



Article

Aroma Release of Olfactory Displays Based on Audio-Visual Content

Safaa Alraddadi ^{1,*}, Fahad Alqurashi ¹, Georgios Tsaramiris ^{2,*} , Amany Al Luhaybi ³ and Seyed M. Buhari ² 

¹ Computer Science Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia; fahad@kau.edu.sa

² Information Technology Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia; mesbukary@kau.edu.sa

³ College of Computing in Al-Qunfudhah, Umm Al Qura University, Makkah 24382, Saudi Arabia; amluhaybi@uqu.edu.sa

* Correspondence: salraddadi0009@stu.kau.edu.sa (S.A.); gtsaramiris@kau.edu.sa (G.T.)

Received: 23 October 2019; Accepted: 12 November 2019; Published: 14 November 2019



Abstract: Variant approaches used to release scents in most recent olfactory displays rely on time for decision making. The applicability of such an approach is questionable in scenarios like video games or virtual reality applications, where the specific content is dynamic in nature and thus not known in advance. All of these are required to enhance the experience and involvement of the user while watching or participating virtually in 4D cinemas or fun parks, associated with short films. Recently, associating the release of scents to the visual content of the scenario has been studied. This research enhances one such work by considering the auditory content along with the visual content. Minecraft, a computer game, was used to collect the necessary dataset with 1200 audio segments. The Inception v3 model was used to classified the sound and image dataset. Further ground truth classification on this dataset resulted in four classes: grass, fire, thunder, and zombie. Higher accuracies of 91% and 94% were achieved using the transfer learning approach for the sound and image models, respectively.

Keywords: audio classification; olfactory display; deep learning; transfer learning; inception model

1. Introduction

The auditory and visual information of computer games is easy to obtain through software means by capturing screenshots and recording audio. It is as easy to recognize the events and the characters in the game as it is to hear the character and the soundtrack in the game. However, the olfactory information related to the game cannot be obtained through the media, either television or any other device, due to the challenges of comparing it digitally with visual and auditory information [1,2].

Olfactory displays have recently been used with virtual reality applications where it imitates reality and allows user interaction with an imaginative world by specific interaction devices [3]. However, the association between virtual content and scents is application specific and cannot be used in other applications. Studies have shown that the information obtained through the sense of smell is lesser than that obtained through the senses of hearing and sight [4]. At the same time, the olfactory information enhances the senses and immersion in reality more than the other senses. Nonetheless, the sense of smell is still the least used to enrich user experience in the virtual world. The literature review covers many studies that have developed olfactory displays that release scents based on a specific time.

Most of the current approaches either have no direct association with the virtual content (releasing scents based on preset timers) or are specific to an application. This makes them inappropriate for gaming and virtual reality applications as it is not possible to predict the user's actions and release the appropriate scents. Recent research [5] associated virtual artifacts with scents, thus allowing olfactory

displays to be used in highly dynamic applications. The work presented in this article builds on [5] by enabling the release of scents based on visual and audio information.

The proposed system uses image recognition classified from [5] and pairs it (Logical OR operator) with a new audio classifier. Transfer learning with Inception v3, which takes the log-Mel spectrogram of a short audio sample as input, is used to recognize the sound. While it is easy for humans to associate sounds with a specific scenario [6], it is challenging for machines, as it requires a significant amount of audio data and can easily be disturbed by undesired noise. In this research, noise was considered as unclassified sounds played at the same time as classified (labeled) sounds, and has a direct negative impact on the accuracy of the recognition.

This study contributes to the areas of gaming and virtual reality as it adds the option of scents to be released based on audio as well as recognized images. This is an important addition as sometimes, some virtual elements are auditory, but with limited or no visual information. For example, it might be raining in the game, but the user cannot see it as it is outside their field of view or due to low lighting conditions. As long as the user can hear the rain, the scent will still be released.

The rest of this paper is organized as follows. Section 2 reviews the related work of olfactory displays and sound recognition techniques. Section 3 describes the methodology of the proposed system. Section 4 presents the data analysis and discusses the experimental results. Finally, the study concludes in Section 5.

2. Literature Review

The literature review is divided into two sections. The first section discusses how recent studies have used convolutional neural networks (CNNs) for sound recognition and justifies the use of CNN in the current research. The second section presents the latest developments in olfactory displays.

2.1. Sound Recognition

In recent years, studies have shown that the CNN model outperforms traditional methods in different taxonomic tasks including sound recognition. For sound recognition, the most common auditory features such as raw waveform, log-Mel spectrogram, or Mel frequency cepstral coefficient (MFCC) are used to train the deep CNN.

A novel end-to-end system to classify raw sound with two conventional layers was proposed in [7]. The experimental results showed that the combination of the proposed model and log-Mel-CNN exceeded the state-of-the-art log-Mel-CNN model with 6.5% improvement in the classification accuracy. However, the model is inappropriate to learn the complex structure of audio due to the presence of only two conventional layers.

Transfer approach called SoundNet used to transfer knowledge from visual recognition network was presented in [8]. The aim was to train a CNN that classified raw audio waveforms from unlabeled videos. The experimental result showed that SoundNet achieved an acoustic classification accuracy of 97%. However, if the CNN is trained on a large scale dataset (around two million samples), it can achieve a similar accuracy.

A very deep conventional network with 34 weight layers that processes the raw audio waveform directly was proposed in [9]. The model applied batch normalization on each output layer while residual learning skipped some fully connected layers and down sampling accurately in the initial layer. All of these contributed to avoiding difficulty in the trained model as well as providing low computational cost. The result showed that the CNN deep architecture outperformed CNN with the log-Mel spectrogram with a 71.8% accuracy.

Another study proposed CNN architecture with three conventional layers to classify sound signals using the log-Mel spectrogram as features to learn the model [10]. Furthermore, different types of audio data augmentation techniques such as time stretching (fast or slow audio), pitch shifting (higher or lower pitch of audio), dynamic range compression (compresses audio sample), and background noise (mix sample sounds with another sound that contains background from different acoustics) were

used to overcome the problem of a lack of data. However, the performance improved in terms of the classification accuracy only in some types of augmentation, while it remained non-progressive in others. This CNN architecture classified short audio by using a log-Mel spectrogram with the same features as that used in [6]. Moreover, the training procedure with two phases non-fully trained and fully trained, improved the accuracy by reaching 86.2% as well as outperformed the accuracy of the Gaussian mixture modeling-Mel frequency cepstral coefficient (GMM-MFCC) by 6.4%.

A fully connected CNN model for partly labeled audio based on a trained large scale dataset (audio set) using the log-Mel spectrogram as input was introduced in [11]. Moreover, a CNN model was used as the framework to transfer and learn audio representation (spectrogram) using different methods where the accuracy of the proposed model reached up to 85%.

In order to overcome the difficulty of distinguishing sounds that come from various sources as well as the missing labels of these sounds, the authors in [12] proposed a deep CNN called AENet. The model processes large temporal input with the data augmentation technique called equalized mixture data augmentation (EMDA), which mixes sounds that belong to the same class and modified frequency of the audio sample by boosting and attenuating in a particular band. Moreover, it applied transfer of learning to extract audio features from AENet and combine them with visual features. The authors claimed that combining AENet features with visual features significantly improved its performance than that by combining MFCC with visual features.

A small number of systems have used spatial features extracted from binaural recordings. In order to obtain the advantages from feature engineering approaches (i-vector) and feature learning methods (CNN), the authors in [13] proposed a multichannel i-vector by computing MFCC for both channels in the audio sample. In addition, they built a CNN model similar to VGG-net (invented by the Visual Geometry Group) architecture that takes spectrogram features as the input. Moreover, combining two models was performed using the score vision technique, which creates the probability scores of each method and then fuses these scores. The performance of this hybrid approach achieved state-of-the-art and obtained first rank in the DCASE-2016 (Detection and Classification of Acoustic Scenes and Events 2016) challenge. However, this approach requires a large set of trainable parameters, which is not possible with our small dataset.

The authors in [14] proposed a CNN that consisted of eight convolutional layers and two fully connected layers using two spectrogram representations, the log-Mel spectrogram and gammatone spectrogram, as input. Traditional data augmentation methods were used to generate a new audio sample such as time stretch and pitch shift, in addition to applying the Mixup method on the training data by mixing two samples randomly selected within or without the same class. It was claimed that Mixup improved performance by 1.5% on the ESC-10 [15] dataset, 2.4% on the ESC-50 [15] dataset, and 2.6% on the UrbanSound8k dataset [16].

Most CNN models need a huge dataset in order to recognize the sound correctly. This makes them difficult to apply on limited datasets. Therefore, we will apply the transfer learning method to recognize sound samples in this research.

2.2. Olfactory Displays

Olfactory displays are devices designed to release scents into the environment. They are classified into two types: “wearable”, which are placed either on-body or on-head, and “environmental”, which are placed in the physical environment [17].

A wearable and fashionable olfactory necklace called Essence was designed in [18]. The Essence is able to release scents automatically based on data from the virtual context such as the location and current time of the users as well as on physiological data such as brain activity and heart rate. Moreover, the necklace can be activated manually, and the intensity of scents can be controlled through the stretch necklace thread. The results of the user experience show that the device is small enough and comfortable to be worn in most daily life activities. However, the device was unable to release multilabel scents at a time, and released one scent for one case based on the chosen user.

A smelling screen is an olfactory display embedded in a Liquid-Crystal Display (LCD) screen to generate and distribute odor along the screen based on the image shown [19]. The proposed device consists of four fans located at the corners of the screen to generate airflow that collides multiple times. Then, the airflow blows toward the user through tubing from the airflow collision point, which is considered as the odor source. However, the time between releasing the scents and its recognition by the user is not synchronized due to the delay of the scent reaching the olfactory organ.

An olfactory display named inScent [20] can be worn as a necklace that enables the user to receive scented notifications. The device was built to hold eight aromas where the scent is exchangeable through a small cartridge. Scents are triggered either manually by the users, or remotely by the instructor via an Android application based on the correlated scenario like scent that reflects the emotional link of the message sender and generates scents by using heating and a small fan to blow airflow toward the user. However, heating a wearable device may cause discomfort to the wearer.

Another study [21] proposed a thermal/heating approach to distribute scents from generation of the olfactory aromas. The device unit was built to hold eight aroma dispensers; each one containing a capillary tube, speed control fan, gas sensor to measure the release rate, and temperature to control the heating elements. The user controls the intensity of the aroma and fan speed as well as selects the aroma to be released by a software application. However, the heating approach might destroy the chemical components, which may limit the range of odors.

The authors in [5] presented a placed-in environment olfactory display that released six scents based on the visible content displayed by using an Inception v3 model for image recognition. However, visual elements were only associated with scents.

Overall, wearable devices can cause discomfort to users, thus hindering immersion into the virtual world. In contrast, environmental olfactory displays do not share this issue, but tend to have synchronization issues.

3. Methodology

3.1. System Overview

The proposed system consists of a Windows application that records the sounds and transforms raw sound into a log-Mel spectrogram while simultaneously taking screenshots from a game called Minecraft [22]. Conceptually, the proposed approach can be applied to any application as long as the classifiers have been trained to associate scents with its visual and auditory virtual phenomenon. The approach was applied on Minecraft as a proof of concept. The image classifier [5] and the sound classifier operate separately and identify which scents are to be released. Their results are then merged (union) and passed to the application that will inform the olfactory display. Only classes with an accuracy of 90% or more will be released. The olfactory display used in [5] was also used in this research. It is worth noting that this research focused on adding the capability of releasing scents based on an audio-visual virtual phenomenon and not on the development of an olfactory display. Figure 1 illustrates the system overview.

3.2. Dataset

The dataset consisted of 1200 audio segments that were distributed equally between four classes: grass, fire, thunder, and zombie. We selected these sounds based on the sound popularity in the Minecraft game as well as the availability of scents. The duration of all audio samples was four seconds with a 44.1 kHz sampling frequency and single audio channel (mono). Due to the similarity of the sound and to avoid overfitting, two deformation methods were used directly on the segments to generate a new sample. First, time stretching (TS) was applied to fast and slow audio samples using the Librosa function [23] (`librosa.effects.time_stretch`). In order to change the stretch, we used two speed factors of 0.5 and 2. Second, pitch shifting (PS) was applied to the high and low pitch of samples through the use of the Librosa function [23] (`librosa.effects.pitch_shift`) to change the pitch randomly.

All audio segments were then converted into a log-Mel spectrogram and used as input representation to train the network. We extracted log-Mel features from raw wave sounds by applying a short-time Fourier transform (STFT) over 40 ms windows with 50% overlap and Hamming windowing. We took the absolute value of each bin to square it and applied a 60-band Mel-scale filter bank. Finally, we computed the logarithmic conversion of the Mel energies using the Librosa library [23]. The log-Mel spectrogram was used to train the network without the need to combine features. Figure 2 illustrates a sample of the log-Mel spectrogram.

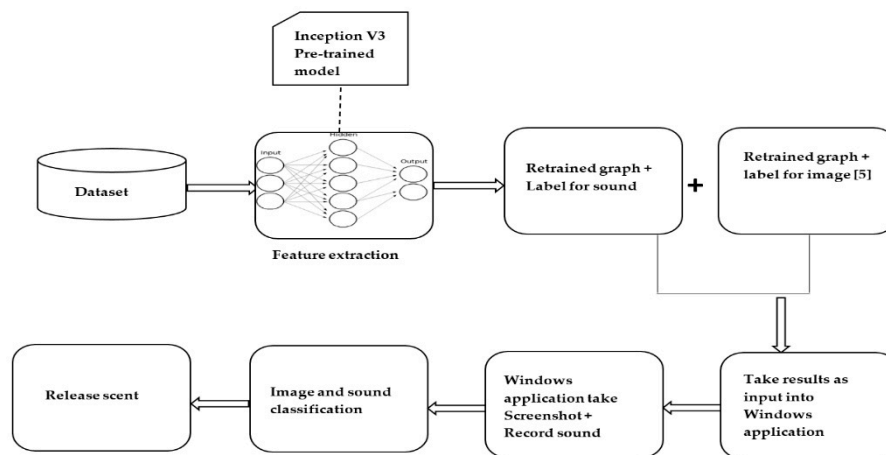


Figure 1. System overview.

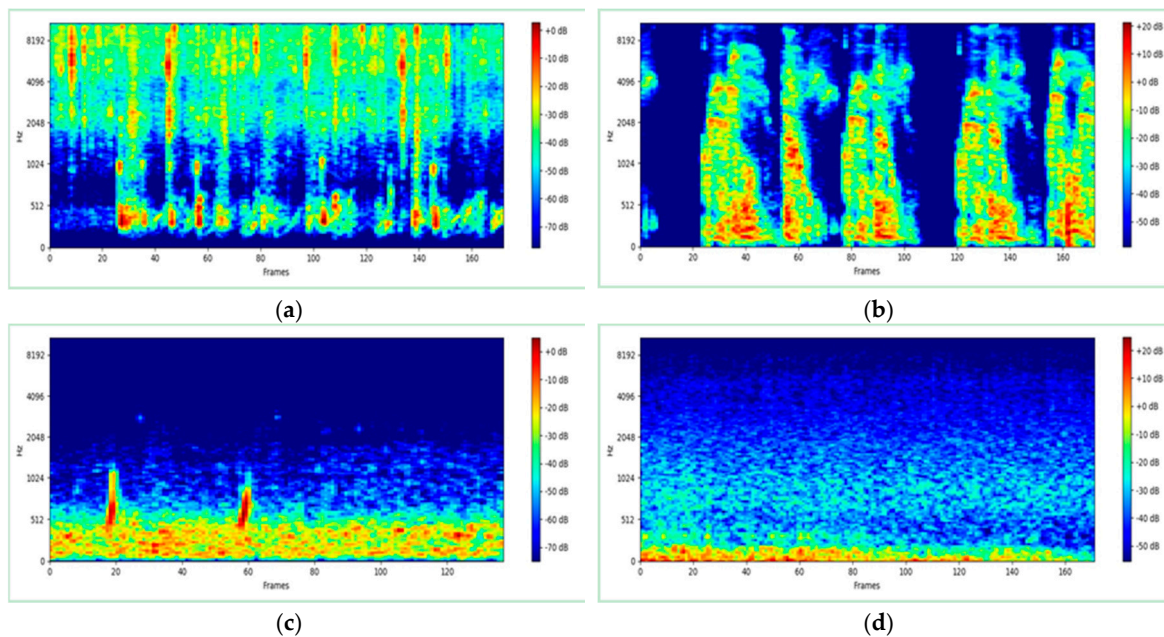


Figure 2. Sample of log-Mel spectrogram to train the model. (a) log-Mel of fire; (b) log-Mel of zombie; (c) log-Mel of ocean; (d) log-Mel of thunder.

3.3. Transfer Learning

Training the deep convolutional network from initialization requires a huge dataset to learn discernable features. The limited availability of data makes automatic image recognition impossible. In such cases, transfer learning makes CNN able to recognize images successfully by transferring knowledge from a model trained on a huge dataset into the target model, which is used for the new task.

Recently, many CNN architectures with a deep layer have been developed. In this research, Inception v3 [24] was adopted as a pre-trained model because of its ability to reduce computational complexity by using different sizes of convolutional filters (e.g., 1×1 , 3×3 , and 5×5) in the same layer and then sending them to the next layer to detect new features. In [25,26], one of these filters had to be chosen to be used first, followed by the max pooling layer, then this operation was repeated with the hope of detecting new features. However, this operation is computationally intensive due to the many operations that occur in each neuron. Despite the complexity of the architecture in Inception v3, it achieved extraordinary performance in terms of accuracy. Inception v3 was trained on an ImageNet dataset [27] that contained 1.2 million images with more than 1000 labels. Inception v3 extracted the features of ImageNet by using a CNN with fully connected layers and a SoftMax layer to classify images based on the ImageNet labels. The transfer learning used all convolutional layers and pooling layers in Inception v3 to extract the input features of the log-Mel spectrogram. Then, it removed the top layer (SoftMax) that classified the original dataset and trained the new layer with our task. Finally, the new model classified the images based on the labels of the new dataset. The process of transfer learning is illustrated in Figure 3.

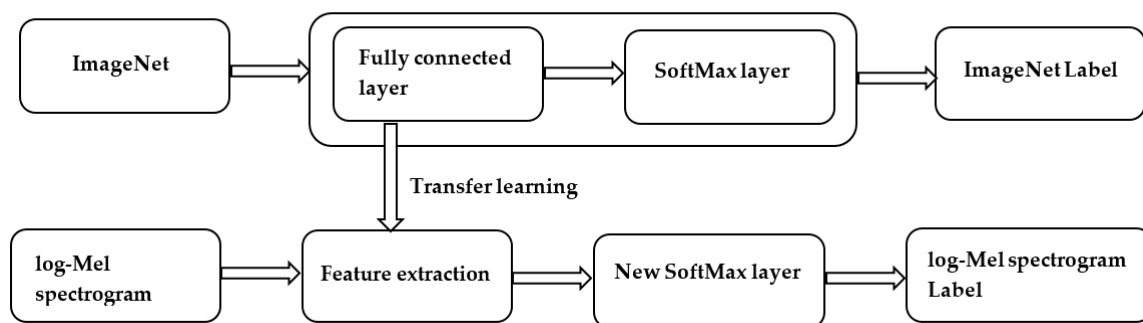


Figure 3. Transfer learning.

4. System Evaluation and Results

The work was evaluated in two stages. First, the capability of the model to identify the various audio classes was tested using a separate dataset of log-Mel spectrograms. Second, the integrated approach, consisting of both the image (re-used from [5]) and the proposed audio classifiers were tested for their consistency. The model was retrained using TensorFlow [28] on an Intel core i7-4720HQ processor with 16.0 GB memory. The dataset was trained with the slandered learning rate of 0.01, the iteration was set as 20,000, and the batch size of each iteration was equal to 100.

The retrained model was evaluated with 30 log-Mel spectrograms for each class. The classification result was obtained from the confusion matrix, as shown in Figure 4. As we can infer, the ocean was the least accurately recognized class by the system. The reason behind this is that the ocean in the game contains other creatures and their sounds overwhelm the sound of the ocean. On the other hand, the model predicts thunder sounds successfully because the sound of thunder is very loud and clear.

The model was evaluated before integrating the application by computing the accuracy, precision, recall, and f1 score for each category from a confusion matrix with 0.5 as the threshold value by using Equations (1)–(4). The prediction of the lowest value was not accepted because the application cannot release the aroma if the prediction is lower than 90%. The performance measurements in Table 1 were computed by using the following equations:

$$Accuracy = \frac{True\ Positives + False\ Negatives}{Total\ Number\ of\ Samples} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

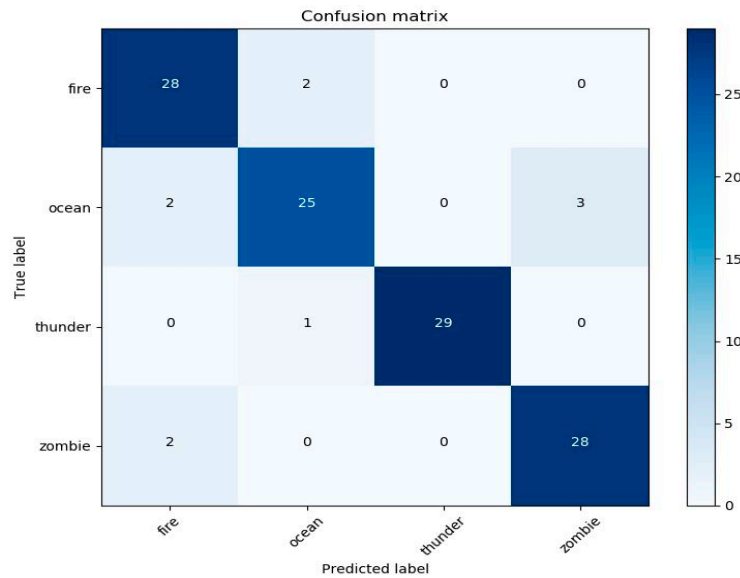


Figure 4. Confusion matrix for the retrained model evaluated based on the log-Mel spectrograms for fire, ocean, thunder, and zombie.

Table 1. Performance Measure for retrained Sound Model before Integrated into Application.

Category	Accuracy	Precision	Recall	F1 score
Fire	0.95	0.8	0.9	0.8
Ocean	0.93	0.8	0.8	0.8
Thunder	0.99	1	0.9	0.9
Zombie	0.95	0.9	0.9	0.9
Average	0.95	0.8	0.8	0.8

As can be seen from Table 1, the accuracy of the thunder outperformed other categories at 99% because it has a high sound that overshadows any other sounds around it, while ocean had less accuracy among the other categories with 93%. Overall, the average accuracy of the model was 95%; this was satisfied in this application due to the limited sounds in the game. The other statistical measures of precision, recall, and F1 score reached 0.9 in most cases, which was satisfied.

4.1. Performance Sound and Image Classifier in the Application

Evaluation performance of the integrated sound classifier with the Windows application was conducted, with audio samples recorded every ten seconds and converted into log-Mel spectrograms. At the same time (10 s), the application took screenshots and passed them to the image classifier. We compared the two classifiers to measure accuracy, precision, recall, and F1 score for the fire, zombie, and ocean categories using a confusion matrix with 0.5 as the threshold value. The predications that scored less than the threshold value were rejected. The following Figure 5 shows a sample of the accuracy of both classifiers within the application at the same time.

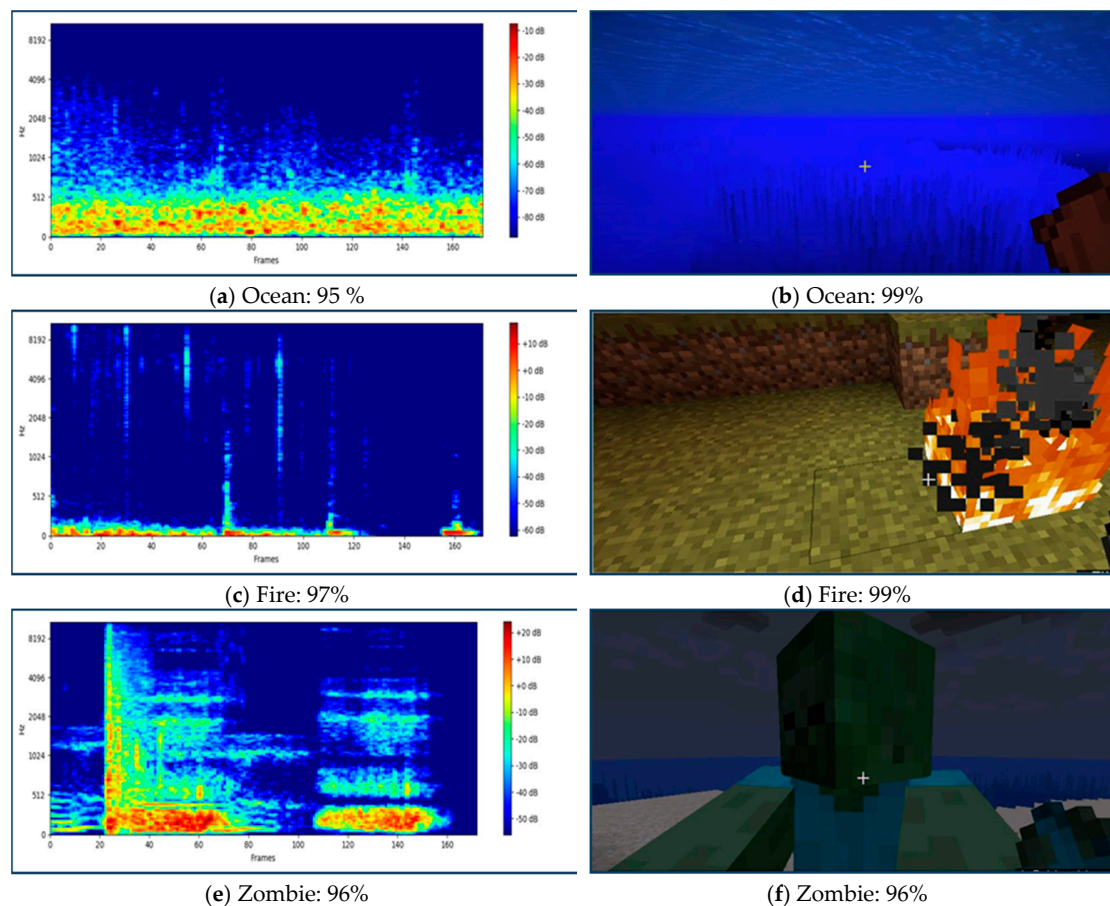


Figure 5. The accuracy of the sound and image classifier inside the application at the same time. (a) The accuracy of the ocean sound; (b) The accuracy of the ocean image; (c) The accuracy of the fire sound; (d) The accuracy of the fire image; (e) The accuracy of the zombie sound; (f) The accuracy of the zombie image.

The confusion matrixes for sound and image classifier performance within the application are shown in Figures 6 and 7, respectively. As can be seen, the ocean was the most misclassified category because the ocean in Minecraft contains other audio sources such as zombies and other creatures, which overlap the sound of the ocean. Additionally, two sounds from fire were classified as the ocean because the lava sound (a type of fire in the game) is similar to the sound of the ocean. In contrast, the ocean in the image classifier were classified correctly. Zombie was the most misclassified class in the image classifier with five images in the fire class because in the game, the zombie burns if exposed to the sun. However, in the sound classifier, zombie was classified with all classes correctly except fire, which was misclassified with two images due to the overlap of fire sounds with zombie sounds when the zombie was burning.

The accuracy performance for both classifiers are illustrated in Tables 2 and 3. It can be seen that the accuracy of the sound classifier decreased after being integrated into the application. It is believed this occurs in circumstances when players move very fast from one scene to another, which makes the sounds overlap and become difficult to recognize. Overall, the average accuracy of the audio classifier (91%) was less than the accuracy of the image classifier (94%). This is because, unlike images, the game produces multiple sounds at the same time (e.g., the sound of the ocean and a pack of zombies), which cannot be predicted. Nevertheless, in some cases such as fire and zombie, the accuracy outperformed fire and zombie in the image classifier. Thus, the smells are released based on the classifier that represents the highest accuracy. Furthermore, the recall result of the image classifier was 0.9, which outperformed the result of the sound classifier. Finally, the average results of precision

and F1 score were 0.8 for both classifiers, which were satisfactory in this application. It is worth noting that the two classifiers are complementary and do not complete each other. Additionally, the audio classifier could identify additional virtual phenomenon (e.g., thunder), even if it is not in the field of view of the player.

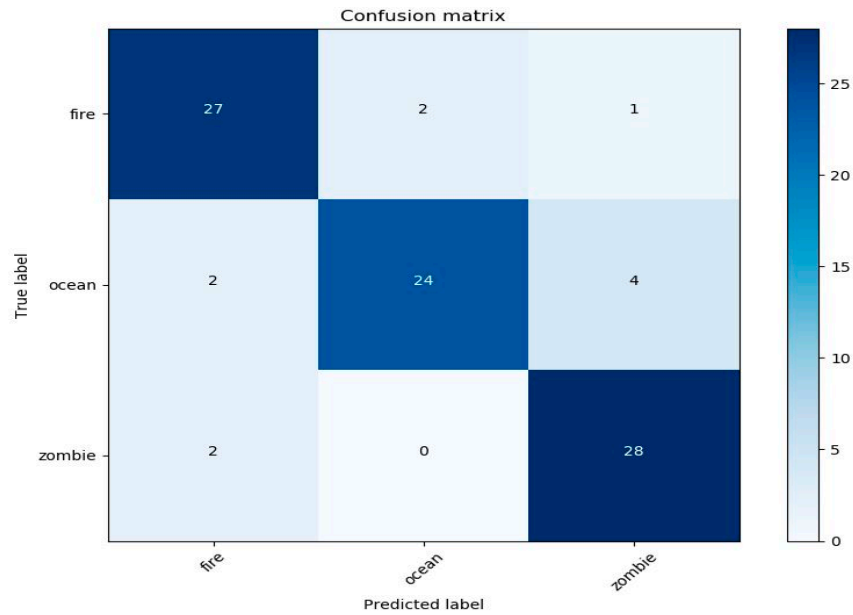


Figure 6. Confusion matrix for the evaluated sound within the application based on the log-Mel spectrogram for fire, ocean, and zombie.

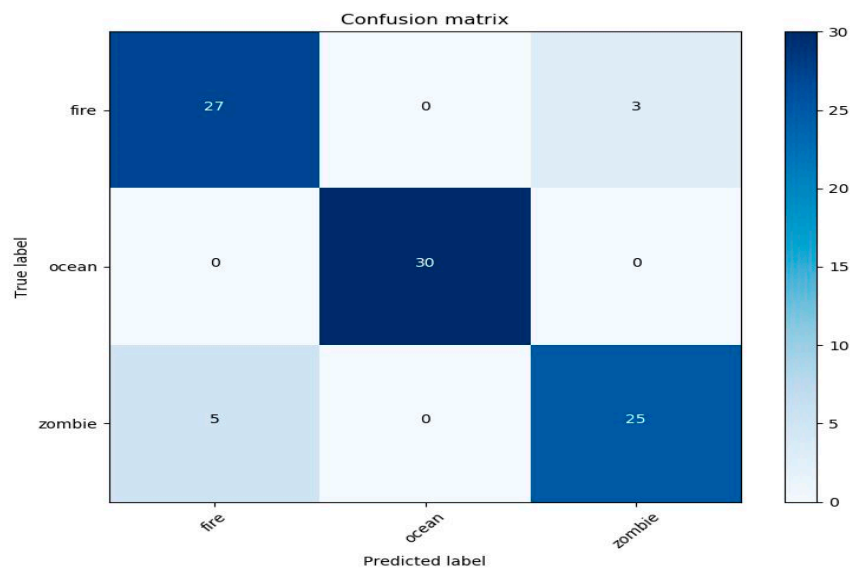


Figure 7. Confusion matrix for the evaluated image within the application based on the log-Mel spectrogram for fire, ocean, and zombie.

Table 2. Performance Measure for Sound Model within Application.

Category	Accuracy	Precision	Recall	F1 score
Fire	0.92	0.8	0.9	0.8
Ocean	0.91	0.9	0.8	0.8
Zombie	0.92	0.8	0.9	0.8
Average	0.91	0.8	0.8	0.8

Table 3. Performance Measure for Image Model within Application.

Category	Accuracy	Precision	Recall	F1 score
Fire	0.91	0.8	0.9	0.8
Ocean	1	1	1	1
Zombie	0.91	0.8	0.8	0.8
Average	0.94	0.8	0.9	0.8

4.2. User Experience

In order to test the impact on the user's experience, we conducted an experiment with five participants. The device was placed under the monitor, at the front of the users. Initially, the device was set to release new scents every three seconds. The synchronization between the game event and the release was acceptable, however, the scent persisted in the air for far longer. However, even after six minutes of game play (average), users could still differentiate the released aromas. Thus, they reported that the atmosphere was uncomfortable. In order to improve the user experience, we modified the release code to prevent an aroma being released more than once per minute. This improved the user experience, but still lacked a way to clean the previously released scent, which was proven to be a major drawback as the new aromas mixed with the old ones. Preventing the release of a new scent for 10 s (used in this research) resulted in a better overall user experience, but at the cost of a lot of missed releases, revealing a trade-off. While out of scope of this research, it is the belief of the authors that a new algorithm to decide when to release a new scent, based on the last release as well as the different persistence rates of various aromas, will have a positive impact on the user experience.

5. Conclusions

This study proposed an approach that combined audio and visual contents to automatically trigger scents through an olfactory device using deep learning techniques. The log-Mel spectrogram sound identification model was built based on a pre-trained Inception v3 model. Moreover, a Windows application was designed to record audio and convert it to a log-Mel spectrogram as well as take a screenshot of the same scene at the same time. In addition, the application controls the release of scents that are identified based on the highest accuracy. The accuracy of the integrated sound model with the application reached 91%, however, the accuracy was lower due to various sound recording situations. For example, sounds may overlap and become difficult to recognize. While the accuracy of the image outperformed that of the sound, sometimes it was misclassified. The sound and image models complement each other: in case one misrecognizes the scene, the higher accuracy will prevail, or the absence of either of them from a scene. The proposed approach can be applied to different virtual environments as long as scents can be associated with visual and auditory content. Further work is required to associate scents automatically with more sounds and images. Additionally, the approach can be tested with other games or virtual reality applications.

Author Contributions: Methodology, S.A., G.T., and A.A.L.; software S.A., A.A.L.; validation, S.A., and G.T.; formal analysis, S.A.; investigation, S.A., and G.T.; data curation, S.A., and A.A.L.; writing—original draft preparation, S.A.; writing—review and editing, S.A., G.T., and S.M.B.; supervision, F.A.

Funding: This research received no external funding.

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

References

1. Hashimoto, K.; Nakamoto, T. Tiny Olfactory Display Using Surface Acoustic Wave Device and Micropumps for Wearable Applications. *IEEE Sens. J.* **2016**, *16*, 4974–4980. [[CrossRef](#)]
2. Hashimoto, K.; Nakamoto, T. Stabilization of SAW atomizer for a wearable olfactory display. In Proceedings of the 2015 IEEE International Ultrasonics Symposium (IUS), Taiwan, China, 21–24 October 2015; pp. 1–4.
3. Steuer, J. Defining virtual reality: Dimensions determining telepresence. *J. Commun.* **1992**, *42*, 73. [[CrossRef](#)]

4. Kadowaki, A.; Noguchi, D.; Sugimoto, S.; Bannai, Y.; Okada, K. Development of a High-Performance Olfactory Display and Measurement of Olfactory Characteristics for Pulse Ejections. In Proceedings of the 2010 10th IEEE/IPSJ International Symposium on Applications and the Internet, Seoul, Korea, 19–23 July 2010; pp. 1–6.
5. Al Luhaybi, A.; Alqurashi, F.; Tsaramirsis, G.; Buhari, S.M. Automatic Association of Scents Based on Visual Content. *Appl. Sci.* **2019**, *9*, 1697. [[CrossRef](#)]
6. Valenti, M.; Squartini, S.; Diment, A.; Parascandolo, G.; Virtanen, T. A convolutional neural network approach for acoustic scene classification. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 1547–1554.
7. Tokozume, Y.; Harada, T. Learning environmental sounds with end-to-end convolutional neural network. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2721–2725.
8. Aytar, Y.; Vondrick, C.; Torralba, A. Soundnet: Learning sound representations from unlabeled video. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 892–900.
9. Dai, W.; Dai, C.; Qu, S.; Li, J.; Das, S. Very deep convolutional neural networks for raw waveforms. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 421–425.
10. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [[CrossRef](#)]
11. Kumar, A.; Khadkevich, M.; Fügen, C. Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 326–330.
12. Takahashi, N.; Gygli, M.; Gool, L.V. AENet: Learning Deep Audio Features for Video Analysis. *IEEE Trans. Multimed.* **2018**, *20*, 513–524. [[CrossRef](#)]
13. Eghbal-zadeh, H.; Lehner, B.; Dorfer, M.; Widmer, G. A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Nairobi, Kenya, 28 August–2 September 2017; pp. 2749–2753.
14. Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. In Proceedings of the Pattern Recognition and Computer Vision, Guangzhou, China, 23–26 November 2018; pp. 356–367.
15. ESC: Dataset for Environmental Sound Classification. Available online: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YDEPUT> (accessed on 9 November 2019).
16. UrbanSound8k. Available online: <https://urbansounddataset.weebly.com/urbansound8k.html> (accessed on 9 November 2019).
17. Murray, N.; Lee, B.; Qiao, Y.; Muntean, G.-M. Olfaction-Enhanced Multimedia: A Survey of Application Domains, Displays, and Research Challenges. *ACM Comput. Surv.* **2016**, *48*, 1–34. [[CrossRef](#)]
18. Amores, J.; Maes, P. Essence: Olfactory Interfaces for Unconscious Influence of Mood and Cognitive Performance. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 28–34.
19. Matsukura, H.; Yoneda, T.; Ishida, H. Smelling Screen: Development and Evaluation of an Olfactory Display System for Presenting a Virtual Odor Source. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 606–615. [[CrossRef](#)] [[PubMed](#)]
20. Dobbstein, D.; Herrdum, S.; Rukzio, E. inScent: A wearable olfactory display as an amplification for mobile notifications. In Proceedings of the 2017 ACM International Symposium on Wearable Computers, Maui, Hawaii, 11–15 September 2017; pp. 130–137.
21. Covington, J.A.; Agbroko, S.O.; Tiele, A. Development of a Portable, Multichannel Olfactory Display Transducer. *IEEE Sens. J.* **2018**, *18*, 4969–4974. [[CrossRef](#)]
22. Minecraft. Available online: <https://minecraft.net/en-us/?ref=m> (accessed on 6 April 2019).
23. Librosa. Available online: <https://librosa.github.io/librosa/> (accessed on 28 March 2019).
24. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Available online: <https://arxiv.org/abs/1409.1556> (accessed on 25 June 2019).
27. ImageNet. Available online: <http://www.image-net.org> (accessed on 20 March 2019).
28. TensorFlow. Available online: <https://www.tensorflow.org/> (accessed on 9 March 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).