Audio Engineering Society

Convention Paper 10375

Presented at the 148th Convention, 2020 June 2-5, Online

Content matching for sound generating objects within a visual scene using a computer vision approach

Dan Turner[1], Chris Pike[2], and Damian Murphy[1]

[1]*University of York*
[2]*BBC Research and Development*

Correspondence should be addressed to Dan Turner (djt530@york.ac.uk)

ABSTRACT

The increase in and demand for immersive audio content production and consumption, particularly in VR, is driving the need for tools to facilitate creation. Immersive productions place additional demands on sound design teams, specifically around the increased complexity of scenes, increased number of sound producing objects, and the need to spatialise sound in 360°. This paper presents an initial feasibility study for a methodology utilising visual object detection in order to detect, track, and match content for sound generating objects, in this case based on a simple 2D visual scene. Results show that while successful for a single moving object there are limitations within the current computer vision system used which causes complications for scenes with multiple objects. Results also show that the recommendation of candidate sound effect files is heavily dependent on the accuracy of the visual object detection system and the labelling of the audio repository used.

1 Introduction

With the rise in popularity of spatial audio content in recent years, the term "immersion" or "immersive" has been used to describe a plethora of content related to new audio experiences. However, it is not uncommon for this term to be used vaguely and interchangeably with others, for example, naturalness, envelopment, presence, or realism [1], and differing definitions can cause confusion [2].

It is helpful to have a clear definition of what is meant by 'immersion' and therefore what would constitute immersive content and therefore immersive sound design. The following definition of immersion is used in this research, as defined by Argrawal et al [2]:

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

Argrawal concluded there are two prominent perspectives on immersion, the first being an individual's psychological state and the second being an objective property of a technological system [2]. The latter is rejected

by Argrawl due to its exclusion of the individual and their experience, which, being highly subjective, is seen of paramount importance.

Immersive content can, therefore, be described as content which is designed to elicit a state of immersion in line with the given definition. While not a prerequisite, many modern day immersive experiences use technologies such as 360° audio and video to provide the subjective sense of being surrounded and, in the case of audiovisual content, provide an experience of multi-sensory stimulation. The production of such content has become increasingly popular in recent years with companies such as BBC [3], Facebook [4], and Google [5], releasing tools and content for such experiences.

In this paper we propose a methodology, based on an initial feasibility study, that aims to streamline immersive sound design workflows through the application of a computer vision approach that facilities the detection, spatial tracking, and content matching of appropriate audio assets to sound generating objects within a visual scene. The current study is undertaken using a simple 2-D scene as a proof of concept with a view to expanding to 360°spatial audio/video in future work.

2 Background

2.1 Sound Design for Immersive Content

Sound design is a well established area of practice and is traditionally associated with the design and production of sound for the purposes of film, TV, radio programes, and video games. Such sounds often fall broadly into the categories of sound effects, dialogue, and music [6]. In the context of this paper we are specifically interested in sound design practice for new immersive film, TV, and radio content.

The aesthetics and practice around these content forms are still new or as yet unknown, but it is clear that workflows are changing from established practices such that sound designers will need to adapt. At present, however, little formal research has been conducted on the subject. Immersive audio production often places additional demands on sound designs teams, with little or no additional time allocated to complete the task. By their very nature, 360° scenes are often more complex and contain many more sound cues or sound generating objects, and require a greater level of detail in order to be plausible and 'fill' the additional visual and hence auditory space. The users have a greater level of interactivity with the scene allowing them to control how they view and focus on individual aspects within it, and often the audience is no longer limited by the framing of a single shot.

Due to this extended visual field, sound spatialisation plays an important role in creating immersive content by facilitating the positioning of audio sources around the complete 360°space. The aim being to create a state of immersion by increasing the sense of realism and the ability to interact with the virtual environment [7]. It also allows for increased engagement (or, more specifically, the lack of disruption caused to a user's engagement) by ensuring consistency within the environment such that visual and auditory information is considered spatially congruent [7]. Furthermore, sound spatialisation can also take advantage of listener expectation to reinforce the sense of immersion [8] with respect to sound location (such as aircraft emanating from above the listener).

Spatialisation of sound can, however, be a time consuming task and one which has seen the development of a multitude of tools for use within Digital Audio Workstations (DAWs) [9, 10, 4]. Automating aspects of the process could allow sound designers to have a greater focus on those tasks requiring creative decision making as opposed to those where the decision making process is simple, but the task itself is more procedural, iterative, or labour intensive.

A common component of many immersive experiences (though admittedly not all) is the presence of visual information, normally in the form of a video. Using this readily available data correctly could provide a wealth of information about the audio scene being constructed and this could then be utilised to inform context-aware automated or semi-automated audio outcomes. The utilisation of computer vision techniques could lend themselves to this application of machine-assisted sound design, especially within the context of automatic content matching and object tracking.

2.2 Visually driven sound design

Computer vision is an established area of machine learning that focuses on making sense of the information contained within digital images and videos. Two of the most common tasks involved in many computer

vision applications are object detection (localisation of objects within a given image) and object classification (estimating which of a given class the object is most likely to belong to). Algorithms for such tasks will often have to deal with evaluating multiple objects within a given image.

Within the field of sound design there are some examples of how visual features can be matched to audio files in a database or used to synthesise sounds from this visual information [11, 12].

Owens *et al.* [11] trained a recurrent neural network (RNN) to map visual features to audio features, which were then transformed into a waveform by either matching them to already existing audio files in a database or by, what the authors describe as, parametrically inverting the features. The sounds synthesised were of people hitting and scratching different surfaces and objects with a drum stick. While this is still somewhat distant from complex soundscape creation, this outlines a general approach that could be used in order to produce other plausible sound objects. Performance of the model were measured via a psychophysical studying using a two-alternative forced choice test where participants were required to distinguish real and machine-generated sounds. Results were mixed, with parametric generation performing well for materials which were considered more noisy, for example leaves and dirt, but performing poorly for harder surfaces such as metal and wood. It was also found that matching the mapped audio features to existing audio files was ineffective for textured sounds such as splashing water.

The online sound installation Imaginary Soundscapes [12] creates 'psuedo' soundscapes by extracting Convolution Neural Network (CNN) features from an image and matches these with corresponding features for sound files from a database of environmental sounds. These features were extracted by using a CNN architecture based on Sound-Net [13].

While the authors of [12] document no formal testing, it appears that the system is capable of extracting relevant sound files based on the location of specific objects within the scene. However, at times the choices made based on the scene's visual content can be slightly ill-suited, such as audio which appears to a recording of a train station (including crowd noise and announcements over a loudspeaker system) matched to an image of a church interior. If the aim of the system is to produce more accurate/plausible soundscapes this could be accomplished by adding a human user in the loop to rate the suitability of the audio content presented for the given scene. Alternately, the system could also provide options for the human user to choose between, which would be more akin to the functionality of a production tool.

A user's ability to interact with the installation in [12] is very limited. They are able to provide an image or select a location but have no control over the resulting soundscape. The aim of this work is to use computer vision as a collaborative agent, much in the same way other machine learning techniques have been utilised to act as collaborators for music composition [14] and sound synthesis [15].

3 Methods

3.1 Google's object detection API

It should be noted that the scope of this feasibility study has been limited to the use of existing visual recognition tools, with the method adopted in this paper being built around Google's object detection API [16]. The API is written in TensorFlow and is used to detect, locate, and classify content from a simple 2D video frame image. The tutorial for the API is easily adapted to run detection on frames of a video following instructions detailed in [17]. The description which follows of how this is applied is illustrated in Fig. 1 and shows flow of information and the different components that make up the wider system.

The model used for this paper is the Single Shot Detection (SSD) meta-architecture, with the inception V2 feature extractor, chosen because it gave a good compromise between speed, accuracy, and memory usage. The model was run on an Intel Core i5-600 CPU @ 3.20GHz, with 8GB RAM, and an Intel HD 530 integrated graphics processor. It should be noted that the specifications of the computing platform being used will greatly influence the time it takes to run the detection. Using the aforementioned specification it took approximately 75 seconds to run the process of detection and information extraction on a 7.97s video filmed at 30 frames per second (fps). This roughly equates to 0.32s per frame. This can be reduced to between 65s (0.27s per frame) if the visualisation of the detection output is bypassed.

The object detection system provides a variety of data related to each frame including number of detections,

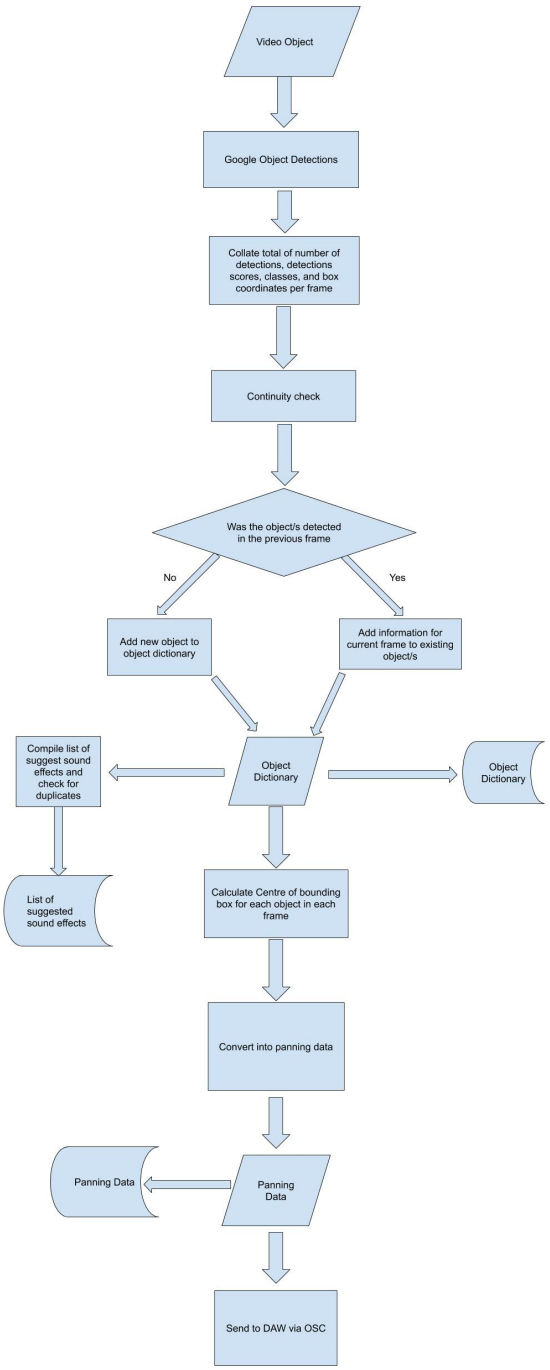


Fig. 1: Flow chart illustrating order of operations and flow of data within the proposed methodology

classes detected, detection scores (confidence), and object bounding box coordinates. The system first collates this data for each frame so it can be used to create the object dictionary. This contains a unique ID number for each detected object, class number of the object detected, and the coordinates relating to the bounding box position of each object. Following the collection of this data it is then used in several processes outlined in the following sections.

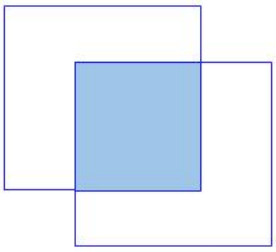
3.2 Continuity between frames

For the proposed system to be successful it must be able to correctly group the data for each detected object across successive frames or create a new object ID if a detected object is considered as being new to the scene. The original API is designed for detection on a single image and this is done iteratively over each frame of video input, but has no internal reference or 'memory' for anything taking place during previous iterations or frames. Each frame is considered as a standalone individual image rather than constituent parts of a wider object where each frame will (usually) have some temporal and/or spatial relationship with those preceding it. This results in a higher chance of misclassification between frames.

The use of single image detectors being used for many video object detectors is an acknowledged problem. For instance, they are unable to take advantage of available temporal information such as objects in adjacent frames being in similar locations [18], which can lead to lower confidence levels and misclassification between frames. There is a wealth of current research within the computer vision community to increase the accuracy of video object detection system by exploiting the available temporal information see, e.g. [18, 19]. In this implementation, to account for between frame misclassifications, and to accurately group object data across multiple frames to their associated object IDs, a basic continuity check has been implemented. An adaption of the Intersection of Union (IoU) evaluation metric was used, which as described in [20], is a common method of comparing the similarity between arbitrary shapes by calculating a normalised measure that focused on the areas of the shapes. This measure is a ratio of the area of intersection (Fig. 2a) over the area of union (Fig. 2b). Traditionally this is a metric used when training object detection systems and is calculated using ground truth boxes (hand labelled bounding

boxes for the testing set specifying the location of the objects) and prediction boxes (boxes generated by the object detection system indication where it predicts the objects are located). Accuracy is deemed sufficient if the IoU value exceeds a user specified amount (usually 0.5<) with value ranges between 0 and 1. This metric is appropriate as it is expected that an object in the current frame will be in a similar location to its position in the previous frame. If the IoU value is above the set threshold the object is defined as being the same as that identified in the previous frame, otherwise it is treated as a new object and is added to the object dictionary.

It should be noted this method has limitations which are discussed in sec. 4.2.2



(a) Area of Intersection

(b) Area of Union

Fig. 2: IoU can be calculated by dividing the area of intersection (the area covered by the overlap of the two boxes) by the area of union (total area covered by the two boxes). Within this work, it is used as a continuity check on objects within the visual scene taking advantage of the similar locations an object will occupy within the current and previous frame

3.3 Sound Effects Suggestions

Once the object dictionary has been compiled it is used to generate a list of suggested sound effects from the chosen repository of audio files, which in this case is the BBC sound effects archive [21]. In this instance each unique object class detected is compared to the metadata tag from the BBC's sound effect archive [21], which is an open source repository made up of 16,011 labelled audio files. The archive is are available to download as WAV files and is subject to terms of use under the RemArc Licence, which permits use for personal, educational, and research purposes. Chosen because it provides a large database of labelled audio files containing a variety of different acoustic scenes and events, with tagging and metadata stored in an associated .CSV file. Table 1 shows examples of tagging and metadata format common to each audio file in the database. Tagging consists of the description of each sound effect (as taken from the original CD) and the category (e.g. Engines: Petrol, Engines: Diesel) to which it belongs. Metadata stored is the length of audio file in seconds, name of the original CD containing the effect, and track number. There are some inconsistencies within the tagging conventions with not all audio files having an associated category and/or CD origin name. Any inconsistencies within a database's tagging convention may impact its effectiveness when used as data for training and evaluating machine listening systems [22].

3.4 Object Tracking

Alongside checking for continuity, object location data can also be used to calculate the trajectory for each object over the course of the video, which can then be utilised to position and pan audio content. Object trajectory is calculated by calculating the centre point of an objects bounding boxes as shown in Fig. 3.The data can then be transmitted to DAWs such as Cockos Reaper [23] via OSC [24] in order to populate automation data for the desired parameter. In the case of stereo panning the horizontal portion of the trajectory data needs to be normalised to between 0 (hard left) and 1 (hard right). Due to the temporal resolution available in Reaper's automation lanes, resolution of location data was reduced by a factor of two, resulting in 15 discrete points per second for a 30fps video.

3.5 Test Material Specification

Two test videos were created to allow for direct and controlled evaluation of simple scenes containing a single and multiple objects. A photographic image containing animals was also sources from the internet in

Description	Duration (s)	Category	CD Number	CD Name	Track #
Two-stroke petrol engine driving small elevator, start, run, stop.	194	Engines: Petrol	EC117D	Diesel and Petrol Engines	4
Single-cylinder Petter engine, start, run, stop. (1 1/2 h.p.)	194	Engines:Diesel	EC117D	Diesel and Petrol Engines	1
Single hen	63		EC31A	Chickens	1
Motorcycle Scrambling: General atmosphere, pre-1965 machines, 250-500cc	194	Motorcycle Scrambling and General Atmosphere	EC5M4		1

Table 1: Examples of metadata format associated with BBC's sound effect archive. Available metadata fields consist of a description, duration in seconds, category, CD number, CD Name, and track number. As shown there is inconsistency within the archive as not all audio files will contain information for within the category and CD name fields

Fig. 3: Single frame taken from a test video with preceding trajectory of the detected object has been plotted and shows a good match.

order to assess capabilities relating to candidate audio file recommendation for non-human objects. All videos were recorded on an iPhone SE at 1080p 30 frames per second at a distance of 5m and have the following conditions

- Video 1 – Single person walking from left to right of scene.
- Video 2 – Two people walking ∼1.5m apart from left to right of scene

Example images from the two video examples are shown in Fig. 4 and Fig. 5.

4 Results

4.1 Candidate Sound Effects Recommendations

The system takes approximately 4s to compile a list of candidate sound effect recommendations for Video 2, returning a total of 36 recommendations (a selection of which are shown in Table 2) of which 6 were considered usable for the given scene. Those deemed unsuitable were for reasons such as a different environment/activity to the one in the example video (e.g. a person exiting a car and a person in an ice skating rink). The current implementation takes the class label as a string of characters and compares this to the tags in the metadata. If an exact match is found it will determine the associated audio file as being a candidate sound effect. A limitations of this method is the reliance on exact matching of tags between the database and the class labels of the detection system. It is therefore unable to recommend audio files which may be suitable but whose tags use different (yet still related) terms, such as 'man', 'woman', or 'human' if detecting the class 'person'. Tagging within the BBC archive is inconsistent (admittedly due to the repository consisting of many decades worth of archived audio files) meaning many potentially suitable sound effects go undetected using the current string comparison method.An alternative method which may alleviate this issue would be to train another machine learning algorithm to detect and

Fig. 4: Image from a single video frame extracted from example Video 1, and used as input for the object detection system to generate candidate audio file recommendations. The detected object's location is indicated by the green bounding box and is assigned the class label of 'person'.

Fig. 5: Image from a single video frame of Video 2 used to derive panning information for two moving objects with a 2D visual scene. The example video is of two people crossing the field of view from left to right approximately 1.5m apart.

recognise synonyms, which is often associated with lexical substitution tasks [25] that require systems to predict alternatives for a target word, while maintaining its meaning with a sentences context. Within our use case this can be used in order to suggest candidate audio files whose tags may not exactly match the detected class but are deemed, by the system, to have similar meaning.

Limitations also exist relating to the type of detection system used. Google's API is for object detection and is limited to detecting the specific objects. It therefore does not allow for prediction of activities taking place within the scene, such as walking as in Videos 1 and 2. As such, the system did not retrieve the 1,484 sound effects containing the term 'footsteps' which may have been suitable as candidate sound effects. It also lacks the functionality of scene recognition systems to predict more generic scene elements such as location e.g. living room, beach, city centre, which may help to inform recommendations for audio files relating to environmental/atmosphere sounds.

4.2 Spatial positioning and trajectory tracking

4.2.1 Single Object

Fig. 3 shows a single frame taken from Video 1 where the trajectory of the detected object has been plotted. The trajectory appears to accurately track the object travelling across the field of view and takes into account the variations in centre point of the bounding box that occurs when walking, and variation in the object's speed, in this case indicated by the non-uniform distribution in spatial proximity of the data points.

Fig. 6a shows horizontal panning data plotted over time in frames derived from the objects positional data. This highlights the precision at which the system will take into account the object's variation in speed as it crosses the field of view as relating with the gradient of the plotted data. It should be noted that it is the distance moved by the centre of the objects associated bounding box between each frame that is being tracked rather than the object itself. For objects whose movement causes bounding boxes of varying sizes (such as a human walking with their arms swinging) this may, produce variable results. Once the object exits the field of view the panning value defaults to 0 which may present problems for objects whose audio needs to remain active for a set time after being no longer visible. However, this is an issue that is a feature of the current 2D only implementation. Field of view in 360°audio/visual content is often dictated by the direction a user is facing, therefore allowing objects outside the field of view to be still be tracked as the video content extends beyond the field of view. There will, however, be additional, alternative limitations and situations that will need to be considered when using 360°audio/visual content.

Fig. 6b shows the horizontal trajectory data translated into panning information within a Reaper stereo track

Candidate Audio File Recommendations
Walking, 1 person in mud
Footsteps, one person walking in mud
Cars: 1.6 GL (Manual) 1982 model Ford Cortina. Interior, door opens, person exits, door closes
Ice Skating, one person circling close, others in the distance on indoor rink
Footsteps, one person walking in water

Table 2: Selection of candidate audio file recommendations generated from Fig. 4. Each file was defined by the system as being a potential candidate if the metadata field 'description' contained an exact match for the detected objects class name, in this case 'person'.

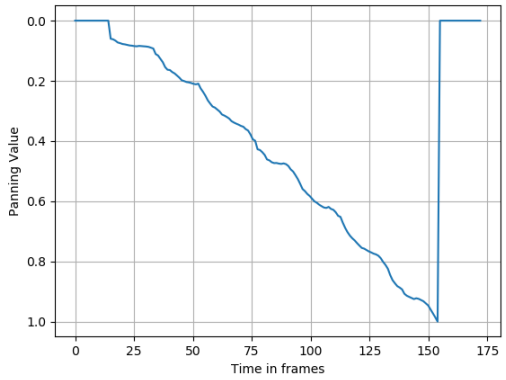
automation lane. Upon visual inspection, the reduction in data does not seem to have had an adverse effect on trajectory trends. The linear interpolation generated by Reaper has little impact on the overall trend due to the size of the time steps but may have a perceptual impact for larger timesteps. The timestep is dependent on the videos fps and is the length of time between each discrete data point of panning data (in this case the timestep is 0.0667s). A reduction in fps results in an increased timestep duration which may introduce greater spatial mismatches between the visual object and the associated auditory material. The results from previous literature vary greatly with respect to the angular offset required in order to create a perceptually noticeable misalignment. Using reaction time (RT) measurements as an indirect method of measurement there was found in [26] to be a significant difference measured from 5° to 10° on wards for the Simon effect (the observation that responses in two-alternative forced-choice-tests, where space is a parameter not relevant to the task, are faster if the stimulus presentation and response side match; responses are slower if the stimulus is presented in the visual hemisphere opposite of the response side) and it was concluded that that for speech signals, even small audio-visual offsets can subconsciously influence the spatial integration of sources. Future work will ascertain the minimum resolution of panning data required to maintain congruence between the objects visual position and the position of the associated audio content.

4.2.2 Multiple Objects

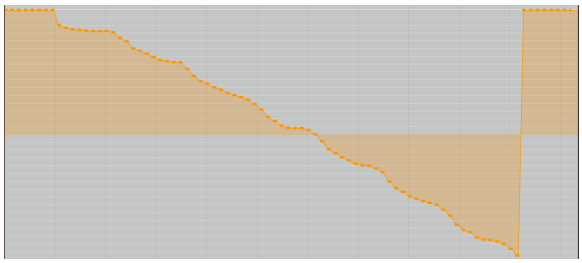
While the system handles the tracking of single objects well, multiple objects introduce extra complexities. When presented with a scene containing multiple objects the API will store and output the data for the detected objects according to the associated confidence scores, beginning with the highest score. This results in the data for specific objects being output in a different order for each frame depending on how the confidence scores change throughout the video. This then affects how the data interacts with the object dictionary compiler. The system compiles the object dictionary, and therefore how the trajectory data is grouped, according to the results of the continuity check. This relies on the detected object's data being output in the same order each time. When presented with the objects in a different order it causes the check to use positional data from a different object, and if the distance between the objects is great enough (which usually it will be), the continuity check fails and the objects in the current frame are defined as new. This causes the panning data from what should be a single object to be spread across multiple entries within the object dictionary. The change in confidence score over the length of the video resulted in a total of 32 objects being added to the dictionary. Due to the object detector in this implementation being based on a single image detector, it is not straightforward to override the ordering method in order to create a more consist output order on a frame-by-frame basis. This presents a problem for projects that require not only accurate location and classification of objects, but the ability to track them through frames and for them be recognised as pre-existing or new objects. It may be possible to address this problem using an alternative object detection system.

5 Conclusions

In this paper, a methodology for detecting, tracking, and matching content for sound generating objects within a simple visual scene has been presented. Work to date allows for successful interaction between a large



(a) Original data output from system. Note the y axis has been flipped to match Reaper's and the data has been normalised to between 0 and 1 to match the values used by Reaper.



(b) Stereo panning data was derived by using every second data point to account for the resolution available in Reaper's automation lanes.

Fig. 6: Horizontal panning data plotted over time as derived from example Video 1.

labelled audio repository and the visual content of a simple 2D scene as taken from a video. Suggested positional data for dynamic audio content can be attached to a single visual object within a scene. However, at present, is unable to support multiple objects due to limitations of the system being used. Erroneous results are also produced from the candidate sound effects search dependent on the accuracy or interpretation of the labels attached to the database of audio files.

6 Future Work

Future work will look into current methodologies for tracking objects throughout a scene ensuring unique identification is maintained, such as the use of Kalman Filters [27]. It would also be of interest to investigate the numerical relationship between an object's position within a 2D visual scene and the panning value a sound designer may assign, as it may not be the case that an object at the extremes of the visual field will be assigned a value at the extremes of the available panning values. This research also raises questions of the impact this kind of technology could have on the current working practices of sound designers working with immersive content and user testing would be carried out once a suitably functioning prototype has been developed. Finally, future work will utilise a GPU in order to reduce time taken to run detection and extract the relevant data.

7 Acknowledgements

This project is support by an EPRSC iCASE PhD Studentship in partnership with BBC R&D

References

[1] Francombe, J., Brookes, T., and Mason, R., "Evaluation of spatial audio reproduction methods (Part 1): Elicitation of perceptual differences," *AES: Journal of the Audio Engineering Society*, 65(3), pp. 198–211, 2017.

[2] Agrawal, S., Simon, A., Bech, S., Bærenstein, K., and Forchhammer, S., "Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences," *AES 147th Convention*, pp. 1–11, 2019.

[3] BBC Research and Development, "Immersive and Interactive Content - BBC R&D," 2019.

[4] FaceBook, "Spatial Workstation ," 2019-12-17.

[5] Google, "Google VR," 2019.

[6] Sonnenschien, D., *Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema*, Michael Wiese Productions, 2001.

[7] Salselas, I. and Penha, R., "The role of sound in inducing storytelling in immersive environments," in *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound - AM'19*, pp. 191–198, ACM Press, New York, New York, USA, 2019.

[8] Chueng, P., Chueng, P., Marsden, P., and Marsden, P., "Designing auditory spaces: the role of expectation," *Proceedings of 10th International Conference on Human Computer Interaction*, pp. 616–620, 2003.

[9] S3A Project Team, "VISR Production Suite," 2019-12-17.

[10] Stitt, Peter, "SSA Plug-ins," 2019-12-17.

[11] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T., "Visually Indicated Sounds," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2413, IEEE, 2016.

[12] Kajihara, Y., Dozono, S., and Tokui, N., "Imaginary Soundscape : Cross-Modal Approach to Generate Pseudo Sound Environments," *Workshop on Machine Learning for Creativity and Design (NIPS 2017)*, (Nips), pp. 1–3, 2017.

[13] Aytar, Y., Vondrick, C., and Torralba, A., "SoundNet: Learning Sound Representations from Unlabeled Video," (Nips), 2016.

[14] Fiebrink, R. A., "Real-time Human Interaction with Supervised Learning Algorithms for Music Composition and Performance," (January), 2011.

[15] Miranda, E., *Sound Design: An Artificial Intelligence Approach*, Phd, University of Edinburgh, 1994.

[16] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K., "Speed/accuracy trade-offs for modern convolutional object detectors," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 3296–3305, 2017.

[17] Vladimirov, L., "Detect Objects Using Your Webcam," n/a.

[18] Zhu, M. and Liu, M., "Mobile Video Object Detection with Temporally-Aware Feature Maps," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5686–5695, 2018.

[19] Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., and Ouyang, W., "T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), pp. 2896–2907, 2018.

[20] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S., "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression," 2019.

[21] BBC, "BBC Sound Effects Archive Resource," *BBC Sound Effects Archive Resource • Research & Education Space*, 2019.

[22] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D., "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, 17(10), pp. 1733–1746, 2015.

[23] Cockos, "REAPER | Audio Production Without Limits," 2019.

[24] Wright, M., "Open Sound Control 1.0 Specification," 2002.

[25] Melamud, O., Levy, O., and Dagan, I., "A Simple Word Embedding Model for Lexical Substitution," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 1–7, Association for Computational Linguistics, Stroudsburg, PA, USA, 2015, doi: 10.3115/v1/W15-1501.

[26] Stenzel, H., Francombe, J., and Jackson, P. J., "Limits of perceived audio-visual spatial coherence as defined by reaction time measurements," *Frontiers in Neuroscience*, 13(MAY), pp. 1–17, 2019, ISSN 1662453X, doi:10.3389/fnins.2019.00451.

[27] Saho, K., "Kalman Filter for Moving Object Tracking: Performance Analysis and Filter Design," in *Kalman Filters - Theory for Advanced Applications*, p. 13, 2017.