PANOPTIC RECONSTRUCTION OF IMMERSIVE VIRTUAL SOUNDSCAPES USING HUMAN-SCALE PANORAMIC IMAGERY WITH VISUAL RECOGNITION

Mincong (Jerry) Huang, Samuel Chabot, Jonas Braasch

Rensselaer Polytechnic Institute
110 8th St, Troy, New York, United States
huangm5@rpi.edu

ABSTRACT

This work, situated at Rensselaer's Collaborative-Research Augmented Immersive Virtual Environment Laboratory (CRAIVE-Lab), uses panoramic image datasets for spatial audio display. A system is developed for the room-centered immersive virtual reality facility to analyze panoramic images on a segment-by-segment basis, using pre-trained neural network models for semantic segmentation and object detection, thereby generating audio objects with respective spatial locations. These audio objects are then mapped with a series of synthetic and recorded audio datasets and populated within a spatial audio environment as virtual sound sources. The resulting audiovisual outcomes are then displayed using the facility's human-scale panoramic display, as well as the 128-channel loudspeaker array for wave field synthesis (WFS). Performance evaluation indicates effectiveness for real-time enhancements, with potentials for large-scale expansion and rapid deployment in dynamic immersive virtual environments.

1. INTRODUCTION

The project is situated at Rensselaer's Collaborative-Research Augmented Immersive Virtual Environment Laboratory (CRAIVE-Lab; see Figure 1). [1] A human-scale room-centered virtual reality system, the CRAIVE-Lab consists of a 360-degree panoramic display, surrounded by a 128-channel loudspeaker array for spatially accurate audio reproduction. Situated as an augmented built environment, the CRAIVE-Lab has enabled audiovisual rendering that could be experienced collectively in a spatially congruent format without any confinement and dislocation to personal sensory functions, while preserving the immediacy, plausibility, and proximity of virtual reality experience.

The CRAIVE-Lab has a track record of environmentally-situated human-scale audiovisual reproduction research, using techniques such as panoptic/ambisonic field recording [2] and machine-learning-based audio classification [3]. While the co-collection of in-situ audiovisual data in these works ensures a high degree of congruence in rendering, it faces a limitation where the information collected becomes incomplete. A notably more common data collection scenario is the capturing of geo-spatial data, which often does not come with co-located audio information.

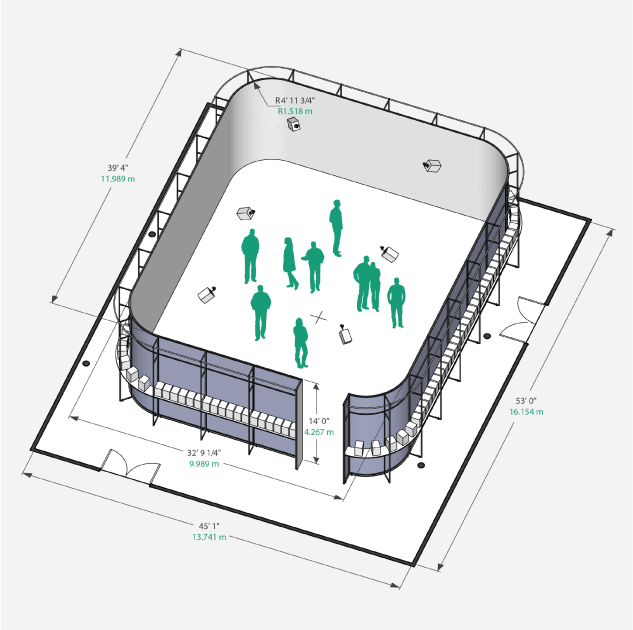

Figure 1: Isometric view of the Collaborative-Research Augmented Immersive Virtual Environment Laboratory (CRAIVE-Lab) at Rensselaer Polytechnic Institute.

Therefore, to effectively render these environments with higher fidelity, a reconstruction method is demanded in the audio domain.

The motivation behind this project originates from both the advancements in visual scene analysis and the recontextualization of soundscape studies that finds its root in the discipline of architecture and urban planning [4]. Specifically, with the systematic construction of visual recognition datasets [5], the ability of visual scene analysis to infer meaningful spatial information becomes systematized and strengthened. Yet such structural robustness has yet to be manifested in the context of acoustic scene analysis. While existing audio data collection has been extensively focusing on semiotically singular sound events, methods of comprehensive environmental acoustic data collection and classification remains a subject of active research [6]. In the context of audio-visual display, this makes full acoustic representation of existing environments difficult to achieve. The project intends to tackle this challenge by combining visual tectonics with human-scale sound mapping.

Recently emerging research efforts also give rise to the con-

https://doi.org/10.21785/icad2021.043

cept of audiovisual fusion, correlating acoustic scene analysis with spatially accurate visual scene analysis, often through the use of techniques related to machine learning [7, 8, 9]. While these works still rely on in-situ audio and visual data, they address and encode the spatial correspondence in their rendering that simple capturing cannot directly represent. This reveals the technical possibility that audio information, if not present, can be reconstructed through a detailed spatial analysis of visual environments with a collection of generic audio resources, especially when visual environments can be presented coherently.

The artificial reconstruction of visually derived soundscape is, in essence, also a form of product-sound design. In the context of product sounds, layers of soundscapes facilitate the formation of listening structures, conveying information that serves contextual, symbolic, and syntactic meanings [10]. To this end, it is highly relevant to the perceptual organization of auditory information, a subject of interest in auditory scene analysis [11] and auditory object formation [12]. Moreover, it also receives interest in the realm of contemporary music theory, where the cognition of sound *objects* has extensively concerned itself with the morphology and typology of the listening environment [13]. The development of approaches that allows for more comprehensive quantitative assessments of these concepts could be significantly benefited from the extensive availability of visual resources.

The project described in this paper is significant in that it introduces coherent audiovisual rendering schema into the context of room-centered immersive virtual reality systems, especially in situations when the only information presented are virtual representations of visual environments (landscape) and corresponding acoustic information is completely absent. With the work involved in this project, we intend to arrive at equally plausible virtual representations of soundscapes that could be experienced in conjunction with a visual ground truth at the CRAIVE-Lab, enabled by its spatial audio reproduction capability. Beyond the novelty of its context, another goal is to further enhance CRAIVE-Lab's capacity for collective experiences of virtual spaces, with minimal intervention to the bodies of immersed individuals, thereby laying the groundwork for further studies of collaborative behaviors in room-centered immersive systems.

2. RESEARCH METHODS

2.1. Overall Framework

Influenced by Schafer's definition of *keynotes*, *signals*, and *soundmarks* [15] in his taxonomy of soundscape elements, existing categorization method for soundscape has been standardized into three components: foreground sounds, background sounds, and contexts [14]. This facilitates a framework of audiovisual correspondence in this research (see Figure 2). Two visual recognition techniques are used: 1) semantic segmentation, which classifies visual imagery into physically meaningful elements on a pixel-by-pixel basis; and 2) object detection, which, with a similar image processing method as semantic segmentation, extracts visual objects with locally precise spatial information. In this research, semantic segmentation is used to configure background sounds, for which the spatial information is encoded as audio display regions *regions*; while object detection is used to translate quantifiable visual entities into meaningful information for foreground sound objects, which are encoded with locally precise spatial positions in virtual space. When executed sequentially (see Figure 3), the spatial

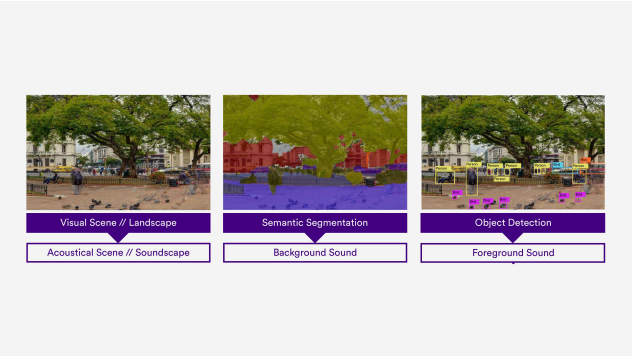

Figure 2: A taxonomic schema of audiovisual correspondence based upon existing visual recognition techniques and the layered soundscape elements as informed in [14] and [15]. In short, if soundscape could be decomposed into foreground and background sound elements, and visual environments should be decomposed under the same framework, then semantic segmentation could address background elements, while object detection serves foreground use.

metadata retrieved through visual processing are then restructured and projected as virtual sound sources.

2.2. Visual Recognition

With the taxonomic schema described in Figure 2, the project aims at retrieving both semantic and spatial information of virtual sound objects for a hand-crafted dataset of 160 high-dynamic-range (HDR) panoramic images. These panoramic images are processed so that it covers the entirety of the visible field of view, a dimensional constraint imposed by the physical structure and display resolution (15360×1200) of the CRAIVE-Lab's panoramic projection system. The visual recognition procedure starts by subdividing the panoramic image horizontally into multiple segments, so that it approximates the aspect ratio of image training datasets prepared for the visual recognition algorithms. To ensure the continuity of analytical performance at boundaries, these panoramic image segments were also given a data augmentation technique that involves a combination of mirror-padding [18] and periphery boundary extension.

The augmented image segments are then processed through the visual recognition system consisting of two pre-trained neural network algorithms: for semantic segmentation, Enet [16] is used with a 20-class subset of the Cityscape dataset [19], which contains urban visual scene objects such as buildings, vegetations, roads, and traffic lights; for object detection, the infamous You Only Look Once (YOLOv3) [17] model is used with the 80-class Microsoft Common Object in Context (COCO) dataset [20] that contains everyday objects such as bicycles, dogs, and clocks. Both neural network algorithms are known for their processing speed and high accuracy, which is beneficial when the visual input volume is significantly larger than non-panoptic visual scene data.

2.3. Audio Object Generation

The output of visual recognition algorithms for this project consists of spatial and symbolic meta-data used for the formation of

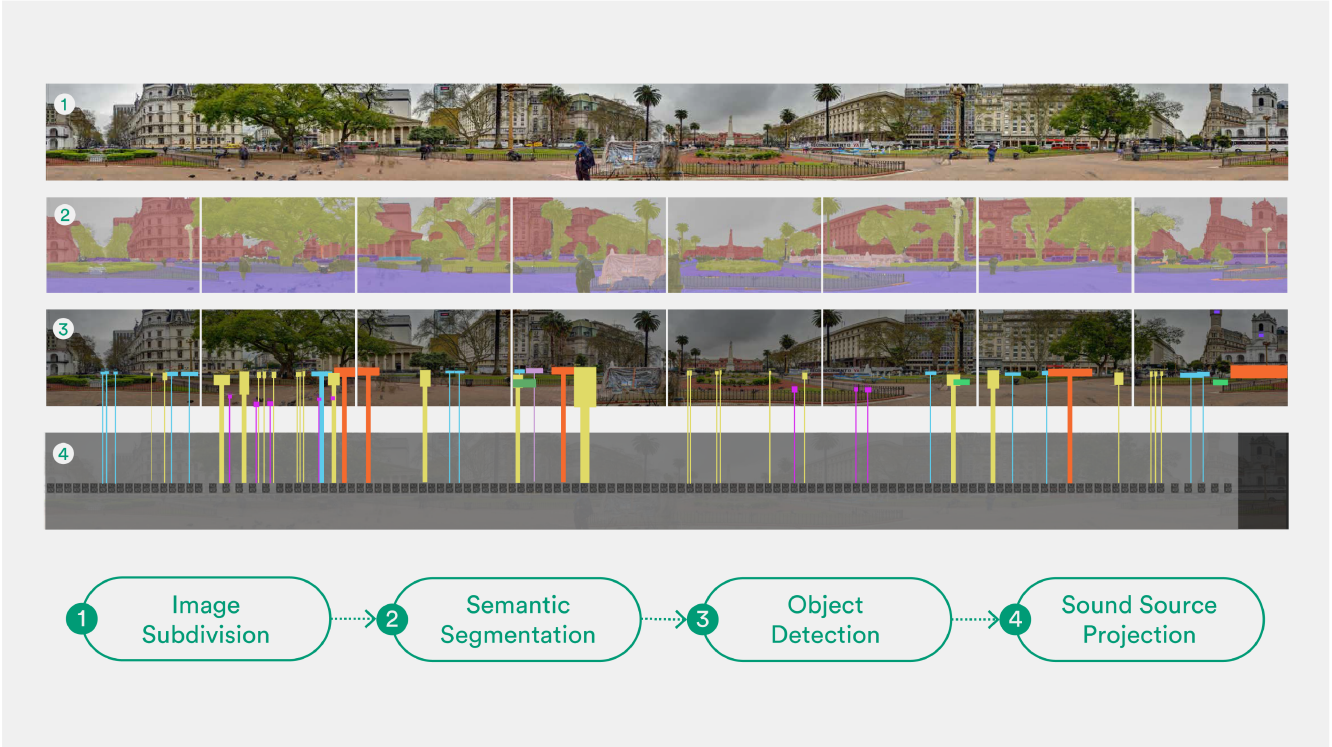

Figure 3: The overall framework of the system implemented for this project. The system subdivides panoramic images, and use pre-trained visual recognition algorithms, such as [16] and [17], to classify these images into spatially-situated semantic categories. The classification results are used to further extrapolate spatial metadata for audio object retrieval.

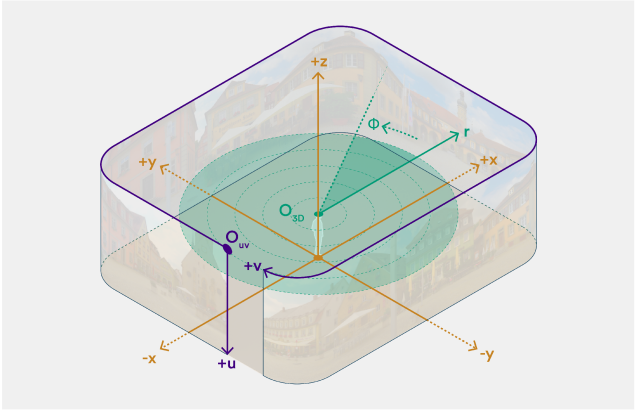

Figure 4: A 3-layered coordinate systems considered for the CRAIVE-Lab. The three layers are: 1) a Cartesian coordinate system that anchors at ground level at the center of the system; 2) an elevated radial coordinate system (as a horizontal cross-section of a spherical coordinate system) situated at human-height (also the height of the system's loudspeaker array); and 3) the surface coordinate that wraps around the entirety of the display system.

audio objects in spatial audio display. Specifically, we are interested in the relationship between the image coordinates (which is also the surface coordinates for the panoramic display system) and the radial coordinates for the spatial audio configuration (see Figure 4), as well as the scaling relationship between image resolution and physical dimensions of the display system. The spatial metadata consists of spatial positions of the foreground audio objects that are represented as bounding boxes, and ambient (background) audio objects that are represented as color-coded regions (again see Figure 3). The area and center positions of foreground sound objects are approximated to radial distances and azimuthal angles using a simple pinhole camera matrix. In addition, the area footprint of foreground sound objects, combined with their respective object classes, are also used to determine relative loudness during calibration. This is further enriched by the ambient (background) audio objects that are displayed using a spatially-distributive approach based upon classification regions given by the semantic segmentation, where the regions of background audio objects are converted to weighted angular ranges based upon vertical object hierarchy, thereby determining the prevalence of each audio object.

Within all 100 classes total from both training datasets, only 32 of them are audio object classes. For these audio object classes, an audio dataset is assembled by combining existing resources. To ensure variability and naturalness of sound effects, this dataset contains both synthetic and recorded audio. For the synthetic sounds, a multi-channel rendition of Farnell's synthesis [21] is reproduced, which could be used to directly populate virtual sound sources (to be discussed in Section 3). This is accompanied by recordings retrieved from open-source audio databases, along with other existing datasets used for audio feature detection and soundscape analysis, such as [22]. Due to the differences in propagation characteristics, each of these audio object classes are labeled with

a broadband calibration gain with respect to their functional components, which predetermines their absolute loudness when being displayed.

Upon receiving audio object meta-data, corresponding audio dataset elements are activated through virtual sound source generation, leading to a spatially-oriented playback within the CRAIVE-Lab. The timing of onsets for soundscape elements differs based upon corresponding functional soundscape components. Specifically, background elements are considered to be consistently present within the same visual context, in which case an amplitude-modulated ambient sound is displayed across all channels of loudspeakers, with respect to the their regions as informed by semantic segmentation. For foreground elements, all classes of sounds are displayed simutaneously, but with stochastic onsets for individual sounds of the same class.

3. IMPLEMENTATION

Figure 5: A visual representation of the calibration procedure for images in the panoramic dataset. *Top*: original panoramic image as output from the Image Composite Editor [23]; *Middle*: re-oriented image for the CRAIVE-Lab's display system, with correction for perspective distortion; *Bottom*: validation of appropriate positioning of horizontal perspective and distortion correction.

In practice, the soundscape reconstruction system is developed using a network of multiple platforms. A hand-crafted dataset of 160 panoramic images is constructed from HDR photography using the Image Composite Editor [23], which stitches image snippets based upon cylindrical projection. Due to the screen's nonuniform geometry (rectangular with rounded corners) a perspective transform must be applied to the images to counter introduced distortion. An important consequence of removing this distortion is ensuring the congruence of onscreen visuals and spatialized audio objects. Without the removal of the distortion, deviations between an original and transformed projection reach beyond 200 pixels [24] at certain points. This translates to over 0.5m of visual deviation on the screen, more than enough to disrupt the congruence of audio-visual presentation. Figure 5 shows an original input image and its corresponding transformed projection. The developed transformation utilizes matrices which define pixel coordinates of a spherical projection and the CRAIVE-Lab projection to interpolate the output from the input image [25]. This process can be applied to other screens of irregular geometry by adjusting the output matrix.

These corrected images are then used as input to a Python script encompassing all visual processing procedures, including the visual recognition algorithms. Once processing of an image is complete, the script outputs the classification and spatial position information for detected objects and scene segmentation analysis. The audio coding environment Max/MSP is used in conjunction with the IRCAM Spat 5 plugin to spatialize audio objects across the loudspeaker array [26]. This is achieved by first defining the relative positions of the loudspeaker array in virtual space. Audio sources are then generated according to the corresponding classifications, and their placement within the virtual space is determined using the position data. The contribution required of each virtual loudspeaker to create the soundfield is determined by the spatialization object. This is subsequently output to each virtual loudspeaker's analog in real space.

Due to the uniformity of audio objects presented in Spat 5, it becomes difficult to distinguish between each audio object classes, as well as their corresponding foreground/background classification. For this reason, a new visualization apparatus is needed for the system, which is proposed in Figure 6. In addition to virtual sound source positioning, this visualization apparatus situates all audio objects into foreground/background categories, and represent them using a variety of color codes to create distinction between audio object classes. In addition, sound intensity regions are also represented accordingly, with the ring radii of foreground audio objects representing relative sound intensity and decay characteristics, and the arcs' distances to room center representing the level of prevalence for ambient audio objects.

This workflow is implemented into a web-based application which provides a simple user interface for uploading content for analysis and display on the screen. Users experience no learning curve and require no training to display their imagery with system-generated soundscapes. Upon uploading, imagery is transformed and formatted for display, run through the visual recognition algorithms, and presented on the screen. The classification and position information is forwarded to the Max spatializer, which automatically generates and places the audio sources. This rapid-prototyping approach renders the CRAIVE-Lab a functional immersive virtual reality system usable by experts and non-experts alike.

4. SYSTEM PERFORMANCE RESULTS

The performance of the implemented system is evaluated with two interests. First, the efficacy of sound object retrieval under the constraints of training datasets used by the visual recognition algorithm is examined statistically. Second, an evaluation of computational performance is also conducted to determine whether there are potentials for real-time application using this approach.

4.1. Efficacy of Sound Object Retrieval

As discussed in Section 2, only 32% of the visual object classes are identified as audio objects classes. While a sizable proportion, the substantial presence of unused visual makes it crucial to analyze how effective the system's method is at generating sufficient amount of audio objects for a plausible rendering of the corresponding soundscape.

The first statistic to observe is the proportion of classified sound objects that are present among all objects in visual recognition. The result could be shown in Figure 7. Two observations could be made from this result. First, despite the fact that

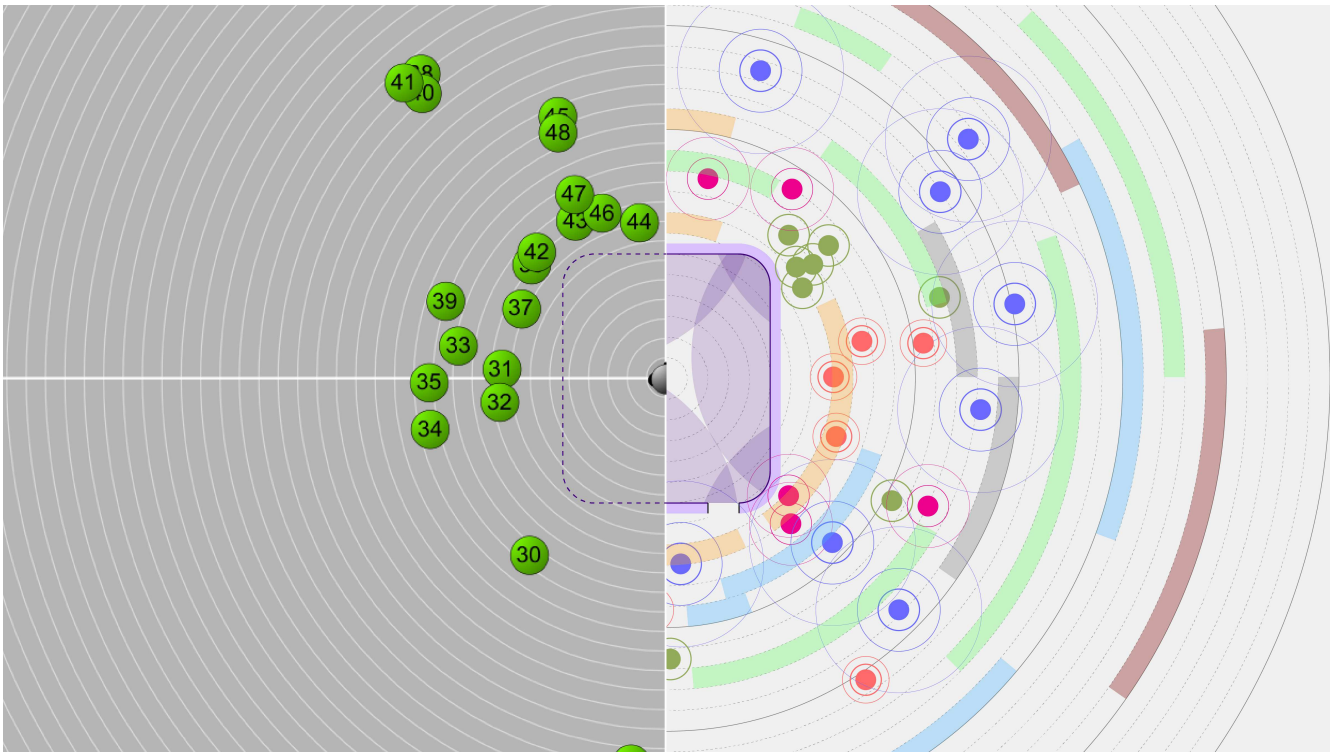

Figure 6: A typical example of the spatial audio mapping outcome for this system. *Left*: current configuration in IRCAM Spat [26] signifying source positions but without apparent visualization of foreground-background classification; *right*: visual interface design for the system that designates source positions, with color-coded foreground (dots) and background (arcs) sound object classes, corresponding to the activation range of the loudspeaker array at the CRAIVE-Lab (with the light purple ring representing the speaker array, while dark purple line representing the panoramic display). Interactive application for this interface is currently under development.

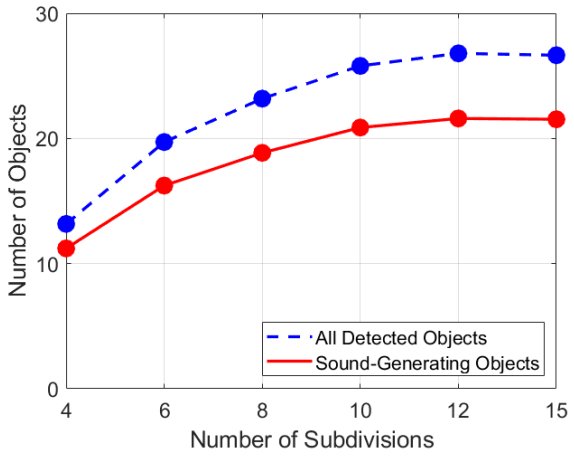

Figure 7: Performance of sound object retrieval dependent on the number of subdivided segments across the assembled panoramic image dataset. The blue dashed line and dots signifies the average amount of detected visual objects per panoramic image, while the red solid line and dots represents the number of sound-generating objects among them.

audio object classes are not the majority of all visual classes, it accounts for a substantial majority (above 80% across all subdivision schemes) of all recognized objects when analyzing instances of the panoramic image dataset. The consistency means that the efficacy of audio object retrieval is independent from the subdivision scheme imposed on the images. Second, the efficacy of visual recognition in general is dependent upon image subdivision. With the pixel resolution of 15360×1200 across all panoramic images, we have found that a subdivision scheme of 12 segments performs the best in both foreground audio object classification and object detection in general, with an aspect ratio of 16:15 (close to the 1:1 aspect ratio need to be enforced for the visual recognition models). This indicates that there is a point of optimization that allows for the most effective generation of virtual sound objects.

Due to the comprehensive nature of semantic segmentation, the resulting background sound object classification involves every class within the training dataset. This is not true for foreground sound object retrieval, in which not all classes are present in every object detection task. Therefore, the frequency of foreground sound object occurrence must be examined. Details could be seen in Figure 8. In general, *person* appears the most frequently as sound objects across the visual scenes depicted in the image dataset, affirming that human presence is largely independent from environmental contexts. This is followed by vehicles and small animals, which to an extent suggests that urban density contributes to the formation of foreground sound objects to a greater extent than

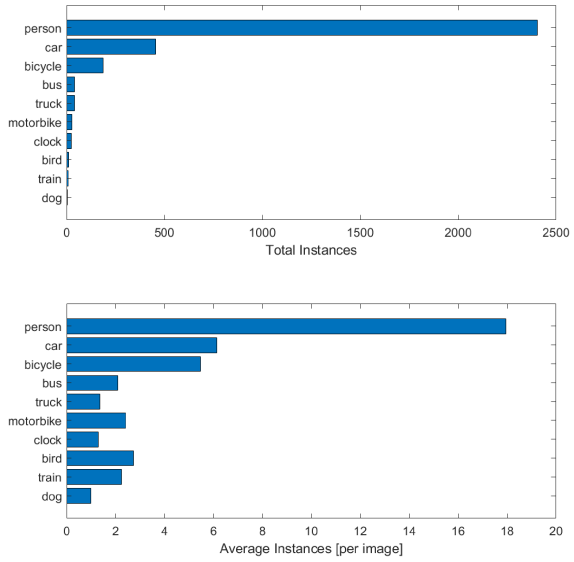

Figure 8: The top constituents of foreground sound objects by frequency of occurrence in the panoramic image dataset. *Top*: the total number of sound objects retrieved across all panoramic images; *Bottom*: average number of object instances when present.

non-urban environments. This further suggests that the contextual neutrality of audio object retrieval is completely dependent on the available visual object classes present, which could be inconclusive in representing all environmental contexts.

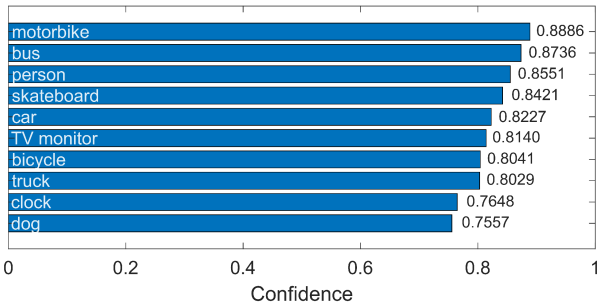

Figure 9: The top constituents of foreground audio objects by confidence of retrieval for each object, with a greater-than-chance confidence thresholding scheme.

The distortion of panoramic images also introduces a degree of uncertainty for the efficacy of this system. For this reason, the retrieval confidence for each sound objects are also analyzed (see Figure 9). The result shows that the subdivision approach neutralizes perceived distortion for the visual recognition models used, thereby influences little to their overall performance, with only minimal deviation from the ground truth confidence of 0.93 [17]. As a result, the approach remains highly effective in extracting spatial metadata for the foreground sound objects. In addition, by juxtaposing this result to the one shown in Figure 8, one could also observe their correlation, associating high confidence with the high frequency of occurrence for foreground sound objects.

4.2. Speed Performance

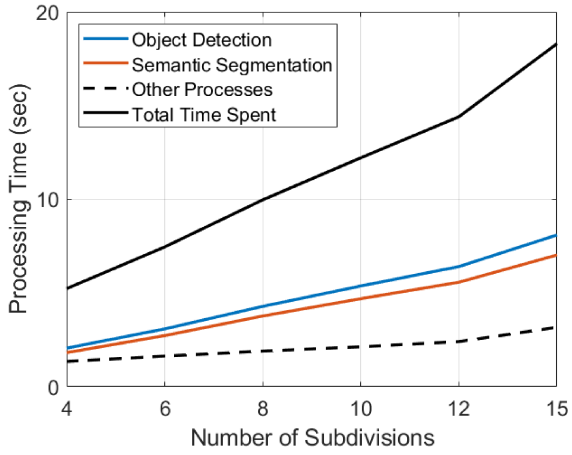

Figure 10: Processing time for the system dependent upon the subdivision scheme used for the panoramic image for the main hardware used for this project.

In addition to efficacy testing, the developed system for this project also undergoes performance testing to examine its speed, under a condition that substantially larger data volumes are present for both visual and audio processing with respect to the platforms used. Using a randomly-selected subset of the panoramic image dataset, the system is evaluated in terms of elapsed computation time using a Dell XPS 15 laptop workstation from 2017, running on a 4-core, 8-thread Intel Core i7 CPU, combined with a GeForce GTX 1050 GPU with 640 CUDA cores (one could say this is equivalent to the processing power of an entry-level gaming laptop). It should be pointed out that only the visual recognition algorithms utilize graphics processing power, while pre-processing and post-processing procedures in the data pipeline remain unparallelized.

The result (see Figure 10) shows a largely linear dependence between the subdivision scheme involved for panoramic image pre-processing, and an average of 1.5 seconds per image segment. Although this indicates the need of data parallelism for optimized performance, it suggests a potential for real-time use if better hardware is used (which would be the case for the serving laptop for CRAIVE-Lab's display system).

5. DISCUSSION

The analysis in Section 4 suggests that the system developed in this project is effective in reconstructing soundscape elements through direct analysis of visual scenes. In particular, it is efficient for the extraction of spatial meta-data in formulating structural representation of sound objects as virtual sound sources. The efficacy of spatial and semantic information retrieval is only contingent upon the limitations imposed by the visual recognition algorithms and

their associated datasets, which indicates that performance could be further improved with some computational optimization.

The most relevant advantage of this system is that it is designed with consideration of modularity. Specifically, although exchange of metadata occurs across platforms, the formation of sound objects in this system is independent from, and adaptive to, the continuing improvement of visual recognition apparatus. This makes it easy to enhance the system as an optimization process with faster rendering speed and increased robustness of audiovisual datasets.

While the system is designed to render soundscape without intentional environmental bias, it has become apparent that there are a number of limitations associated with the system. First, even though the plausibility of soundscape rendering could be achieved purely from the semiotic presence of each sound object [10], the aggregated sound object presented in this system does not meet with an ability to contextually filter out incoherent sound sources. This may contribute to an incongruent attention to the rendered sound field, especially with movement [27]. Second, while the method serves particularly well in the context of outdoor environments, where the peripheral acoustic conditions are largely uncontrolled, it does not take into account any room acoustics parameters, which results in inaccurate auditory experiences of indoor environments. While this system could serve as a good foundation for acoustic content generation for indoor spaces, any realistic spatial impression, such as reverberance, must be coupled with real-time auralization techniques [28] to be attained. This would most likely require accurate 3D reconstruction of virtual spaces and their respective acoustic simulation in the context of room-centered immersive systems, which must remain a separate research topic.

6. CONCLUSIONS AND FUTURE RESEARCH

In this project, a virtual soundscape reconstruction system is developed for room-centered immersive virtual reality systems such as the CRAIVE-Lab, in which virtual sound sources are projected and populated based upon spatial information retrieved with machine-learning-based visual recognition models. With optimization, this approach could facilitate realistic audiovisual rendering of visually-captured physical spaces, with potentials for sonically augmenting navigation of dynamic environments (e.g., 360° videos).

There are a number of future research directions that could be pursued. Among them, the perceptual accuracy of the generated soundscapes needs to be investigated. This is of particular interest because of its implications in achieving adequate place illusion and plausibility that could be experienced collectively [29], for which a system of quantitative evaluation has not been developed. One possible approach to such assessment include user studies through blind testing, in which only auditory cues (the reconstructed soundscapes) are presented without visual information, so that test subjects could determine the environmental context based solely upon listening. Such investigation can further incorporate human movement control within the CRAIVE-Lab by examining the soundscape across various local listening positions. System design for this research could also be further optimized in three aspects. Among them, parallelism could be employed to drastically reduce computational effort, leading to faster rendering and potential extension of soundscape reconstruction using panoramic video analysis. This could further incorporate sufficient consideration of reverberant conditions, so that it could also be effectively deployed for dynamic rendering of indoor environments. The audio dataset could be labelled contextually with geo-tagging, so that GPS metadata from images (such as EXIF) could be used to recognize and inherit more socio-culturally-oriented site-specific knowledge.

7. REFERENCES

[1] J. Braasch, Radke, R., B. Cutler, J. Goebel, and B. Chang, "Mri: Development of the collaborative-research augmented immersive virtual environment laboratory (CRAIVE-Lab)," 2012–2015, NSF #1229391.

[2] J. P. Carter and J. Braasch, "Cross-modal soundscape mapping: Integrating ambisonic environmental audio recordings and high dynamic range spherical panoramic photography," in *The 20th International Conference in Auditory Display (ICAD-2014)*, June 2014.

[3] M. Morgan, "Automatic classification and immersive representation of environmental soundscape," Master's thesis, Rensselaer Polytechnic Institute, July 2018.

[4] M. Southworth, "The sonic environment of cities," *Environment and Behavior*, vol. 1, pp. 49 – 70, 1967.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR09*, 2009.

[6] D. E. Tuomas Virtanen, Mark D. Plumbley, *Computational Analysis of Sound Scenes and Events*. Cham, Switzerland: Springer Nature, 2018.

[7] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *Proceedings of European Convention for Computer Vision*, 2018.

[9] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360-degree video," in *Neural Information Processing Systems (NIPS)*, 2018.

[10] U. Jekosch, "Assigning meanings to sound: Semiotics in the context of product-sound design," in *Communication Acoustics*, J. Blauert, Ed. Berlin and Heidelberg: Springer-Verlag, 2005, ch. 8, pp. 193–221.

[11] A. Bregman, *Auditory Scene Analysis: the perceptual organization of sound*. Cambridge, Massachusetts: MIT Press, 1990.

[12] J. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nat Rev Neurosci 14, 693707*, vol. 14, no. 693-707, 2013.

[13] P. Schaeffer, *Treatise on Musical Objects (Trait des objets musicaux)*. Oakland, California, USA: University of California Press, 2017 (1966).

[14] ISO, "Iso 12913-1:2014: Acoustics soundscape part 1: Definition and conceptual framework," https://www.iso.org/standard/52161.html, 2014.

[15] R. M. Schafer, *The Soundscape: our sonic environment and the tuning of the world*. Rochester, Vermont, USA: Destiny Book, 1977.

[16] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 2016.

[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft coco: Common objects in context," 2014.

[21] A. Farnell, *Designing Sound*. Cambridge, Massachusetts: MIT Press, 2010.

[22] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[23] Microsoft, "Image composite editor," 2015, [Online] Available https://www.microsoft.com/en-us/research/product/computational-photography-applications/image-composite-editor/ (accessed June 3, 2020).

[24] J. P. Carter, "Immersion: A framework for architectural research," Ph.D. dissertation, Rensselaer Polytechnic Institute, 2019.

[25] S. Chabot and J. Braasch, "Interactive framework to control and rapid-prototype for collaborative immersive environments," *J. Audio Eng. Soc*, 2021, [submitted].

[26] T. Carpentier, "A new implementation of Spat in Max," in *15th Sound and Music Computing Conference (SMC2018)*, Limassol, Cyprus, July 2018, pp. 184 – 191. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02094499

[27] B. Cohen-Lhyver, S. Argentieri, and B. Gas, *Audition as a Trigger of Head Movements*. Cham: Springer International Publishing, 2020, pp. 697–731. [Online]. Available: https://doi.org/10.1007/978-3-030-00386-9_23

[28] L. Savioja and U. Svensson, "Overview of geometrical room acoustic modeling techniques." *The Journal of the Acoustical Society of America*, vol. 138 2, pp. 708–30, 2015.

[29] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments." The Royal Society, 2009, pp. 3549–3557.