Technical Disclosure Commons

Defensive Publications Series

July 2022

Enriching Audio Signals in Augmented Reality with Spatial Attributes of Sound

Corville Allen

Andy Lavery

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Allen, Corville and Lavery, Andy, "Enriching Audio Signals in Augmented Reality with Spatial Attributes of Sound", Technical Disclosure Commons, (July 05, 2022)
https://www.tdcommons.org/dpubs_series/5240

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Enriching Audio Signals in Augmented Reality with Spatial Attributes of Sound

ABSTRACT

This disclosure describes augmented reality (AR) techniques that spatially track audio signals to enable AR interactivity based on audio and on the relationship of the objects that the audio is attached to. An AR object responds interactively based on the properties and the classification of spatially-tracked audio as well as the intent and sentiment inferred from the audio. Spatially separated microphones are leveraged to localize sounds to objects. Alternatively, video object detection is combined with audio processing to localize sounds to objects. AR objects are tagged with spatial audio classification information as they move such that audio information of the object can be used for natural interactions in the scene.

KEYWORDS

- Augmented reality (AR)
- Virtual sound
- Virtual reality (VR)
- Immersive experience
- Audio localization
- Sound classification
- Microphone array
- Direction finding
- Audio ranging
- Object detection
- Audio tagging

BACKGROUND

Augmented reality (AR) incorporates virtual objects into the visual display of objects in the real world, typically through a single viewport attached or associated with a physical viewing instrument such as AR glasses, headset, etc. AR works by maintaining a visual model of a physical environment onto which virtual objects are layered. Currently, AR technologies rarely incorporate sound as a way to interact with the AR objects in a realistic way. This results in a

user experience that is less engaging than would be possible with accurate sound interactions. Further, spatial attributes of audio signals are not currently taken into account in AR. Current AR techniques can play virtual sounds or modify environmental sounds, e.g., cancel out certain types of sounds or noises. However, this does not constitute user interaction with audio.

Audio signals sensed by multiple microphones can be triangulated to determine the spatial location of the source of the audio. Sound can also be localized by analyzing the audio feed that accompanies video, e.g., by correlating audio with objects found in the video. Sound classification can also be done through machine learning, e.g., by associating the sound of a video with objects detected within the video. Speech-to-text (STT) techniques can be used to classify speech and detect intent.

DESCRIPTION

This disclosure describes augmented reality (AR) techniques that spatially track audio signals to enable AR interactivity based on audio and on relationships between the objects within the AR environment that the audio signals are attached to. User interactivity with spatially tracked audio is a modality of interaction that is additional to the traditional visual modality of interaction. An AR object can interactively respond to the properties and classification of spatially-tracked audio as well as the intent and sentiment inferred from the audio. The response of an AR object to audio can include a set of predetermined activities based on sounds and the relationships of the object to other objects. To determine the activity for a given sound, audio can be received and spatially tracked from multiple perspectives, e.g., the location of the AR object (either derived or actual), the user location, etc.

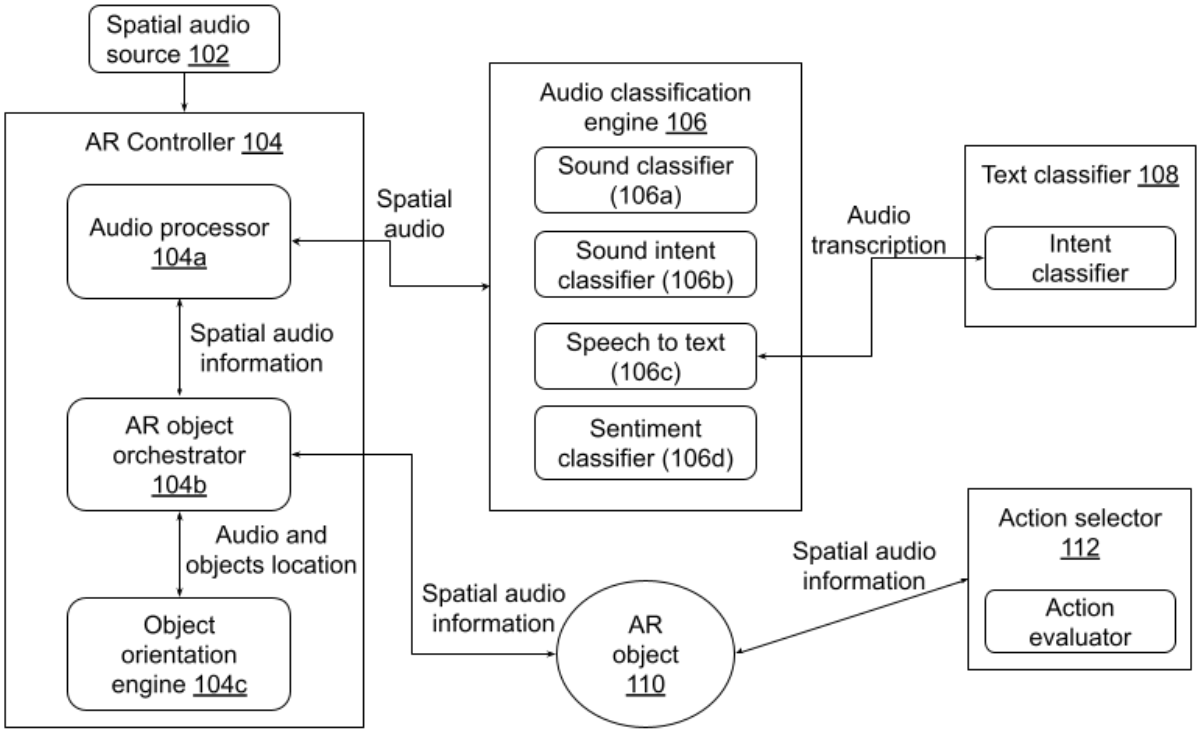


Fig. 1: Enriching audio signals in augmented reality with spatial attributes of sound

Fig. 1 illustrates the enrichment of audio signals in augmented reality with spatial attributes of sound. A source (102) emits audio with spatial properties. An AR controller (104) receives the audio. The AR controller includes an audio processor (104a), an AR object orchestrator (104b), and an object orientation engine (104c).

Spatial properties of the audio including the direction and position of the source are determined by the audio processor and passed on to the AR object orchestrator, which in turn passes it on to the AR object (110). The behavior of the AR object depends, among other things, on the spatial properties of audio. The audio and location of the object is also sent to the object orientation engine, which orients the object based on the relative positions of the audio source and the object. The orientation engine selects the perspective to use for spatial audio data. An example perspective can be the perspective of the user or of the microphone. Another example perspective can be a perspective derived by adjusting the distance of the AR object from the

microphone and by adjusting the directional aspects for the spatial properties. Here, direction can refer to the vector between the microphone and the point of reception of the sound. The direction can be correlated with the location of the AR object in two or three-dimensional space and rotated accordingly.

The spatial location of the audio source can be determined using, for example, by one or more of the following.

- Leveraging the directionality of microphones on a single device such as a mobile phone, an AR headset, or other device used to view the AR scene.
- Using individual directional microphones placed in a scene and feeding a central control unit that determines the location of the source.
- Using a directional microphone array that can localize sound to multiple spatial locations within a scene.

Audio from the source is also sent to an audio classification engine (106), which can include a sound classifier (106a), a sound-intent classifier (106b), a speech-to-text engine (106c), a sentiment classifier (106d), etc. Properties of the audio, including amplitude, wavelength, frequency, period, velocity, speech intent (using a speech classifier), sound intent (using a sound classifier), sound classification (e.g., car, truck, airplane, animal, human speech, coffee pot whistling, door closing, ambient sound, etc.), text transcription of speech, sentiment of speech, sentiment (cheerful, calming, energetic, dangerous, etc.) of musical sounds, etc., are fed to the object.

Classification of audio can take place in predetermined snippets (e.g., 2, 4, 5 seconds). For sound classified as human speech, intent can be determined (with user permission) using a text classifier (108) by determining utterance length, e.g., up to ten seconds; by transcribing

speech to text up to the pause of utterance; by categorizing the text into intents; etc. For non-speech sounds, data structures are built representing the classification of sound and of the producer, e.g., car, train, animal, airplane, whistle, music, water running, etc. Intent and sound properties (amplitude, frequency, wavelength) can be used to determine a sentiment. Sound classification and spatial audio properties are encapsulated in data structures to send to the AR object.

The spatial and other properties of the audio, including perspective, intent, sentiment, sound classification, etc., are sent to the AR object. The response of AR objects to audio streams and their properties can vary across objects, depending on the type of object, the relationship of the object to other objects in the viewport space, the classification attributes of the audio, the spatial characteristics of the audio, the spatial characteristics of objects in the viewport space, etc.

The response of an AR object to audio is determined by an action selector module (112), which can include an action evaluator. The response of an AR object to audio can be pre-programmed or based on dynamic lookup of key-action types. For example, the AR object can use the spatial and other properties of audio information to perform a simple acknowledgement. As another example, the AR object can use spatial and other properties of audio to perform a relatively complex response such as selecting the closest matched intent, sentiment, or sound classification from its list of actions. Sound capture and the resulting AR actions are looped over predetermined time frames.

Examples

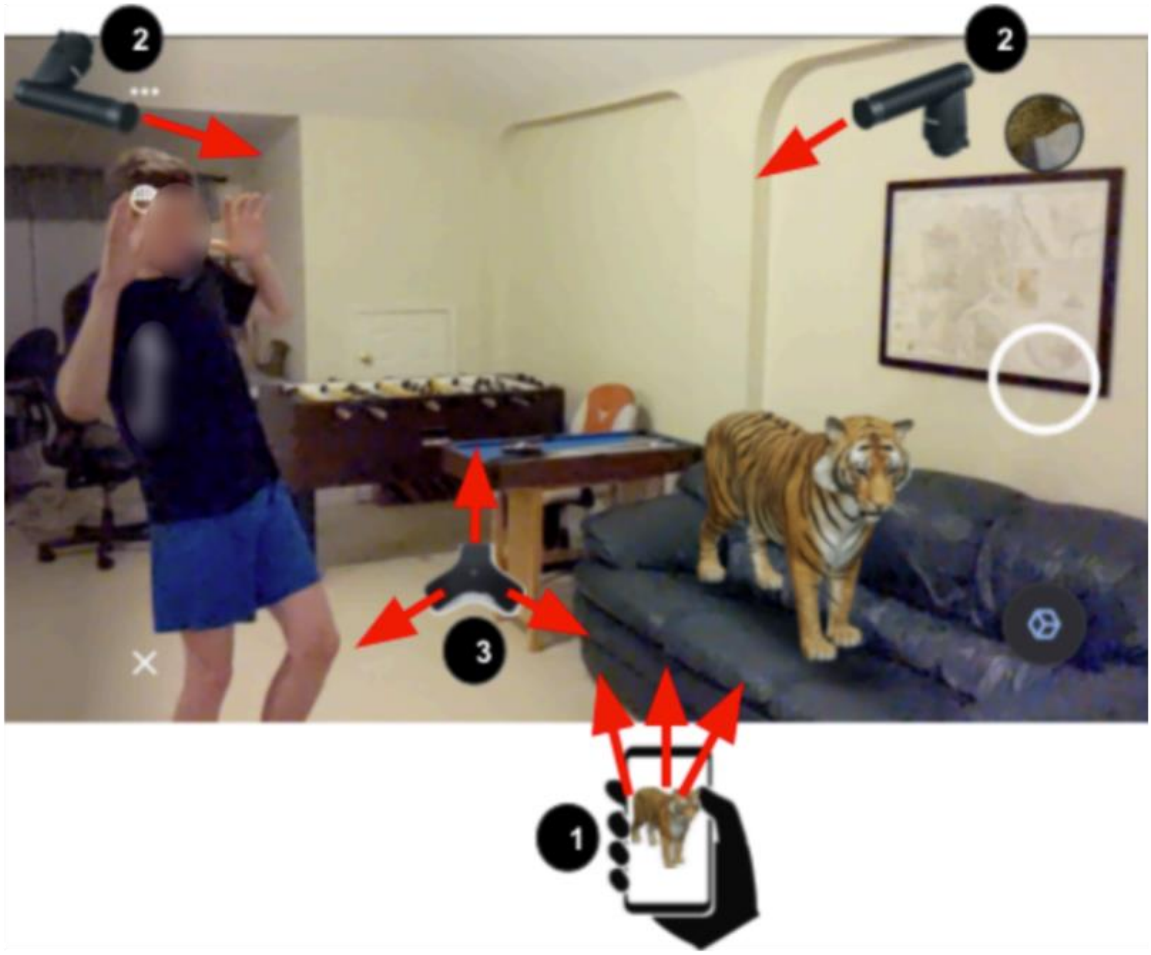


Fig. 2: Response of an AR object can depend on the spatial properties of audio

Fig. 2 illustrates an example where the response of an AR object is dependent on the spatial properties of audio. In this example, a virtual tiger stands on the sofa of a living room, while a real (physical) person gesticulates and speaks to it. The spatial location of the audio source (the person) is determined by analyzing audio streams collected from microphones that can be, e.g., embedded within a smartphone ❶; directional microphones located in various parts of the scene ❷; a microphone array that can detect audio in multiple spatial areas ❸; etc.

If the person yells “go away” at the tiger, the tiger object can receive information on the audio signal classified in many ways such as amplitude (90 dB, loud); duration of audio (one

second); sound classification (speech); text of speech (“go away”); speech intent (move away); sentiment of speech (aggressive, angry); etc. The virtual tiger can react to the audio in pre-programmed ways based on its properties, e.g., it can turn its head in the direction of the person and roar at the person; it can lie down; it can take a step back; etc.

An AR object can have a simplified response depending on the designed level of interactivity and pre-programed based on the sound. In the example of Fig. 2, simplified responses can be, e.g., acknowledging the sound by animated focusing; changing color; displaying a text response; etc.

Example: A virtual assistant provides information by reacting to a sound

A loud cheer erupts and its sound is classified along dimensions such as amplitude (93 dB, loud); location (multiple /all around); duration (3 seconds); sound classification (people); text of spoken language (inaudible); sentiment of spoken language (not applicable); speech intent (not applicable); sound intent (stadium celebration); etc. The AR object, a virtual assistant, utilizes audio properties to prompt the user to look up for a score, while turning to the left and changing its expression to a smile.

Example: An inanimate AR object provides information by reacting to a sound

A loudspeaker announces “touchdown by the home team," and the sound is classified along dimensions such as amplitude (50 dB, medium); location (left of user); duration (2 seconds); sound classification (human speech); text of speech (“touchdown by the home team”); speech intent (game score); sound intent (not applicable); speech sentiment (happy); etc. An inanimate AR object, e.g., a companion display, looks up the current score for any home-team game, and displays the current score with a happy face.

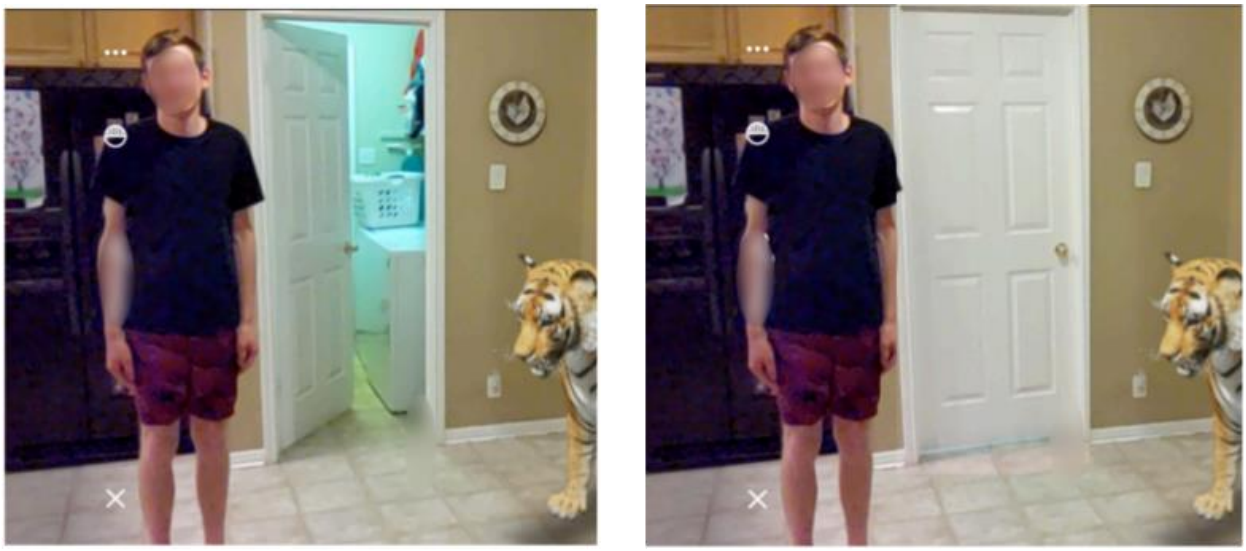
The described techniques apply to a wide variety of AR settings, including events, meetings, logistics warehouses, virtual companions, advertisements, maps, navigation, augmented live navigation, etc. The techniques generally enable the user to participate more fully in augmented reality with audio controls.

Determining the location of moving audio sources by correlating emitted audio with associated objects in an accompanying video

Multiple, spatially separated microphones may not always be available for determination of the location of an audio source. In what follows, techniques are described to determine the spatial location of multiple audio sources within an AR scene by correlating their emitted audio with associated objects in an accompanying video. The techniques are usable even if the audio sources move. The techniques enable tagging of audio streams to objects found in the video feed of the AR scene, such that users can interact with objects in an AR viewport in a natural way.

A three-dimensional model of the (virtual and real) objects in a dynamic AR scene is built, such that as a user moves around, they can see the scene from different vantage points. Audio received from the environment is run through a classifier and video of the environment is run through an object detector to identify objects in the scene. Distances and locations of objects within the AR scene are estimated. Audio is associated with an object in the scene. The location of the object is adjusted to provide spatial information for enriched audio. The object is tagged with the sound. The enriched spatial information is sent to the AR object for a better experience.

Example



(a) **(b)**

Fig. 3: An AR scene with (a) an open door (b) a shut door

In the AR scene example of Fig. 3, a door is initially open (Fig. 3a) and then slams shut (Fig. 3b), causing a sound. The AR device monitors the video and audio of the scene. Various objects, including person, door, (virtual) tiger, etc., are detected. The change between video frames (open door versus shut door) is used to determine the source of the slamming sound, e.g., the door or the general area of the door. The slamming door audio signal is tagged with the door in the scene. The audio is enriched with spatial characteristics relative to the scene. The location for the objects in the scene can be determined, and the spatial information (localized) sent to the AR objects for interactions.

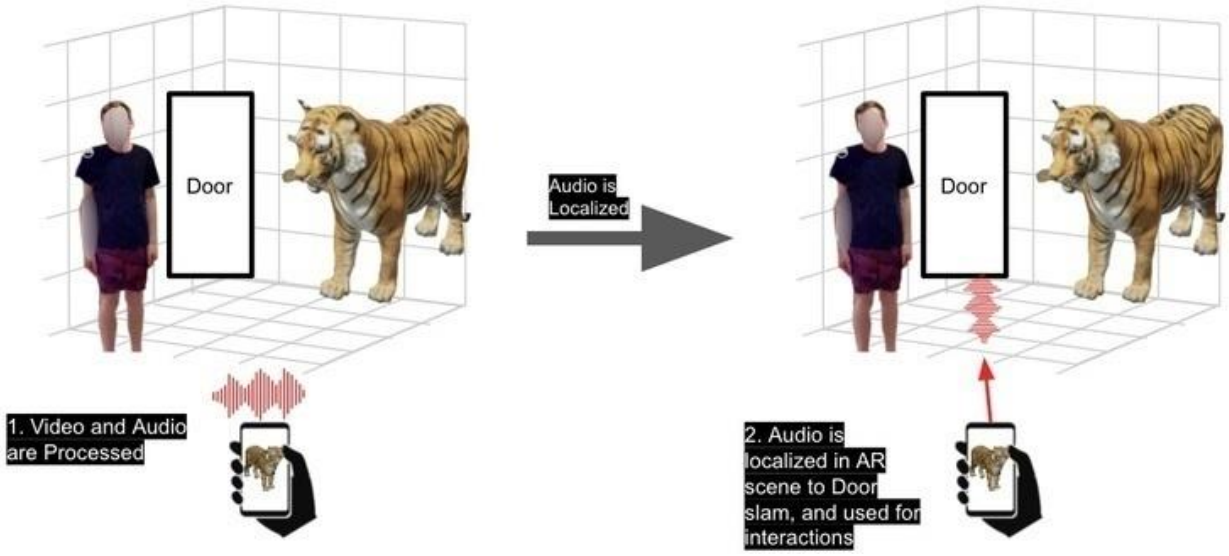


Fig. 4: Associating an object detected in a video with a sound

Fig. 4 illustrates in greater detail a process for associating an object detected in a video with a sound. The video and audio from an AR scene are processed ❶ to identify a three-dimensional location for the sound source. The location is mapped to an object that created the sound at that location. Audio is localized in the AR scene, and the mapping is used for user interactions with AR objects ❷. The enriched audio is classified and provided to AR objects in the scene for them to respond to the enriched audio information for the scene. Enriched audio can include information about the audio such as sound properties (amplitude, wavelength, frequency, period, velocity, text/sentiment of spoken language, etc.); base location information; spatial information; spatial information relative to AR object; etc.

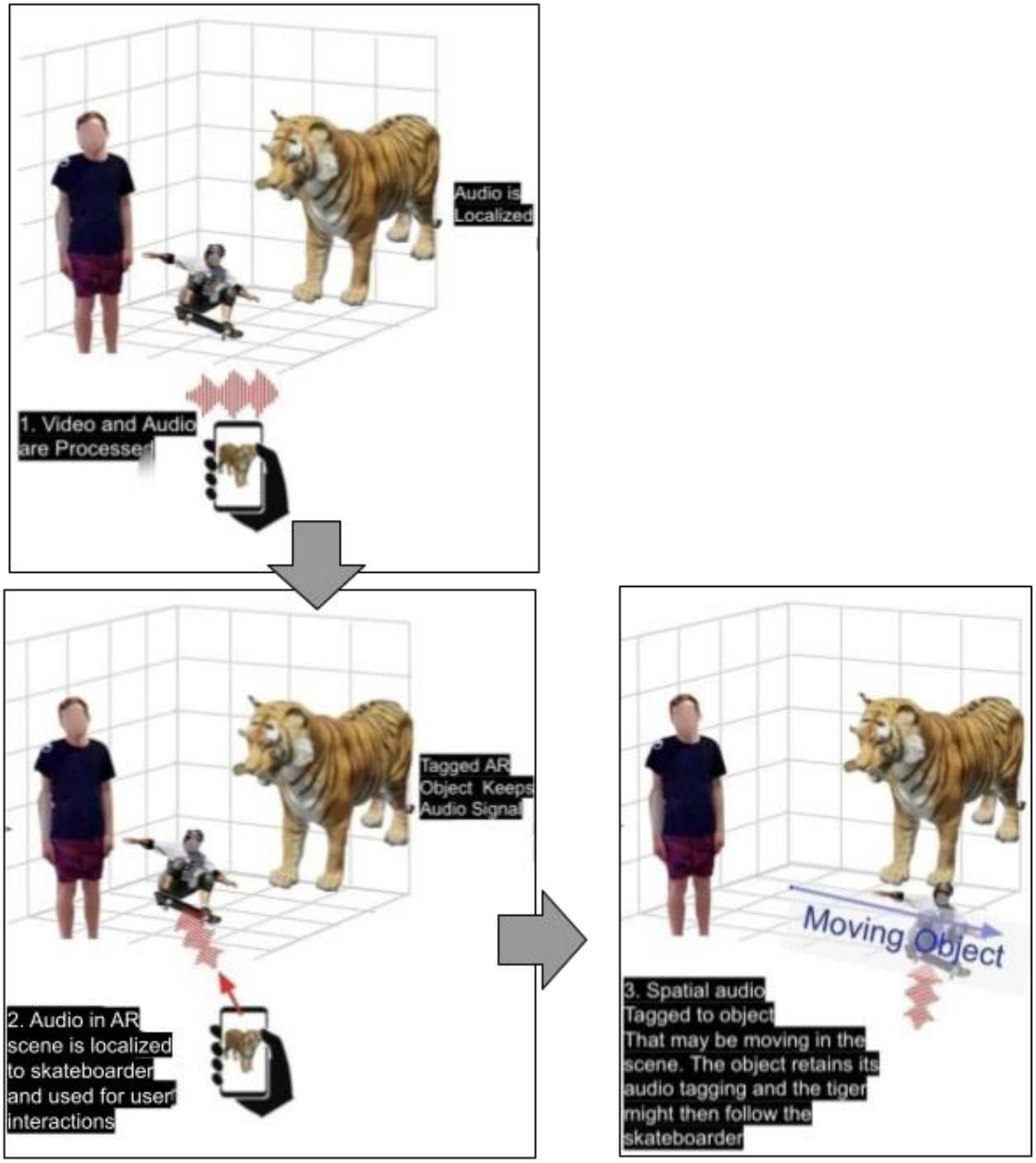


Fig. 5: Audio tagged to a moving object remains tagged even after the object exits the viewport

Illustrated in Fig. 5, the object tagged with spatial audio and additional classification attributes can be a moving (dynamic) object (a skateboarder, in the example of Fig. 5). The spatial audio tagging remains tagged to the object even as it moves within or exits the AR scene.

In addition to tagging the object with its audio, periodic updates are sent of the object and whether its audio has stopped or is continuing. In the example of Fig. 5, the skateboarder crashes into the scene ❶, makes a loud sound, and continues moving through the scene after the crash. Having tagged the crashing sound with the skateboarder ❷, the tiger can track the skateboarder even after the crash ❸.

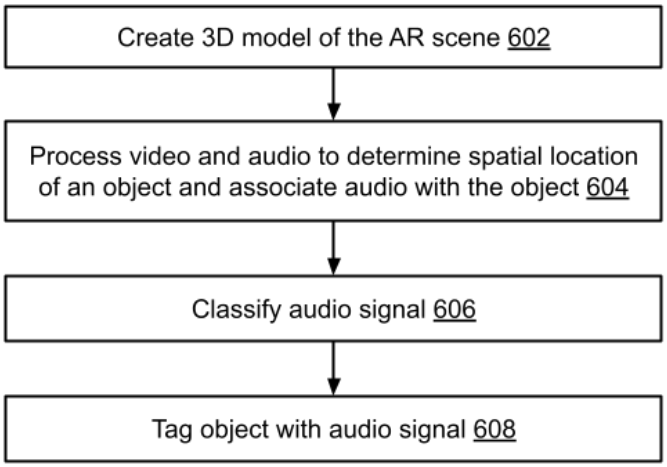


Fig. 6: Enriching audio signals in augmented reality with spatial attributes of sound determined by associating audio with objects detected in video

Fig. 6 illustrates an example method for enriching audio signals in augmented reality with spatial attributes of sound determined by associating audio with objects detected in video. A three-dimensional (3D) model of the AR scene is created to track physical and virtual objects within the scene (602). The AR scene, including its video and audio feed, is recorded or monitored. The video (or frame sequences) of the scene and the accompanying audio signal is processed to determine the spatial location and object association of the audio in that scene (604), as follows.

Audio from the environment is run through a classifier, and video of the scene is analyzed to identify objects in the scene. Distances and locations of physical and virtual objects in the scene are estimated. Audio is associated with an object in the scene. The location of the

object is adjusted to become the spatial information for an enriched audio. The audio signal is classified (606). The object in the 3D model is tagged with the observed audio signal (608). The objects in the scene are provided with classification information along with the object that was identified as the source of the audio signal.

Objects in the AR scene can react to newly generated sound and can base their reaction on the dynamic spatial location of objects in the scene. Since the object that created the sound is tracked, even if the object moves, other AR objects in the scene can respond appropriately based on that movement. The procedure of Fig. 6 can be run in a loop to track changes in the AR scene and in the audio, so that the objects in the scene can continue to respond to spatial audio signals.

In this manner, the described techniques provide enriched interactivity with an AR scene through sound. Spatially separated microphones, if present in the AR scene, are leveraged to localize sounds to objects. In the absence of sound-locating spatial microphones, video object detection is combined with audio processing to localize sounds to objects. AR objects can be tagged with spatial audio classification information even as they move, so that audio information of the object can be used for natural interactions in the scene.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein may enable the collection of user information (e.g., information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location

information is obtained (such as to a city, ZIP code, or state level) so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes augmented reality (AR) techniques that spatially track audio signals to enable AR interactivity based on audio and on the relationship of the objects that the audio is attached to. An AR object responds interactively based on the properties and the classification of spatially-tracked audio as well as the intent and sentiment inferred from the audio. Spatially separated microphones are leveraged to localize sounds to objects. Alternatively, video object detection is combined with audio processing to localize sounds to objects. AR objects are tagged with spatial audio classification information as they move such that audio information of the object can be used for natural interactions in the scene.

REFERENCES

[1] “Acoustic location - Wikipedia” available online at https://en.wikipedia.org/wiki/Acoustic_location, accessed 16 Jun 2022.

[2] Vahedian Mazloum, Abedin. “Identification of sound source in machine vision sequences using audio information.” In *4th EURASIP conference focused on Video/Image and Multimedia Communications*. 2003.

[3] Senocak, Arda, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. "Learning to localize sound source in visual scenes." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358-4366. 2018.