Look and Listen: A Multi-modality Late Fusion Approach to Scene Classification for Autonomous Machines

Jordan J. Bird[1,2], Diego R. Faria[1,2], Cristiano Premebida[3], Anikó Ekárt[1] and George Vogiatzis[1,2]

***Abstract*— The novelty of this study consists in a multi-modality approach to scene classification, where image and audio complement each other in a process of deep late fusion. The approach is demonstrated on a difficult classification problem, consisting of two synchronised and balanced datasets of 16,000 data objects, encompassing 4.4 hours of video of 8 environments with varying degrees of similarity. We first extract video frames and accompanying audio at one second intervals. The image and the audio datasets are first classified independently, using a fine-tuned VGG16 and an evolutionary optimised deep neural network, with accuracies of 89.27% and 93.72%, respectively. This is followed by late fusion of the two neural networks to enable a higher order function, leading to accuracy of 96.81% in this multi-modality classifier with synchronised video frames and audio clips. The tertiary neural network implemented for late fusion outperforms classical state-of-the-art classifiers by around 3% when the two primary networks are considered as feature generators. We show that situations where a single-modality may be confused by anomalous data points are now corrected through an emerging higher order integration. Prominent examples include a water feature in a city misclassified as a river by the audio classifier alone and a densely crowded street misclassified as a forest by the image classifier alone. Both are examples which are correctly classified by our multi-modality approach.**

I. Introduction

'Where am I?' is a relatively simple question answered by human beings though it requires exceptionally complex neural processes. Humans use their senses of vision, hearing, temperature etc. as well as past experiences to discern whether they happen to be indoors, outdoors, and geolocate in general. This process occurs, for all intents and purposes, in an instant. Visuo-auditory perception is optimally integrated by humans in order to solve ambiguities; it is widely recognised that audition dominates time perception while vision dominates space perception. Both modalities are essential for awareness of the surrounding environment [1]. In a world rapidly moving towards autonomous machines outside of the laboratory or home, environmental recognition is an important piece of information which should be considered as part of interpretive processes of spatial awareness.

[1]J.J. Bird, D.R. Faria, A. Ekart, and G. Vogiatzis are with the School of Engineering and Applied Science, Aston University, Birmingham, United Kingdom. Emails: {birdj1, d.faria, a.ekart, g.vogiatzis}@aston.ac.uk

[1]J.J. Bird, D.R. Faria, and G. Vogiatzis are also with the Aston Robotics Vision and Intelligent Systems (ARVIS) Lab https://arvis-lab.io/

[3]C. Premebida is with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal. Email: cpremebida@isr.uc.pt

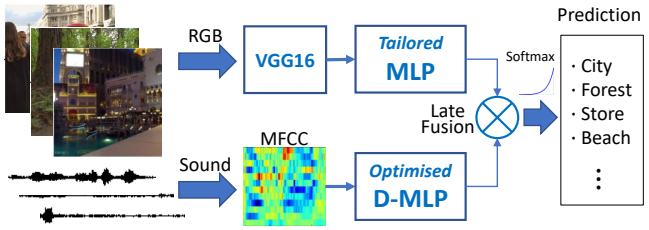

Fig. 1: The proposed multi-modality (video and audio) approach to scene classification. MFCC are extracted from audio-frames as input to an optimised DNN, while VGG16 and an ANN classify the images. We propose a higher-order function to perform late fusion.

Current trends in Robotic Vision [2]–[5] indicate two main reasons for the usefulness of scene classification. The most obvious reason is simply the ability of an awareness of where one currently is, but furthermore, and in more complex situations, the awareness of one's surroundings can be further used as input to learning models or as a parameter within an intelligent decision making process. Just as humans *'classify'* their surroundings for every day navigation and reasoning, this ability will very soon become paramount for the growing field of autonomous machines in the outside world such as self-driving cars and self-flying drones, and possibly, autonomous humanoid androids further into the future. Related work (Section II) finds that although the processes of classification themselves are well-explored, multi-modality classification is a ripe area enabled by the rapidly increasing hardware limits faced by researchers and consumers. With this finding in mind, we explore a bi-modal sensory cue combination for environment recognition as illustrated in Figure 1. This endows the autonomous machine with the ability to *look* (Computer Vision) and to *hear* (Audio Processing) before predicting the environment with a late fusion interpretation network for higher order functions such as anomaly detection and decision making. The main motivation for this is to disambiguate the classification process; for example, if a person were to observe busy traffic on a country road, the sound of the surroundings alone could be misclassified as a city street, whereas vision enables the observer to recognise the countryside and correct this mistake. Conversely, a densely crowded city street confuses a strong vision model since no discernable objects are recognised at multiple scales, but the sounds of the city street can still

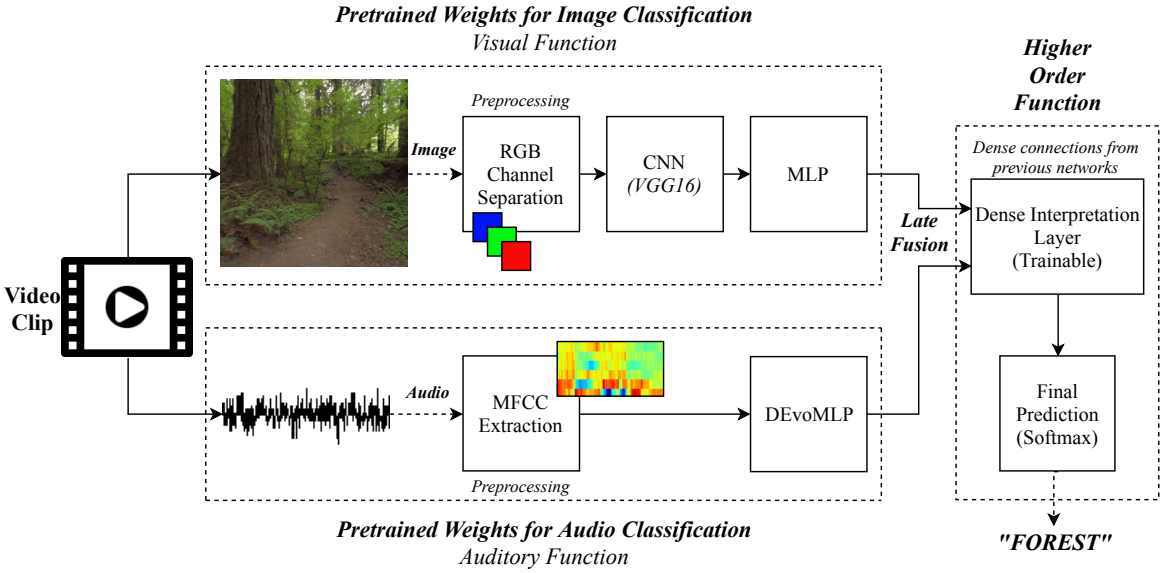

Fig. 2: Overview of the multi-modality network. Pre-trained networks without softmax activation layer take synchronised images and audio segments as input, and classify based on interpretations of the outputs of the two models.

be heard. Though this anomalous data point has confused the visual model, the interpretation network learns these patterns, and the audio classification is given precedence leading to a correct prediction. The main contributions of this study are centred around the proposed multi-modality framework illustrated in Figure 2 and are the following: (1) A large dataset encompassing multiple dynamic environments is formed and made publicly available.[1] This dataset provides a challenging problem, since many environments have similar visual and audio features. (2) Supervised Transfer Learning of the VGG16 model towards scene classification by training upon the visual data, together with engineering a range of interpretation neurons for fine-tuning, lead to accurate classification abilities. (3) The evolutionary optimisation of a deep neural network for audio processing of attributes extracted from the accompanying audio leads to accurate classification abilities, similarly to the vision network. (4) The final late fusion model combines and interprets the output of previously trained networks in order to discern and correct various anomalous data points that led to mistakes (examples of this are given in Section IV-D). The multi-modality model outperforms both the visual and audio networks alone, therefore we argue that multi-modality classification is a better solution for scene classification.

II. Related Work

Much state-of-the-art work in scene classification explores the field of autonomous navigation in self-driving cars. Many notable recent studies [6]–[8] find dynamic environment mapping leading to successful real-time navigation and object detection through LiDAR data. Visual data in the form of images are often shown to be useful in order to observe and classify an environment; notably 66.2% accuracy was achieved on a large scene dataset through transfer learning from the Places CNN[2] compared to ImageNet transfer learning and SVM which achieved only 49.6% [9]. Similarly Xie, et al. [10] found that through a hybrid CNN trained for scene classification, scores of 82.24% were achieved for the ImageNet dataset.[3] Though beyond the current capabilities of autonomous machine hardware, an argument has recently been put forward for temporal awareness through LSTM [11], achieving 78.56% and 70.11% pixel accuracy on two large image datasets. A previous single-modality study found improvement of scene classification ability by transferring from both VGG16 and scene images from videogames to photographic images of real-life environments [12] with an average improvement of +7.15% when simulation data was present prior to transfer of weights. In terms of audio, the usefulness of MFCC audio features in statistical learning for recognition of environment has recently been shown [13], gaining classification accuracies of 89.5%, 89.5% and 95.1% with KNN, GMM, and SVM methods respectively. Nearest-neighbour MFCC classification of 25 environments achieved 68.4% accuracy compared to a subject group of human beings who on average recognised environments from audio data with 70% accuracy [14]. It is argued that a deep neural network outperforms an SVM for scene classification from audio data, gaining up to 92% accuracy [15].

Researchers have shown that human beings use multiple parts of the brain for general recognition tasks, including the ability of environmental awareness [16], [17]. Though in many of these studies a single-modality is successful, we argue that, since the human brain merges the senses into a robust percept for recognition tasks, the field of scene

[1]Full dataset is available at:
https://www.kaggle.com/birdy654/scene-classification-images-and-audio

[2]http://places.csail.mit.edu/downloadCNN.html

[3]http://www.image-net.org

classification should find some loose inspiration from this process through data fusion. We explore visual and audio in this experiment due to accessibility, since there is a lot of audio-visual video data available to researchers. We propose that in the future further sensory data are explored, given the success of this preliminary experiment (Section V).

III. Problem and Method

The state-of-the-art is to interpret real-world data considering a single input. Conversely, the idea of multi-modality learning is to consider multiple forms of input [18].

Simply put, the question posed to a classifier is *'where are you?'*. Synchronised images and audio are treated as inputs to the classifier, and are labelled semantically. A diagram of this process can be observed in Figure 2;[4] visual and auditory functions consider synchronised image and audio independently, before a higher order function occurs. The two neural networks are concatenated into an interpretation network via late fusion to a further hidden layer before a final prediction is made. Following dataset acquisition of videos, video frames and accompanying audio clips, the general experimental processes are as follows. (i) *For audio classification:* the extraction of MFCCs of each audio clip to generate numerical features and evolutionary optimisation of neural network topology to derive network hyperparameters. (ii) *For image classification:* pre-processing through a centre-crop (square) and resizing to a 128x128x3 RGB matrix due to the computational complexity required for larger images, and subsequent fine tuning of the interpretation layers for fine-tune transfer learning of the VGG16 trained weight set. (iii) *For the final model:* freeze the trained weights of the first two models while benchmarking an interpretation layer for synchronised classification of both visual and audio data. This process is described in more detail throughout Subsections III-A and III-B.

A. Dataset Acquisition

Initially, 45 videos as sources are collected in varying length for 9 environmental classes at NTSC 29.97 FPS and are later reduced to 2000 seconds each: *Beach* (4 sources, 2080 seconds), *City* (5 sources, 2432 seconds), *Forest* (3 sources, 2000 seconds), *River* (8 sources, 2500 seconds) *Jungle* (3 sources, 2000 seconds), *Football Match* (4 sources, 2300 seconds), *Classroom* (6 sources, 2753 seconds), *Restaurant* (8 sources, 2300 seconds), and *Grocery Store* (4 sources, 2079 seconds). The videos are dynamic, from the point of view of a human being. All audio is naturally occurring within the environment. It must be noted that some classes are similar environments and thus provide a difficult recognition problem. To generate the initial data objects, a crop is performed at each second. The central frame of the second of video is extracted with the accompanying second of audio, an example of data processing for a city is shown in Figure 3. Further observation lengths should be explored in future. This led to 32,000 data objects, 16,000 images (128x128x3 RGB matrices) accompanied by 16,000 seconds (4.4 hours) of audio data. We then extract the the Mel-Frequency Cepstral Coefficients (MFCC) [20] of the audio clips through a set of sliding windows 0.25s in length (ie frame size of 4K sampling points) and an additional set of overlapping windows, thus producing 8 sliding windows. From each audio frame, we extract 13 MFCC attributes, producing 104 attributes per 1 second clip. MFCC extraction consists of the following steps: The Fourier Transform (FT) of the time window data ω is derived as $X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$. The powers from the FT are mapped to the Mel scale, the psychological scale of audible pitch [21]. This occurs through the use of a triangular temporal window. The Mel-Frequency Cepstrum (MFC), or power spectrum of sound, is considered and logs of each of their powers are taken. The derived Mel-log powers are treated as a signal, and a Discrete Cosine Transform (DCT) is measured. This is given as $X_k = \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right]$ where $k = 0, ..., N-1$ is the index of the output coefficient being calculated and x is the array of length N being transformed. The amplitudes of the spectrum are known as the MFCCs.

The learning process we present is applicable to consumer-level hardware (unlike temporal techniques) and thus accessible for the current abilities of autonomous machines.

B. Machine Learning Processes

For audio classification, an evolutionary algorithm [22] was used to select the amount of layers and neurons contained within a MLP in order to derive the best network topology. Population is set to 20 and generations to 10, since stabilisation occurs prior to generation 10. The simulation is executed five times in order to avoid stagnation at local minima being taken forward as a false best solution. Activations of the hidden layers are set to ReLu. For image classification, the VGG16 layers and weights [19] are implemented except the dense interpretation layers beyond the Convolutional layers, which is then followed by $\{2, 4, 8, \cdots, 4096\}$ ReLu neurons for interpretation and finally a softmax activated layer towards the nine-class problem. In order to generate the final model, the previous process of neuron benchmarking is also followed. The two trained models for audio and image classification have their weights frozen, and training concentrates on the interpretation of the outputs of the networks. Referring back to Figure 2, the softmax activation layers are removed from the initial two networks in order to pass their interpretations to the final interpretation layer through concatenation , a densely connected layer following the two networks and $\{2, 4, 8, \cdots, 4096\}$ ReLu neurons are benchmarked in order to show multi-modality classification ability. All neural networks are trained for 100 epochs with shuffled 10-fold cross-validation.

IV. Experimental Results

A. Fine Tuning of VGG16 Weights and Topology

Figure 4 shows the tuning of interpretation neurons for the image classification network. The best result was 89.27% 10-fold classification accuracy, for 2048 neurons.

[4]VGG Convolutional Topology is detailed in [19]

Fig. 3: Example of extracted data from a five second timeline. Each second, a frame is extracted from the video along with the accompanying second of audio.

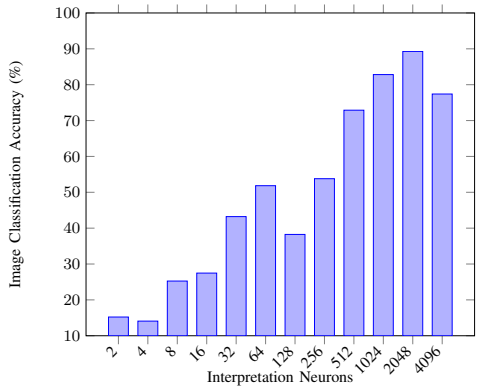

Fig. 4: Image 10-fold Classification Accuracy corresponding to interpretation neuron numbers.

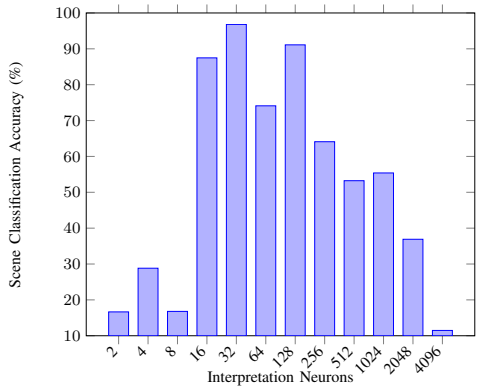

Fig. 5: Multi-modality 10-fold Classification Accuracy corresponding to interpretation neuron numbers.

TABLE I: Final results of the five Evolutionary Searches sorted by 10-fold validation Accuracy. *Conns.* denotes the number of connections in the network.

Simulation	Hidden Neurons	Connections	Accuracy
2	977, 365, 703, 41	743,959	**93.72%**
4	1521, 76, 422, 835	664,902	93.54%
1	934, 594, 474	937,280	93.47%
3	998, 276, 526, 797, 873	1,646,563	93.45%
5	1524, 1391, 212, 1632	2,932,312	93.12%

B. Evolving the Sound Processing Network

Regardless of initial (random) population, stabilisation of the audio network topology search occurred around the 92-94% accuracy mark. The best solution was a deep network of *977, 365, 703, 41* hidden-layer neurons, which gained 93.72% accuracy via 10-fold cross validation. All final solutions are presented in Table I. Interestingly, a less complex solution scores a competitive score of 93.54% accuracy with 79,057 fewer network connections.

C. Fine Tuning the Final Model

With the two input networks frozen at the previously trained weights, the results of the multi-modality network can be observed in Figure 5. The best interpretation layer was selected as 32, which attained a classification accuracy of 96.81% as shown in Table II. Late fusion was tested with other models by treating the two networks as feature

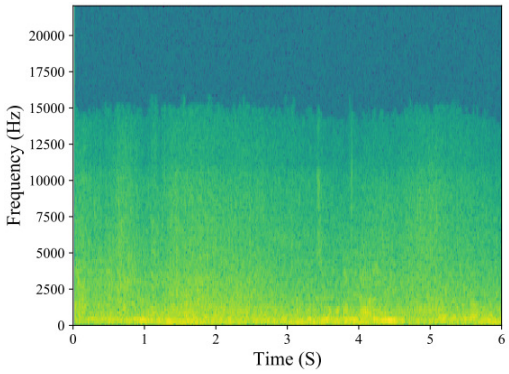

Fig. 6: Beach sonogram (with speech at 3s-4.5s)

generators for input, a Random Forest scored 94.21%, Naive Bayes scored 93.61% and an SVM scored 95.08%, which were all outperformed by the tertiary deep neural network.

D. Comparison and Analysis of Models

For final comparison of classification models, Table II shows the best performances of the tuned vision, audio, and multi-modality models, through 10-fold cross validation. Though visual classification was the most difficult task at 89.27% prediction accuracy, it was only slightly outperformed by the audio classification task at 93.72%. Outperforming both models was the multi-modality approach (Figure 2), when both vision and audio are considered

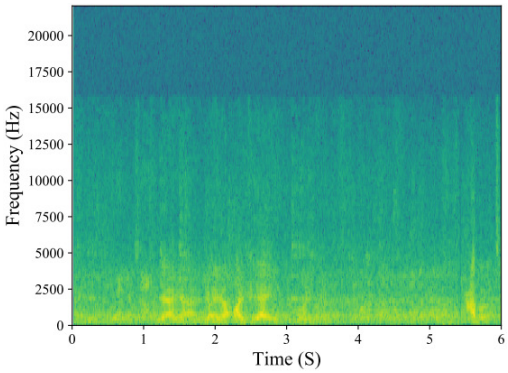

Fig. 7: Restaurant sonogram (with speech throughout)

Image: "CITY"
Audio: "CITY"
Multi-modality: "CITY"

Image: "FOREST"
Audio: "CITY"
Multi-modality: "CITY"

Fig. 8: An example of confusion of the vision model, which is corrected through multi-modality. In the second frame, the image of hair is incorrectly classified as the "FOREST" environment through Computer Vision.

through network concatenation, the model learns not only to classify both network outputs concurrently, but more importantly calculates relationships between them. An example of confusion of the audio model can be seen by the sonograms in Figures 6 and 7. Multiple frames of audio from the beach clip were mis-classified as *'Restaurant'* due to focus on the human speech audio. The image classification model on the other hand correctly classified these frames, and the multi-modality model did also. The same was observed several times within the classes *'City'*, *'Grocery Store'*, and *'Football Match'*. We note that these Sonograms show the

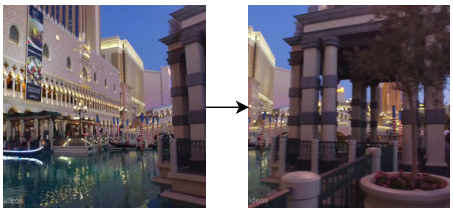

Image: "CITY"
Audio: "RIVER"
Multi-modality: "CITY"

Image: "CITY"
Audio: "RIVER"
Multi-modality: "CITY"

Fig. 9: An example of confusion of the audio model, which is corrected through multi-modality. In both examples, the audio of a City is incorrectly classified as the "RIVER" environment due to the sounds of a fountain and flowing water by the audio classification network.

TABLE II: Scene Classification ability of the three tuned models on the dataset

Model	Scene Classification Ability
Visual	89.27%
Auditory	93.72%
Multi-modality	**96.81%**

TABLE III: Results of the three approaches applied to completely unseen data (9 classes)

Approach	Correct/Incorrect	Classification Accuracy
Audio Classification	359/1071	33.52%
Image Classification	706/1071	65.92%
Multi-modality	**856/1071**	**79.93%**

frequency of the raw audio (stereo averaged into mono) for demonstrative purposes and MFCC extraction occurs after this point. Another example of this can be seen in Figure 8, in which the Vision model has been confused by a passerby. The audio model recognises the sounds of traffic and crowds etc. (this is also possibly why the audio model outperforms the image model slightly), the interpretation network has learnt this pattern and thus has *'preferred'* the outputs of the audio model in this case. Since the multi-modality model outperforms both single-modality models, this confusion also occurs in the opposite direction; observe that in Figure 9, the audio model has inadvertently predicted that the environment is a river due to the sounds of water, yet the image classifier correctly predicts that it is a city, in this case, Las Vegas. The multi-modality model, again, has learnt such patterns and has preferred the prediction of the image model, leading to a correct recognition of environment.

The results of applying the models to completely unseen

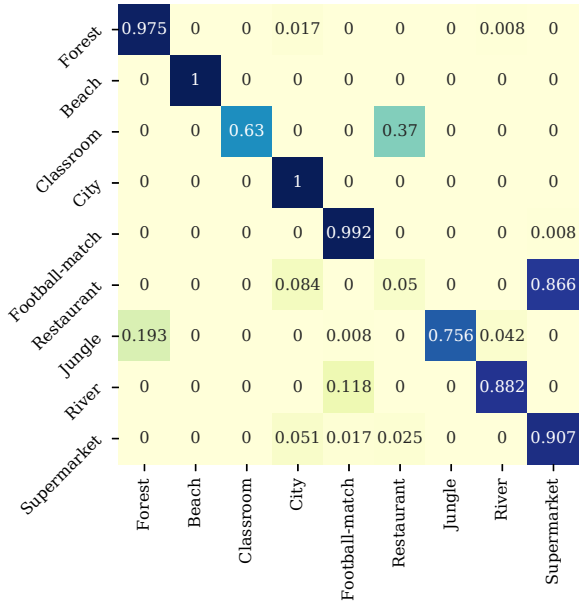

Fig. 10: Confusion matrix for the multi-modality model applied to completely unseen data

data (two minutes per class) can be seen in Table III. It can be observed that audio classification of environments is weak at 33.52%, which is outperformed by image classification at 65.92% accuracy. Both approaches are outperformed by the multi-modality approach which scores 79.93% classification accuracy. The confusion matrix of the multi-modality model can be observed in Figure 10; the main issue is caused by 'Restaurant' being confused as 'Supermarket', while all other environments are classified strongly. On manual observation, both classes in the unseen data both feature a large number of people with speech audio, we conjecture that this is possibly most similar to the supermarkets in the training dataset and thus the model is confident that both of these classes belongs to supermarket. This suggests that the data could be more diversified in future in order to feature more minute details and thus improve the model's abilities for discerning between the two.

V. Conclusions and Future Work

This study presented and analysed three scene classification models: (a) a vision model through fine-tuned VGG16 weights for classification of images of environments. (b) a deep neural network for classification of audio of environments and (c), a multi-modality approach, which outperformed the two original approaches through the gained ability of detection of anomalous data through consideration of the outputs of both models. The tertiary neural network for late fusion was compared and found to be superior to Naive Bayes, Random Forest, and Support Vector Machine classifiers. We argue that since audio classification is a relatively easy task, it should be implemented where available to improve environmental recognition tasks. This work focused on the context of autonomous machines, and thus consumer hardware capability was taken into account through temporal-awareness implemented within the feature extraction process rather than within the learning process. In future, better results could be gained from attempting to enable a neural network to learn temporal awareness in recurrence. Since the model was found to be effective with the complex problem posed through our dataset, future studies could concern other publicly available datasets in order to explore the applicability more widely. With the available hardware, evolutionary selection of network topology was only possible with the audio classifier. In future and with more resources, this algorithm could be applied to both the vision and interpretation models with the expectation to achieve a better set of hyperparameters beyond the tuning performed in this study. The model could also be applied in real-world scenarios. For example, it has recently been shown that autonomous environment detection is useful in the automatic application of scene settings for hearing aids [23]. Future works could also consider optimisation of the frame segmentation process itself as well as exploration of the possibility of multiple image inputs per task. Additionally, given the success of late fusion in this work, applications to video classification tasks could be considered through a similar approach.

References

[1] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017.

[2] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D traffic scene understanding from movable platforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 5, pp. 1012–1025, 2014.

[3] M. Cordts, T. Rehfeld, M. Enzweiler, U. Franke, and S. Roth, "Tree-structured models for efficient multi-cue scene labeling," *IEEE TPAMI*, vol. 39, no. 7, pp. 1444–1454, 2017.

[4] J. Xue, H. Zhang, and K. Dana, "Deep texture manifold for ground terrain recognition," in *IEEE/CVF CVPR*, pp. 558–567, 2018.

[5] K. Onda, T. Oishi, and Y. Kuroda, "Dynamic environment recognition for autonomous navigation with wide FOV 3D-LiDAR," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 530–535, 2018.

[6] F. Yu, J. Xiao, and T. Funkhouser, "Semantic alignment of LiDAR data at city scale," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1722–1731, 2015.

[7] C. Zach, A. Penate-Sanchez, and M. Pham, "A dynamic programming approach for fast and robust object pose recognition from range images," in *IEEE CVPR*, pp. 196–203, 2015.

[8] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *IEEE/CVF CVPR*, pp. 244–253, 2018.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems (NIPS)*, pp. 487–495, 2014.

[10] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid cnn and dictionary-based models for scene recognition and domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1263–1274, 2015.

[11] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3547–3555, 2015.

[12] J. J. Bird, D. R. Faria, P. P. Ayrosa, and A. Ekárt, "From simulation to reality: Cnn transfer learning for scene classification," in *2020 International Conference on Intelligent Systems (IS)*, IEEE, 2020.

[13] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *IEEE International conference on multimedia and expo*, pp. 885–888, 2006.

[14] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1941, 2002.

[15] Y. Petetin, C. Laroche, and A. Mayoue, "Deep neural networks for audio scene recognition," in *IEEE 23rd European Signal Processing Conference (EUSIPCO)*, pp. 125–129, 2015.

[16] M. P. Mattson, "Superior pattern processing is the essence of the evolved human brain," *Frontiers in neuroscience*, vol. 8, p. 265, 2014.

[17] M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student's handbook*. Psychology press, 2015.

[18] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Advances in Neural Information Processing Systems (NIPS)*, (USA), pp. 2222–2230, 2012.

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

[20] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[21] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[22] J. J. Bird, A. Ekárt, C. D. Buckingham, and D. R. Faria, "Evolutionary optimisation of fully connected artificial neural network topology," in *Intelligent Computing-Proceedings of the Computing Conference*, pp. 751–762, Springer, 2019.

[23] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *Journal on Advances in Signal Processing*, no. 18, pp. 387–845, 2005.