applied sciences

Article

Exploiting Balcony Sound Atmospheres for Automatic Prediction of Floors with a Voted-Majority Approach Based on Neural Networks

Hengameh Pirhosseinloo † and Massih-Reza Amini *,†

Computer Science Laboratory, LIG, Université Grenoble Alpes, 38400 Saint-Martin-d'Hères, France
* Correspondence: massih-reza.amini@univ-grenoble-alpes.fr
† These authors contributed equally to this work.

Abstract: This paper tackles the challenging task of automatically predicting the floor of a balcony based on a sound bound of two minutes recorded on that balcony. The sound fragments are typical of the environment, as nothing out of the usual can be heard in them. However, there is a good probability that, when hearing a fragment in quiet surroundings with consistent metropolitan background noise, it would not be straightforward to determine or estimate the floor height. In our experiments, it was found that sound chunks lasting 5 s can be identified with high accuracy even with a small number of training samples. In addition, when using a late fusion strategy to combine the outputs of classifiers trained on two modalities of the sound tracks, the floors of these bands are perfectly correctly classified. This result was consistent throughout all twenty tests when training and test sets were chosen at random, supporting the viability of the suggested method.

Keywords: soundscape; balcony; machine learning; sound recognition

Citation: Pirhosseinloo, H.; Amini, M.-R. Exploiting Balcony Sound Atmospheres for Automatic Prediction of Floors with a Voted-Majority Approach Based on Neural Networks. *Appl. Sci.* **2023**, *13*, 5834. https://doi.org/10.3390/app13105834

Academic Editor: Masayuki Takada

Received: 19 March 2023
Revised: 5 May 2023
Accepted: 6 May 2023
Published: 9 May 2023

1. Introduction

The Organization for Economic Co-operation and Development (OECD) (https://www.oecd.org/fr/env/indicateurs-modelisation-perspectives/, accessed on 5 May 2023) predicts that over 70% of the world's population will reside in cities by 2050. However, given that space is finite and bounded and because driving is a significant contributor to global warming, urban expansion is reaching its limits. In this context, cities must choose alternative development paths, particularly in the housing sector, to boost density and promote the growth of new ecological neighborhoods.

A dense habitat, however, runs against the generally held notion of private residence and is by no means favored by inhabitants. Furthermore, from the standpoint of communal housing, especially social housing, density frequently corresponds with a poor impression of the environment, as it fosters a closeness that makes intimacy difficult while imposing exterior restrictions. Thus, one of the main issues facing modern urban architecture is to suggest changes that would improve communal living conditions in collective housing, thereby lessening its poor reputation [1–3].

From different studies, it is apparent that residents would abandon the dream of the single-family home with a garden if collective housing were to provide expansions to the exterior, allowing beautiful attractive uses, for example, deep balconies, covered terraces, patios, indoor courtyards, and shared gardens [4]. In order to create more aesthetically pleasing communal living while satisfying demands for thermal, visual, olfactory, and acoustic comfort, optimal housing criteria need to be reassessed and new standards need to be introduced.

Related to the above, the European Union's member states are presently using the instruments required to address the issues with sound environments on an agglomerational scale while developing mapping strategies and related action plans. These tools have the privilege of using mapping and presenting regional quantitative criteria to ensure the reproducibility of studies, offering a common European standard for comparing different neighborhoods and cities. These strategic sound environment maps include action plans to reduce the sound levels of major noise sources and set limits in specific areas affected by severe nuisance. Therefore, despite the creation of tactical noise maps and the selection of a single quantifiable indication, these maps are unable to identify all the sound sources that are present around us in daily life [5].

In order to better meet the expectations of citizens, it is necessary to approach urban development projects by quantitatively measuring noise as well as by integrating the quality of the sound environment. For more than 20 years, from the first environmental labels to the more recent eco-districts (on which our research is based) [4,6], many operations have proposed new models of living in which thick facade devices (balconies, loggias, terraces, corridors) offer answers to this complex equation between density, intimacy, and sociability, as depicted in Figure 1.

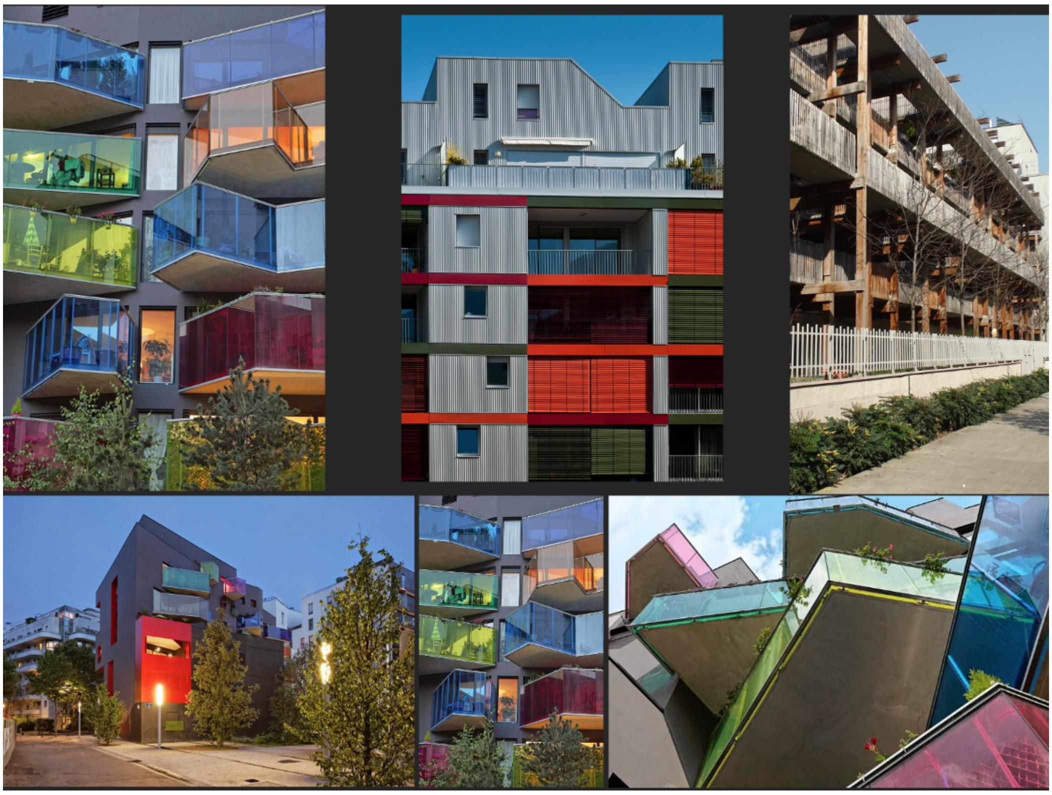

Figure 1. An example of eco-district collective housing extracted from [4].

The balconies of new constructions are places rich in sensory practices that provide users with this feeling of well-being. In the field of sensation, the perceptive dimension of sound is typically less directly worked on by architects as a means of design in comparison to the visual field [7].

This lack of concern for the sound environment in most architectural projects is certainly explained by its non-determinable and evolving character in space and time. In fact, sound travels in all directions through materials, which makes it much harder to channel and direct compared to light.

Indeed, as shown by different studies, in particular those of Cresson (https://www.esquissons.fr/, accessed on 5 May 2023) [8], the limits of sound are rarely identical to those of the visual; obviously, sound phenomena are not observable and rarely stop due to architectural techniques such as facades, balconies, railings, walls, and openings. It is a conceivable hypothesis that a dearth of control tools (prediction, simulation, etc.) contributes to the lack of value placed on sound perception in architectural and urban design.

1.1. Motivation and Contribution

The factors that influence sound perception on the balconies are the morphology of the balcony, the materials, the height of the balcony's location on the facade, and the surrounding urban configurations and use [9]. In this direction, the height of the floor where a balcony is located is an essential parameter that allows for different types of atmospheres, as it can affect the visual and sensory experience. For example, if a balcony is located on a high floor, the view can be expansive, providing a sense of openness and freedom. The height can make the balcony feel more secluded and private, as it is removed from street-level activity. Additionally, being at a greater height can create a sense of thrill or excitement, as one feels a sense of being suspended in the air. On the other hand, if a balcony is located on a lower floor, the view may be obstructed by buildings or trees, limiting the sense of openness and freedom. The balcony may feel more exposed to street-level activity, reducing the sense of privacy. Being at a lower height can create a more grounded feeling connected to the surrounding environment. In this sense, the height of a floor where a balcony is located can greatly impact the visual and sensory experience of being on the balcony, and can create different types of atmospheres.

In this work, our goal is to determine whether sound perception in the form of sound fragments can be used to automatically determine the floor on which a balcony is located. If this identification can be done efficiently, new components could be included to the design of balconies, as designers could better suggest the exposure, morphology, and arrangement tailored to balconies with regard to the impression of a sound environment connected with the height of its floor.

In this work, the problem of floor identification of a balcony is tackled by analyzing its sound environment as an audio classification problem using the associated soundtrack to characterize this environment. This task involves learning to classify sounds and predicting the class (in this case, the floor height) of a new sound.

Although machine learning has made significant advancements in the field of audio categorization recently, this issue remains open and challenging, as ambient everyday noises includes a great deal of information [10]. Recent research suggests using machine learning for sound and acoustic analysis [11–13]. To the best of our knowledge, no research has been done on the application of learning strategies for taking advantage of balcony soundscapes. In this paper, a novel machine learning-based approach is proposed to automatically classify recordings with soundtracks from balconies into the respective floors.

1.2. Machine Learning

Machine learning (ML) is a branch of artificial intelligence that systematically applies algorithms to synthesize the underlying relationships between data and information [14]. Machine learning approaches are traditionally divided into three main categories, which correspond to learning paradigms depending on the nature of the "signal" or "feedback" available to the learning model, as described below.

Supervised Learning [15] is a framework in which a Machine Learning algorithm is trained to predict outputs (or "labels") based on a given set of input data and their corresponding output labels. On the other hand, Unsupervised learning [16] is a framework in which an algorithm is trained to identify patterns or relationships within a given set of data, without any explicit guidance or labels provided. In between, there is the Semi-Supervised Learning [17] framework, which trains an algorithm using both labeled and unlabeled datasets. The goal is to use the structure of the data with the label information to find a predictor that performs better than one that uses only the labeled training data. Finally, Reinforcement Learning [18] involves an agent learning to make decisions in an environment in order to maximize a cumulative reward signal.

In our study, a set of sound fragments associated with stages of thirty intermediate spaces from the Esquis'sons! project was used. This project was studied in six sustainable neighborhoods in Europe (Germany, Spain, France, and Sweden) using the sound qualities of intermediate spaces located on the facade of buildings, such as balconies, loggias, terraces, and corridors (BLTC). Figure 2 shows images of these six-districts. This space has drawn a lot of attention, as it presents a new architectural language that has emerged in this type of neighborhood. A sonic approach was privileged in studying this architectural element for evaluating its potential in different urban and climatic contexts in Europe. These sites were chosen because they support outstanding and avant-garde BLTC architectural forms while emphasizing the morphological development of this typology of spaces. The different eco-districts were selected along a north–south axis in Europe that traverses various cultural, climatic, and urban environments in order to ensure a variation of urban form.

For this purpose, several sound recordings were made in situ on each balcony, each lasting 10 min, to better understand the differences in listening and the effect of the spatial layout, on both the architectural and urban scales, on the sound atmosphere. Many micropositions, including standing and sitting postures, in front of the railing, or close to the facade, were chosen in consultation with the inhabitants to reflect the natural and frequent usage of the balcony. Please refer to [19,20] for more details.

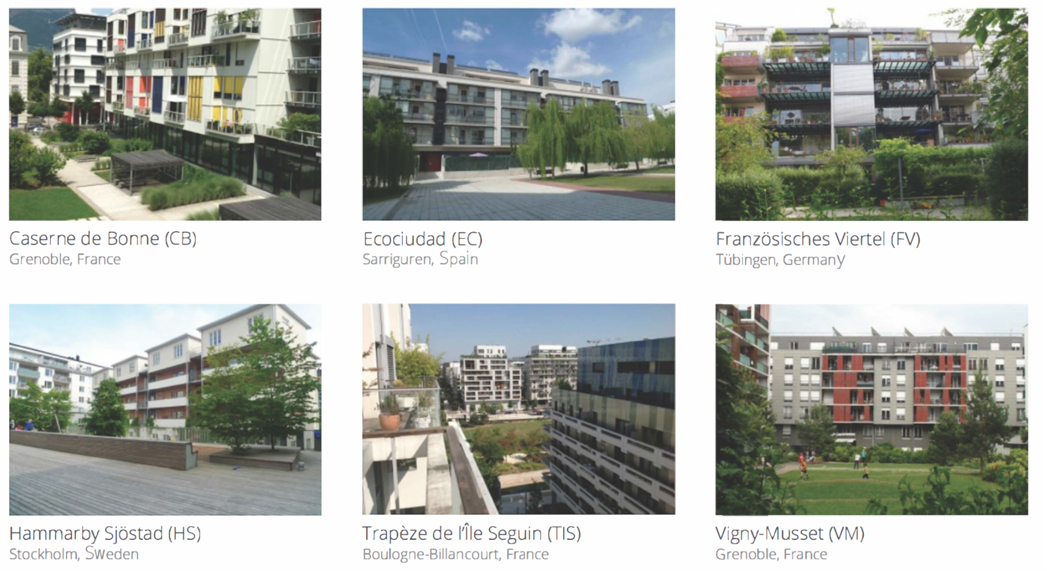

Figure 2. The six eco-districts in Europe considered in our study [20].

Our goal was to determine from the soundtracks of a subset of these buildings whether it is possible to predict the floors or height of the soundtracks in the rest of the buildings. Therefore, the framework of our study is that of supervised learning.

1.3. Outline

In the following, Section 2 presents the classical sound processing models as well as the convolutional neural network models used in this work. Section 3 presents our approach for classifying floors by their soundtracks. The data from the Esquis'sons! project are used to illustrate our findings in Section 4, and Section 5 concludes the work and identifies its next steps.

2. Soundtrack Representations

VanDerveer [21] offers the following criteria for defining a sound of the environment:

- It is the product of an event;
- It is the reflection of one or a series of causal events;
- It does not fall under speech recognition.

Thus, the sounds of an environment are categorized according to several categories: noise, natural sound, artificial sound, speech, music.

Recent techniques have been developed to process sound for a variety of purposes, including removing noise components from the signal.

In this way, one of the most popular techniques is Independent Component Analysis (ICA), which separates a multivariate signal into independent non-Gaussian components [22]. The basic idea behind ICA is to find a set of statistically independent source signals from a set of mixed signals. In the case of sound signals, the mixed signals may contain noise components that interfere with the desired signal. By applying ICA to the mixed signals, it is possible to separate the noise components from the desired signal, allowing for a cleaner and clearer output.

To apply ICA to sound signals, the first step is to acquire a set of mixed signals. This can be done by recording the desired sound along with any interfering noise. The mixed signals are then preprocessed to ensure that they are centered, scaled, and have a consistent sampling rate. Next, the mixed signals are passed through an ICA algorithm which separates them into independent components. The ICA algorithm works by maximizing the statistical independence of the components while minimizing their mutual information. After the independent components have been identified, the noise components can be removed from the desired signal by simply subtracting them from the mixed signal. The result is a cleaner and clearer sound signal that is free of noise interference.

One important consideration when applying ICA to sound signals is the choice of algorithm [23]. There are many different ICA algorithms available, each with its own strengths and weaknesses. It is important to choose an algorithm that is appropriate for the specific application and signal characteristics.

As the sound signals in our scenario are not standardized and originate from a range of neighborhood soundscapes (such as the street, the school, the inside of the building, etc.), it is impossible to pinpoint the ideal interfering noise for any situation. As a result, our aim is to develop a stand-alone method that can identify the sound signal with its incorporated noise.

In our work, the sound is considered as a one-dimensional vector with a large number of samples per second; in our case, the sampling frequency is 16 kHz, with a sample being an integer value. From this one-dimensional vector, it is possible to calculate different representations such as the MFCC and the Spectrogram, which translate different aspects of the sound fragment.

2.1. MFCC

Mel frequency cepstral coefficients (MFCCs) are a feature extraction technique used in speech recognition and other audio signal processing applications. They are based on a nonlinear frequency scale called the Mel scale, which is designed to mimic the way the human ear perceives sound [24]. The Mel scale is divided into a number of evenly spaced frequency bands, which are logarithmically spaced at lower frequencies and linearly spaced at higher frequencies. The Mel scale is typically used to map the frequency spectrum of an audio signal onto a set of Mel frequency bins, which can then be used to calculate MFCCs following a set of pre-processing steps from a windowing method known as the Hamming window.

2.2. Spectrogram

A classic representation of a spectogram is a three-dimensional graph, with the x-axis representsing time and the y-axis representing frequency; the intensity or color of each point on the graph provides a third dimension that indicates the amplitude of a certain frequency at a given time [25].

Spectrograms are often used in audio signal processing to analyze sounds such as music or speech and identify patterns or features within them. They can be used in other fields as well, such as in the analysis of seismic data, where they can help to identify earthquakes and other seismic events.

Figure 3 depicts MFCC coefficients and the spectrogram of a sound obtained with the modules of the `PyTorch` (https://pytorch.org/audio/stable/transforms.html, accessed on 5 May 2023) library on one of the sound tracks in our study.

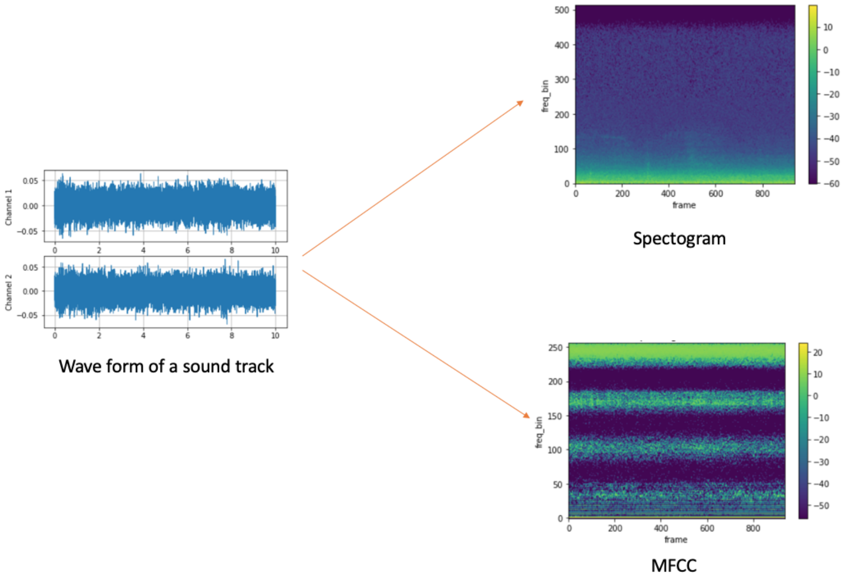

Figure 3. MFCC and Spectogram representations of a soundtrack considered in our work.

3. Proposed Approach

The MFCC and spectrogram transforms of a sound are represented by matrices of values which are considered as matrices of pixels, allowing the use of convolutional neural networks (CNNs) developed for image classification. In our case, CNNs were trained to find the association between these transforms and the floors of the balconies from which the corresponding sounds were recorded.

3.1. The CNN Model and Its Variants

Convolutional neural networks (CNNs) are inspired by the organization of the visual cortex. Inspired by the work of neuroscientists, ref. [26] proposed a neural network model called a neocognitron that has two basic types of layers (see Figure 4):

- Convolution layers
- Subsampling (or max-pooling) layers.

A convolutional layer contains units with receptive fields that cover part of the previous layer; the weight vector is often called a filter or kernel. Downsampling layers contain units with receptive fields that cover parts of previous convolutional layers. Such a unit generally calculates the average of the activations of the units of its set. This downsampling allows input forms to be correctly classified in visual scenes even when they are moved.

Based on this, in [27] the authors introduced the first convolutional network capable of successfully capturing spatial and temporal dependencies in an image through the application of learned kernels. This network architecture is better suited to the image dataset thanks to the reduction in the number of parameters involved and the reusability of the weights.

The three types of layers in a convolutional network are convolution, pooling (or pruning), and fully connected layers.

3.1.1. Convolution Layer

In the convolution layer, an input (usually a tensor) is modified by a kernel or filter that is learned. A convolution layer has three attributes:

- Convolution kernels, which are defined by width and height;
- The number of input channels and output channels (hyperparameters);
- The depth of the convolution filter (the input channels), which must be equal to the number of channels (depth) of the input map.

The example below shows the convolution of a 5 × 5 × 1 image with a depth of 1 using a 2 × 2 × 1 kernel of depth 1 in order to obtain 4 × 4 × 1 convolved features:

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}}_{\text{Image}} + \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}}_{\text{Core,}2\times 2} = \underbrace{\begin{bmatrix} 2 & 1 & 3 & 2 \\ 0 & 2 & 3 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 0 & 2 & 3 \end{bmatrix}}_{\text{Convolved image}}$$

In this example, the kernel shifts nine times, from top left to bottom right, each time performing a matrix multiplication operation between K and the part of the image to which it is applied. In the case of images with multiple channels, e.g., RGB, the kernel has the same depth as that of the input image. Matrix multiplication is performed between the kernel and image stacks, and all results are summed using bias to output a one-channel convolved image.

The purpose of the convolution operation is to extract high-level features. Conventionally, the first convolution layer is responsible for capturing low-level features such as edges, color, gradient orientation, etc. With additional layers, the architecture accommodates higher-level functionality. The convolution result is of two types: either the convolved feature is reduced in dimensionality relative to the input, which is the general case, or the dimensionality is increased or remains the same.

3.1.2. Pruning Layer

The pruning layer is responsible for reducing the spatial size of the convolved feature to decrease the computing power required to process the data through this dimensionality reduction. In addition, it is used to extract dominant features that are invariant in rotation and position. There are two types: maximum pooling (or max-pooling) and average pooling. The first returns the maximum value of the part of the image covered by the kernel. In contrast, an average kernel returns the average of all values of the part of the image covered by the kernel. For the example below, the result of max-pooling is as follows:

$$\underbrace{\begin{bmatrix} 2 & 1 & 3 & 2 \\ 0 & 2 & 3 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 0 & 2 & 3 \end{bmatrix}}_{\text{Convolved image}} \rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 3 \end{bmatrix}$$

3.1.3. Fully Connected Layer

Fully connected layers at the end of the network allow nonlinear combinations of the high-level features represented by the output of the convolutional layer to be learned. They are generally few in number, typically one to three. There should be a transition between the last grouping layer and the first fully connected layer. This is done by simply serializing the representations produced by this last grouping layer. Our baseline model, called CNN$_{782}$, was successfully proposed for the classification of images [28]. Its architecture is composed of two successive convolutions and $8 \times 8, 2 \times 2, 5 \times 5$, and 8×8 max-pooling kernels, followed by a fully connected network with an input representation vector of size 782. This model is shown in Figure 4.

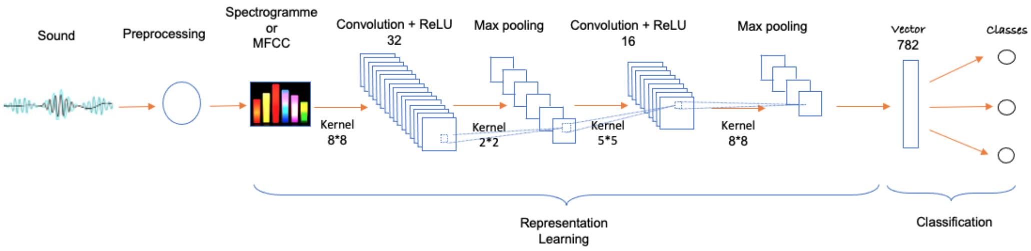

Figure 4. Diagram of the CNN_{782} convolutional network used in our experiments, with two successive convolution and max-pooling layers.

In order to see the impact of the filter size and the stride length used in the convolutional layers, two other variants of this model, named CNN_{392} (Figure 5) and CNN_{16384} (Figure 6), respectively, were considered. These models are defined as follows:

- CNN_{392}: Two successive convolutions and max-pooling with 8 × 8, 5 × 5, 5 × 5, and 4 × 4 kernels followed by a fully connected network with an input representation vector of size 392.

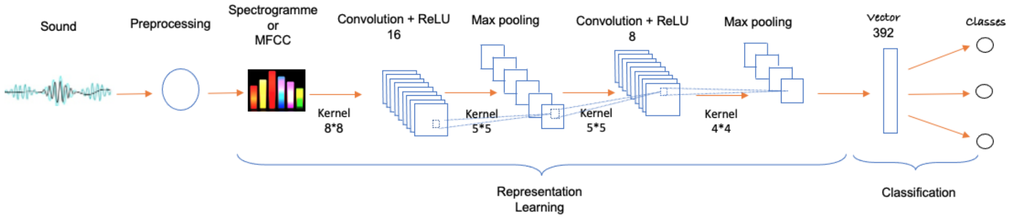

Figure 5. Diagram of the CNN_{392} convolutional network used in our experiments.

- CNN_{16384}: Three successive convolution and max-pooling layers, then a final convolution with seven 3 × 3 kernels, followed by a fully connected network with an input representation vector of size 16,384.

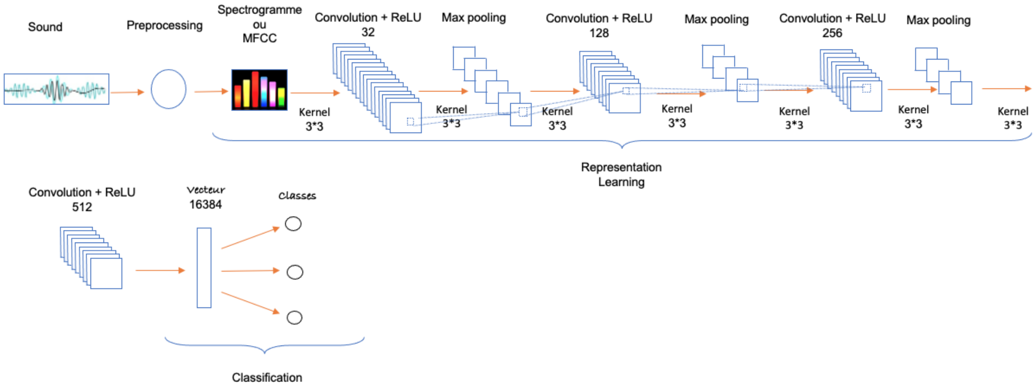

Figure 6. Diagram of the CNN_{16384} convolutional network used in our experiments.

In our experiments, the modules of the NN PyTorch library (https://pytorch.org/docs/stable/nn.html, accessed on 5 May 2023) were used to define these models. The sizes of the models based on their number of parameters are provided in Table 1.

3.2. Late Fusion

For the final prediction of our system, late fusion [29] was employed. This technique is mainly used in multi-view learning, and combines information from multiple views of a dataset in order to improve the accuracy of a machine learning model.

Table 1. The number of parameters of the different models.

Model	Params
CNN_{392}	29 k
CNN_{782}	227 k
CNN_{16384}	8 M

In multi-view learning, data are represented by multiple “views” or “modalities” that contain different information about the same underlying entities. For example, in image classification one view could be the raw pixels of the image, while another view could be a set of hand-crafted features extracted from the image.

Late fusion involves training separate models on each view of the data, then combining their predictions in a later stage. This can be done using a variety of techniques, such as averaging, voting, or weighted combinations.

The intuition behind late fusion is that each view contains complementary information about the underlying entities, meaning that combining them can lead to more accurate predictions.

In our work, the MFCC and the spectogram constitute the two modalities of a sound track, providing complementary information on the characteristics of the latter. Our purpose is to exploit this information by combining the predictions of two models, each operating on one of these modalities. Our approach consists of learning a convolutional model using the MFCC representation of the sounds and another convolutional model using the spectrogram representation of these sounds. After these two models are learned, the sounds in the training set are represented by a vector made up of the outputs of these two models. Each output is a class probability membership corresponding to a floor. Finally, a third model is learned in order to find the association between the predictions of these two initial models and the desired outputs associated with the sounds. This procedure is shown in Figure 7.

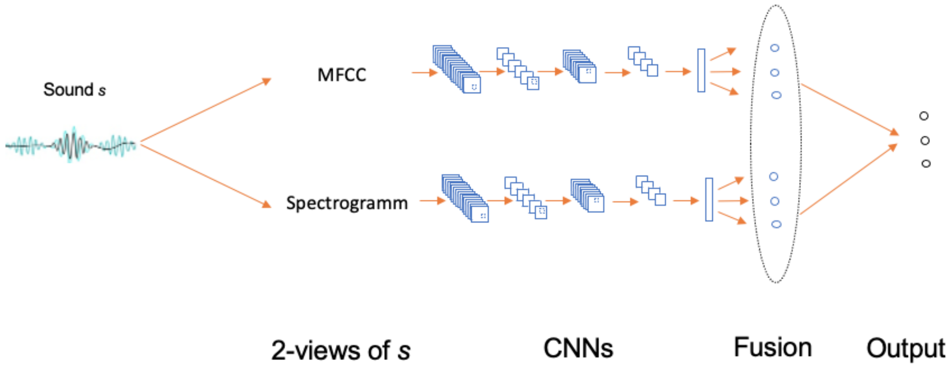

Figure 7. Our strategy of merging the decisions of each convolutional network using the MFCC or Sound Spectrogram representation.

It should be noted that other decision fusion techniques have been proposed recently to combine the decisions made by multiple classifiers. Among them, Alpha integration [30] has been applied to a variety of applications, for example, the classification of archaeological ceramics using ultrasound [31].

In comparison, late fusion involves independently making decisions with each classifier and then combining them at the end of the classification process, while Alpha integration involves fusing the decisions of the classifiers at an earlier stage by considering the degree of agreement between them. Alpha integration assigns a weight to each classifier based on its performance on the training data and the level of agreement among the classifiers. This weight is then used to combine the decisions of the classifiers, with greater weights assigned to classifiers that are more accurate and have higher agreement with the other classifiers.

Alpha integration can lead to better results than late fusion when the classifiers are highly correlated and have high agreement. However, it requires more computation and may be more complex to implement.

4. Experiments

This section presents the sound database used in our experiments as well as the results obtained with the three convolutional networks on the two modalities (MFCC and sound spectrogram) and with the proposed late fusion approach.

4.1. Data Collection

The sounds used in our research were sound recordings in MP3 format obtained by the Esquis'sons project teams. These sounds come from thirty recordings of thirty different intermediate spaces, and were classified as the 1st, 3rd, and 5th floors, designated respectively by R + 1, R + 3, and R + 5 in the following. Each sound recording had a duration of 2 min and was made on a specific floor of a building. In all, 27 records (90%) were considered when training the models and the rest were used for testing. The CNN models had many parameters (cf. Table 1), and using only 27 examples to learn these parameters would not lead to conclusive results. Thus, in order to increase the training samples, the sound recordings were cut every 5 s. Thus, the models learned to make associations between these cut-offs (or chunks) and the floors of their corresponding soundtracks. The matrices of MFCC and spectogram chunks provided in the input of the CNN models were 512 × 938 and 128 × 938, respectively. These chunks were obtained using the make_chunks function of the library pydub.utils.

For the test, the floors that were most predicted for the chunks of a band were used to determine the floor of the associated balcony. Table 2 presents the distribution of bands and chunks per floor for the collection that is considered.

Table 2. Distribution of soundtracks and chunks by floor.

		R + 1	**R + 3**	**R + 5**	**Total**
Soundtracks	Train	8	9	10	27
	Test	1	1	1	3
	Total	9	10	11	30
Chunks	Train	258	268	250	776
	Test	19	38	30	87
	Total	277	306	280	863

In all of our experiments, an NVIDIA GeForce RTX 3070 GPU with 5888 CUDA cores and 8 GB of GDDR6 memory was used. The learning rate was chosen by grid search over the set $\{10^{-3}, 10^{-2}, 10^{-1}\}$, the batch size was fixed to 64, and Adam was used as the optimization algorithm. The running times for training the convolutional neural networks were from 2 to 4 h. After the models were trained, the test times were in the range of a few milliseconds to a few hundred milliseconds per sound track.

4.2. Experimental Results

The training and testing of the models was repeated twenty times, and four performance measures were used for evaluation: Accuracy, Recall, Precision, and F_measurement, calculated on the basis of the confusion matrix $\mathbf{A} = (a_{ij})_{1\leq i,j\leq 3}$ of the predicted classes of the chunks of the test samples compared to their desired classes as follows.

Accuracy = $\frac{a_{11}+a_{22}+a_{33}}{\sum_{i=1}^{3}\sum_{j=1}^{3} a_{ij}}$; Recall = $\frac{1}{3}\sum_{k=1}^{3}\frac{a_{kk}}{\sum_{i=1}^{3} a_{ki}}$; Precision = $\frac{1}{3}\sum_{k=1}^{3}\frac{a_{kk}}{\sum_{i=1}^{3} a_{ik}}$; and F1 = $2\frac{\text{RecallPrecision}}{\text{Recall+Precision}}$ [32].

Table 3 presents the Accuracy of the three models learned using the MFCC and the spectrogram of the chunks separately on the training and the test sets. The best performance per modality is shown in bold; $^{\downarrow}$ indicates statistically significantly worse performance than the best result according to the Wilcoxon rank-sum test ($p < 0.01$) [33].

Table 3. Accuracy of different variants of CNNs with Spectogram and MFCC representations. The best performance per representation is shown in bold.

	Spectogram		**MFCC**	
Model	**Train**	**Test**	**Train**	**Test**
CNN_{392}	$0.561^{\downarrow} \pm 0.017$	$0.404^{\downarrow} \pm 0.014$	$0.566^{\downarrow} \pm 0.012$	$0.468^{\downarrow} \pm 0.011$
CNN_{782}	$\mathbf{0.643} \pm 0.019$	$\mathbf{0.569} \pm 0.013$	$0.638^{\downarrow} \pm 0.012$	$0.582^{\downarrow} \pm 0.014$
CNN_{16384}	$0.4423^{\downarrow} \pm 0.012$	$0.316^{\downarrow} \pm 0.015$	$\mathbf{0.769} \pm 0.011$	$\mathbf{0.67} \pm 0.013$

The CNN_{782} and CNN_{16384} models provide the best results with the Spectogram and MFCC representations, respectively. The CNN_{782} model is the second best model when considering the MFCC representation, while CNN_{16384} is the worst model with the spectogram representation. The table shows that the CNN_{782} model has the best compromise between performance and complexity compared to the two other models. Thus, this model was used in the late fusion approach; two distinct CNN_{782} models were trained on the MFCC and spectrogram representations, the outputs of these models were used to represent the sounds, and a third classifier was then trained to combine these predictions for a final classification. This third model used a random forest approach [34], which is one of the most efficient models for classification in cases where the data are represented by vectors. This model learns multiple random trees and combines their results based on majority vote [35]. Table 4 reports the average of the results over twenty training/test runs of this approach.

Table 4. Late fusion performance on CNN_{782} classifier results using MFCC and Spectrograms as two views of sounds.

Model	**Accuracy**	**Recall**	**Prec.**	**F1**
Random Forest	0.823	0.821	0.812	0.816

Table 5 presents the confusion matrix of one of these experiments. From these results, it is apparent that the true positives on each diagonal of the confusion matrix are greater than the false positives, meaning that the chunks of the different floors are 100% correctly predicted. This result is the same for the nineteen other experiments, which validates the effectiveness of this approach for automatically determining the floors of a building.

Table 5. The confusion matrix of the test set of one of our experiments.

		Predicted		
		R+1	R + 3	R + 5
Desired	R + 1	16	3	2
	R + 2	4	32	1
	R + 3	2	4	24

Furthermore, these results suggest that by using two views of the sound signal it is possible to combine the outputs of CNN_{782} in an efficient manner in order to precisely determine the height of the floors based on their related sound tracks. This paves the way for the use of more elaborate fusion techniques in situations where there are a higher number of floors and a more noisy environment.

5. Conclusions and Discussion

In usual urban environments with a constant background noise, it is likely that listening to a fragment is not necessarily easy to apprehend or use in guessing the height of the corresponding floor. For the Esquis′sons study, the sound engineer and a jury of listeners that were formed for the research had selected fragments between 1′30 and 2′ that represented the situation well.

In this work, the difficult task of automatically identifying the floor of a balcony based on a sound bounded at 2 min recorded on that balcony is considered.

In our experiments, it was found that sound chunks of 5 s can be identified with high accuracy, even though the number of training samples is not very much; furthermore, with the proposed late fusion strategy, the floors of these bands are 100% correctly classified. This result is consistent throughout all twenty experiments when training and test sets were chosen randomly, which demonstrates the viability of this method for automatically identifying balcony floors.

In general, these algorithms open the way to tools that can help designers to understand the consequences of design choices by intervention on the sound material itself, where other acoustic tools may remain too crude.

For future work, the present study can shed light on the development of software that can predict the class of new sound fragments. As a first step, it would be interesting to determine the extent to which this recognition is effective with sound fragments that are radically different (i.e., in event density, intensity, and frequency) from those that serve to train the model. This problem was tackled as a classification problem with a fully annotated training set. It would be interesting to consider other learning approaches, such as ranking [36], and to consider other types of training the model, for example using unlabeled samples in the training process [37].

Author Contributions: All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this research come from https://www.esquissons.fr/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Madanipour, A. Why are the Design and Development of Public Spaces Significant for Cities? *Environ. Plan. B Plan. Des.* **1999**, *26*, 879–891. [CrossRef]
2. Jacobs, J. *The Death and Life of Great American Cities*; Vintage: New York, NY, USA, 1993.
3. Gehl, J.; Mortensen, L. *Cities for People*; Island Press: Washington, DC, USA, 2010.
4. Pirhosseinloo, H. Habiter la Façade: La Conquête D'une Épaisseur Sensible: Les Dispositifs de Façade Épaisse dans les Logements Collectifs des Écoquartiers: Conception Architecturale et Ambiances. Ph.D. Thesis, Université Grenoble Alpes, Saint-Martin-d'Hères, France, 2019.
5. Accon; AECOM; Directorate-General for Environment (European Commission); The Centre for Strategy & Evaluation Services LLP (CSES). *Evaluation of Directive 2002/49/EC Relating to the Assessment and Management of Environmental Noise: Final Report*; Publications Office: Luxembourg, 2016. [CrossRef]
6. Cunha, A.D. Les écoquartiers, un laboratoire pour la ville durable: Entre modernisations écologiques et justice urbaine. In *Espaces et Sociétés*; Érès: Paris, France, 2011; pp. 193–200.
7. Spence, C. Senses of place: Architectural design for the multisensory mind. *Cogn. Res. Princ. Implic.* **2020**, *5*, 46. [CrossRef] [PubMed]
8. Rémy, N.; Pirhosseinloo, H.; Bardyn, J.L.; Chelkoff, G.; Said, N.G.; Marchal, T. *Esquis'sons! Outils d'aide à la Conception d'Environnements Sonores Durables*; Research Report 88, Cresson; ADEME, Direction Villes et territoires durables; ENSAG: Grenoble, France, 2016.
9. Lee, P.; Kim, Y.; Jeon, J.; Song, K. Effects of apartment building façade and balcony design on the reduction of exterior noise. *Build. Environ.* **2007**, *42*, 3517–3528. [CrossRef]
10. Sehili, M.E.A. Environmental Sounds Recognition in a Domotic Context. Ph.D. Thesis, Institut National des Télécommunications, Paris, France, 2013.
11. Font, F.; Serra, X. Urban sound tagging with attention-based recurrent neural networks. *Appl. Sci.* **2019**, *15*, 31–61.

12. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6. [CrossRef]
13. Ali, Y.H.; Rashid, R.A.; Hamid, S.Z.A. A machine learning for environmental noise classification in smart cities. *Inst. Adv. Eng. Sci.* **2022**, *25*, 1777–1786. [CrossRef]
14. Awad, M.; Khanna, R. Machine Learning. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, USA, 2015; pp. 1–18. [CrossRef]
15. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
16. Ghahramani, Z. Unsupervised Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, 2–14 February 2003, Tübingen, Germany, 4–16 August 2003, Revised Lectures*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 72–112.
17. Amini, M.R.; Usunier, N. *Learning with Partially Labeled and Interdependent Data*; Springer: Berlin/Heidelberg, Germany, 2015.
18. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
19. Pirhosseinloo, H.; Said, N.G. Towards a typology of listening situations: The balcony as a sonic interface in evolution. In Proceedings of the International Conference for Sustainable Design of the Built Environment (SDBE), London, UK, 20–21 December 2017; pp. 305–317.
20. Esquissons. Présentation d'EsquisSons! 2015–2022. Available online: https://www.esquissons.fr/outils/presentation-desquissons/ (accessed on 18 September 2022).
21. Vanderveer, N.J. Ecological Acoustics: Human Perception of Environmental Sounds. Ph.D. Thesis, Cornell University, Ithaca, NY, USA, 1979.
22. Comon, P. Independent component analysis, A new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
23. Lee, T.W. *Independent Component Analysis: Theory and Applications*; Kluwer Academic Publishers: Norwell, MA, USA, 1998.
24. Zheng, F.; Zhang, G.; Song, Z. Comparison of Different Implementations of MFCC. *J. Comput. Sci. Technol.* **2001**, *16*, 582–589. [CrossRef]
25. Flanagan, J. *Speech Analysis, Synthesis and Perception*; Springer: Berlin/Heidelberg, Germany, 1972.
26. Fukushima, K. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biol. Cybern.* **1980**, *36*, 193–202. [CrossRef] [PubMed]
27. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
28. Hasanpour, S.H.; Rouhani, M.; Fayyaz, M.; Sabokrou, M. Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures. *arXiv* **2016**. [CrossRef]
29. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6. [CrossRef]
30. Safont, G.; Salazar, A.; Vergara, L. Vector score alpha integration for classifier late fusion. *Pattern Recognit. Lett.* **2020**, *136*, 48–55. [CrossRef]
31. Salazar, A.; Safont, G.; Vergara, L.; Vidal, E. Pattern recognition techniques for provenance classification of archaeological ceramics using ultrasounds. *Pattern Recognit. Lett.* **2020**, *135*, 441–450. [CrossRef]
32. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
33. Wolfe, D.A. Nonparametrics: Statistical Methods Based on Ranks and Its Impact on the Field of Nonparametric Statistics. In *Selected Works of E. L. Lehmann*; Springer: Boston, MA, USA, 2012; pp. 1101–1110. [CrossRef]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2008.
35. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282. [CrossRef]
36. Usunier, N.; Amini, M.R.; Goutte, C. Multiview Semi-supervised Learning for Ranking Multilingual Documents. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases ECMLPKDD, Athenes, Greece, 5–9 September 2011; pp. 443–458.
37. Maximov, Y.; Amini, M.R.; Harchaoui, Z. Rademacher Complexity Bounds for a Penalized Multi-class Semi-supervised Algorithm. *J. Artif. Intell. Res.* **2018**, *61*, 761–786. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.