Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230664

Identifying Patterns Between Acoustic Environment and Visual Landscape Through Semantic Segmentation Based on Deep Learning

Riccardo Giacalone [a,1], Carlos Guerrero-Mosquera [a,1] and Xavier Sevillano [a,1]
[a] *HER - Human-Environment Research Group, La Salle - Universitat Ramon Llull, Barcelona, Spain, Quatre Camins, 30, 08022 Barcelona, Spain*
ORCiD ID:
Riccardo Giacalone https://orcid.org/0009-0008-0581-4990
Carlos Guerrero-Mosquera https://orcid.org/0000-0001-8265-3651
Xavier Sevillano https://orcid.org/0000-0002-6209-3033

Abstract. This work is part of the research project "Sons al Balcó" conducted by La Salle - Universitat Ramon Llull, which examines the impacts of noise pollution on human perception and mental health, specifically focusing on the perception of noise in Catalonia during the lockdown in 2020 and the return to normalcy in 2021. The purpose of this research is to identify patterns between the soundscape and the visual landscape of participants' environments. To achieve this, we have developed a pipeline to automatically analyse the visual landscape of participants' environments by semantically segmenting the keyframes of their videos using deep neural networks. Specifically, we use the SegFormer model, a Transformer-based framework for semantic segmentation that integrates Transformers with lightweight MLP decoders. This pipeline facilitates the efficient and accurate identification of different objects, to understand the complex relationships among the acoustic environment, visual landscape, and human perception. We expect that our findings will offer insights into the design of urban and suburban areas that promote well-being and quality of life.

Keywords. Artificial Neural Networks, Artificial Vision, Image Processing, Machine Learning, Deep Learning, Transformers, Semantic Segmentation

1. Introduction

In recent years, there has been a rise in interest in the effects of noise pollution on human perception and mental health. The project "Sons al Balcó" conducted by La Salle - Universitat Ramon Llull seeks to investigate the effects of the lockdown during the COVID-19 pandemic on human perception of the acoustic environment in Catalonia [1]. This paper aims to understand the impact of COVID-19 lockdowns on noise perception

[1] Corresponding Author:
Riccardo Giacalone; E-mail: ricardo.giacalone@students.salle.url.edu
Carlos Guerrero-Mosquera; E-mail: carlos.guerrero@salle.url.edu
Xavier Sevillano; E-mail: xavier.sevillano@salle.url.edu

due to changes in human activity and mobility patterns and analyzes the relationship between participants' perception of the acoustic environment and the landscape of their surroundings using semantic image segmentation.

2. Datasets

Two databases were used in this study: Cityscapes and Sons al Balcó. Cityscapes is a benchmark dataset for semantic segmentation [2], which comprises high-resolution images of urban environments captured across 50 cities in Germany. This dataset includes pixel-level annotations for 30 distinct object classes [3]. For this study, we used the 19 classes and two files which contain RGB images and ground truth annotations. To prepare the data and convert the annotations the label images, we assign a unique grayscale value to each class label and create grayscale label images where each pixel corresponds to a specific class label.

The Sons al Balcó database is a comprehensive dataset that was collected during the COVID-19 lockdown in Catalonia. The data was obtained through a socio-acoustic digital participatory survey, which was made available to Catalonian citizens via social media and the press. The survey included a series of questions regarding socio-demographics, residential soundscape quality, and individual positive and negative perceptions prior to and during the lockdown. Participants were also given the option to contribute videos of their residential soundscapes.

3. Experimental setup

To ensure the quality of the data, each video was manually reviewed before analysis to eliminate those with insufficient lighting, private information, or a singular focus on an object. After this filtering process, 188 videos from 2020 and 165 videos from 2021 were selected for further analysis.

Next, we evaluated the performance of five semantic segmentation models on the Cityscapes dataset using the MMSegmentation framework [4]. To develop the models, we built on top of the MMSegmentation library, which offered the training and testing pipelines needed for this task and was commonly used in the development of these models. The models were evaluated based on the mean Intersection over Union (mIoU) metric, a commonly used metric in the evaluation of segmentation models.

After evaluating the models, we selected the best performing semantic segmentation model to perform inference on videos from both the 2020 and 2021 Son al Balcó datasets. For each video, the average presence of every class was computed by averaging the predicted classes from the segmented frames. Then, we created mappings between video files and segmentation results to process the data.

Finally, the survey data was merged with the segmentation data for each video using a Python Pandas DataFrame, and Pearson and Spearman correlations were plotted to examine the relationships between the variables. Finally, an attempt was made to interpret the correlations using a logistic regression model and examining its coefficients.

4. Results and Discussion

To compare the selected Cityscapes models, they were all evaluated on the 500 validation images set and ranked by the mean Intersection over Union metric.

Table 1. Comparison of Cityscapes Models on Validation Set - Comparison of selected Cityscapes models based on Mean Intersection over Union (mIoU) scores on the 500-image validation set. Models are ranked in descending order of mIoU scores: SegFormer (82.25), OCRNet (81.35), HRNet (80.67), DeepLabv3 (80.48), and DeepLabv3+ (80.46).

Model name	mIoU
SegFormer	82.25
OCRNet	81.35
HRNet	80.67
DeepLabv3	80.48
DeepLabv3+	80.46

The table shows that SegFormer is the clear winner in terms of performance with an mIoU of 82.25%.

The next result shows the Spearman correlation between the noise perception reported in the survey and the percentage of presence of the class labels in the recorded videos. The first figure corresponds to the surveys taken in 2020, while the second in 2021.

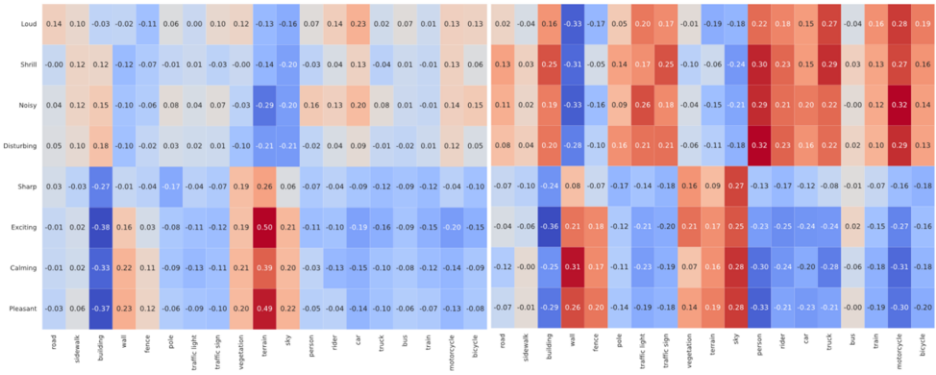

Figure 1. Spearman correlation between the 2020 (left) and 2021 (right) segmentation classes and perception adjectives. Note that the Spearman correlations do not show strong relationships; however, it can be noticed that the relationships differ significantly between the 2020 and 2021 surveys. In 2020 the strongest positive correlations are between positive attributes and "terrain". In 2021 the highest correlations occur with the class "person" and with vehicles and they are all with negative attributes.

To introduce more insights into the analysis with significant correlations, another approach was performed by training a logistic regression model to classify the class label data into a positive or negative perception.

The results indicate that SegFormer performs better in terms of performance metrics. However, in addition to its top-performance, this model offers other range of features that make it an interesting choice for this type of problem, including: (i) It is not a CNN based model, but it is based on an encoder-decoder architecture, (ii) it uses Mix-FFN layers as positional embeddings, addressing the drawbacks of resolution-dependence, redundancy, and inefficiency, (iii) uses outputs multiscale features due to hierarchically structured transformer encoder, (iv) avoids complex decoders combining both local attention and global attention to render powerful representations.

The step of fitting a logistic regression model and analyzing the coefficients was performed after observing a high correlation between the 4 negative attributes ("Loud",

"Shrill", "Noisy", and "Disturbing") and the four positive attributes ("Sharp", "Exciting", "Calming", "Pleasant"). A negative coefficient means that it influences negatively in the calculation of the perception. A positive coefficient is the opposite, it positively affects the overall perception. The results from the logistic regression model trained on the 2020 and 2021 datasets show that the coefficients for 2020 are different than the coefficients for 2021. This is because the class labels can include more than one object, like "rider", which could be a person riding a motorbike or a bicycle.

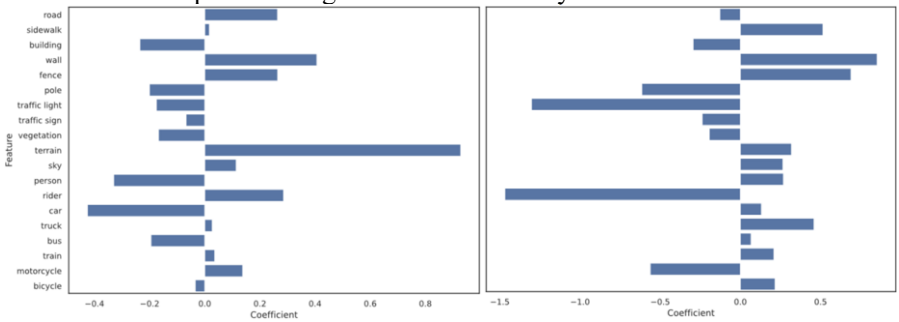

Figure 2. Logistic regression model coefficients for 2020 data (left) and 2021 data (right). Note that 2020 the feature with highest coefficient is "terrain", which contributes the most to a positive perception. About the negative perceptions, "car" is the feature with the lowest coefficient, meaning that negatively influences the overall perception of noise if present in the video. In 2021 the strongest coefficients are negative. "Traffic light" and "rider" are twice and thrice greater than the highest positive coefficients. Notice how rider in 2020 was a positive feature, while in 2021 it is the most negative coefficient; this is because rider includes several categories. In 2020 there were probably more bicycles and electric scooters than people riding a motorbike, due to the confinement. Also, in 2020 "car" is the most negative feature while in 2021 it is slightly positive. There could be several reasons for this divergence, for instance a car in 2020 could be heard from further since there were less cars and could impact the noise perception more than a car in 2021 as there were more cars on the road.

In conclusion, this study indicates that noise perception is not solely dependent on the visual features, as evidenced by the correlations and coefficients of the logistic regression model. Rather, it is also influenced by contextual factors. Specifically, our results suggest that the year 2020 played a significant role in shaping the importance of the various features. To improve the project, it would be interesting to create a model that predicts noise perception more reliably by taking into consideration the sound present in the videos.

Acknowledgment

The research in this paper was supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya (grant 2021 SGR01396) and the Secretaria d'Universitats i Recerca from the Departament d'Empresa i Coneixement (Generalitat de Catalunya) and Universitat Ramon Llull (grant 2020-URL-Proj-054).

References

[1] LaSalle. Ramon Llull University. SONS AL BALCÓ - Map of the Soundscape of the Lockdown in Catalonia. 20 May 2020. [Online]. Available: https://www.salleurl.edu/en/research/project/sons-al-balco.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. 6 Apr 2016. [Online]. Available: https://arxiv.org/abs/1604.01685.

[3] Cityscapes Dataset – Semantic Understanding of Urban Street Scenes. 17 Oct 2020. [Online]. Available: https://www.cityscapes-dataset.com/.

[4] M. Contributors. OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. [Online]. Available: https://github.com/open-mmlab/mmsegmentation.