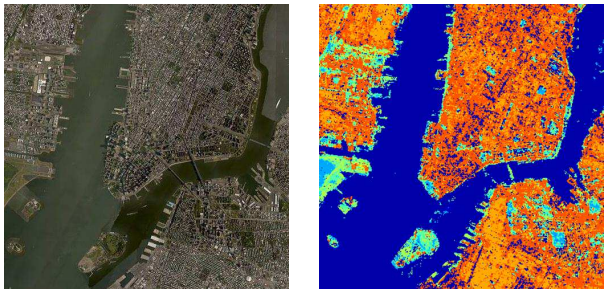

Figure 6. City-level audio mapping: (left) An overhead image covering New York City. (right) A per-pixel labeling of sound clusters.

every pixel in the image, we downloaded the corresponding overhead image and used our network (Section 2.4) to predict the distribution over sound clusters. We show the results of our approach in Fig. 5 (right), as a per-pixel labeling where the color represents the most likely sound cluster (e.g., blue = water-related sounds and orange = traffic sounds). The color coding is the same for the next two spatial scales.

City level: Here we apply the same technique to a larger geographic area. Fig. 6 shows the aural atlas for a portion of New York City. Note how the majority of the urban areas are colored orange and the water areas are dark blue.

Country level: Finally, we demonstrate the results of our method at the country level. We used 500,000 overhead images randomly sampled from the CVUSA dataset and extracted the sound cluster prediction with our trained model. Fig. 7 shows the results. Note the orange regions covering the major metropolitan areas.

4. Conclusion

We created a location-dependent model of sound conditioned on overhead imagery. We showed how our model could be used for sampling a set of sounds that you would hear at a given location and to generate maps of soundscapes at varying spatial scales. To the best of our knowledge, our work is the first to model the relationship between overhead imagery and sound. In the future, we will extend our work to include time, as the sounds you might hear at a location are highly time dependent.

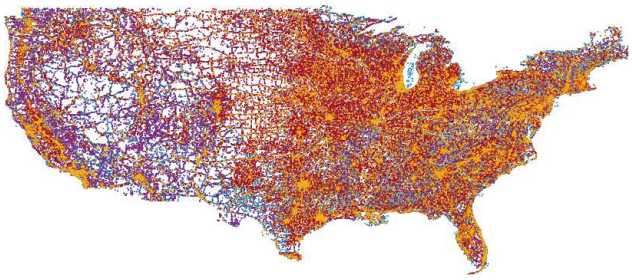

Figure 7. Country-level audio mapping: visualizing the sound clusters over USA. Gaps (white) are regions where the CVUSA dataset does not have imagery.

Acknowledgments We gratefully acknowledge the support of an NSF CAREER award (IIS-1553116).

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 3

[2] L. M. Aiello, R. Schifanella, D. Quercia, and F. Aletta. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science*, 3(3):150690, 2016. 1

[3] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016. 1, 2

[4] M. Basner, W. Babisch, A. Davis, M. Brink, C. Clark, S. Janssen, and S. Stansfeld. Auditory and non-auditory effects of noise on health. *The Lancet*, 383(9925):1325–1332, 2014. 1

[5] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416. ACM, 2000. 3

[6] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016. 1

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2

[8] S. Stansfeld, M. Haines, and B. Brown. Noise and health in the urban environment. *Reviews on environmental health*, 15(1-2):43–82, 2000. 1

[9] S. Workman, R. Souvenir, and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 2

[10] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017. 2