Towards a Systematic Scene Analysis Framework for Audiovisual Data Representations

Elias Elmquist
Linköping University
Sweden
elias.elmquist@liu.se

Alexander Bock
Linköping University
Sweden

Anders Ynnerman
Linköping University
Sweden

Niklas Rönnberg
Linköping University
Sweden

CCS Concepts

•**Human-centered computing** → **HCI theory, concepts and models;**

Author Keywords

Audiovisual integration, Visualization, Sonification, Scene graph, Scene analysis.

1. INTRODUCTION

The way humans make sense of the world is largely due to the ability to perform scene analysis, a cognitive process that organizes the sensory inputs that are perceived [9]. The auditory system enables the analysis of a mixture of sounds to identify coherent perceptual objects corresponding to individual sound sources, a process known as auditory scene analysis [4]. Similarly, sensory input from the visual system is used to identify visual objects from the background through edge detection and figure-ground separation [13], enabling, for example, the identification of an animal from a dense forest landscape.

The scene analysis process results in a perceptual organization of sensory inputs consisting of several levels. In computer vision [11] and computer graphics [14], the components and levels of a virtual scene are often illustrated as a scene graph. The scene graph has a tree structure, where the root node represents the entire scene, which is divided into the objects of the scene, where the objects themselves are also divided into their components. The scene graph in this context is mainly used as a data structure to arrange the logical and spatial components of the scene. 1 shows an example of a scene graph represented with images, and also a scene graph represented with words. Similarly, auditory scene analysis utilizes a hierarchical description of a scene, which segregates parts of a scene at one level, and merges into a single unit in another [4]. However, no illustration systematically lists the different levels of organization and listening.

2nd Audiovisual Symposium 2024, 6 December, Falun, Sweden.

Considering the cognitive integration of auditory and visual scene analysis, and the resulting perceptual organization, this paper proposes a systematic conceptual framework for displaying the organization of an audiovisual representation, named Audiovisual Scene Graph. The framework illustrates and highlights how visual and auditory elements merge to become higher-level objects, and how they can also integrate to become audiovisual objects. Furthermore, the framework has the potential to assist in the design of an audiovisual representation, as well as in analyzing existing representations to better understand how they are perceived.

2. AUDIOVISUAL SCENE GRAPH

The Audiovisual Scene Graph takes inspiration from scene graphs in computer vision and graphics, and applies them to the cognitive process of scene analysis to create a systematic framework of perceptual organization. Similarly to how a parse tree displays the syntactic structure of a sentence, the Audiovisual Scene Graph displays how an audiovisual scene can be parsed. The scene graph can be read from top to bottom to deconstruct a scene into objects and marks, or from bottom to top to construct it. To set up an audiovisual scene framework, it is necessary that the levels it contains apply to both the visual and auditory modalities, while also working for each modality independently.

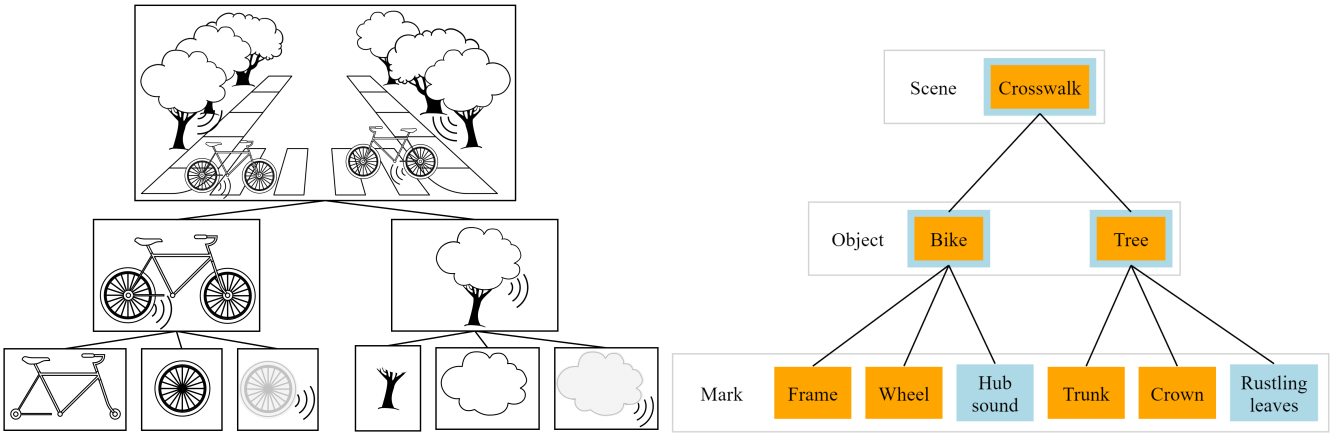
2.1 Scene

The top level of the Audiovisual Scene Graph corresponds to the perceived audiovisual scene. It is a conceptual spatiotemporal frame where objects can be placed, and gives a general description of how these are perceived from a holistic perspective. It is therefore the combined landscape and soundscape of an audiovisual representation, where it is experienced as a single composition. As shown in 1, the top level represents the scene where all objects are visible. See also the first half of the linked video for a demonstration of the images and sounds[1]. Some parts of the scene do not necessarily have to be deconstructed into objects but might instead be part of the general scenery, which in this case is the sidewalk.

2.2 Object

Objects are the entities or things that populate the scene. Although the word object is mainly associated with the visual perspective, there is research that points to the defini-

[1]Link to video: `https://youtu.be/tMykj3hySEU?si=5kSHqnjk8HRAofnz`



(a) Scene graph represented with images. The scene consists of the objects of a bike and a tree, which are instantiated into the scene.

(b) Scene graph represented with words. Orange color represents the visual modality, and blue represents the auditory modality.

Figure 1: The Audiovisual Scene Graph, shown through two different approaches.

tion of auditory objects, which extends the more abstract notion of an auditory stream [7, 10, 2], a term more commonly used in auditory scene analysis [4]. An object comprises other components, just as a bicycle consists of two wheels and a bike frame (as shown in 1). An auditory object is the composition of sounds that originates from the same auditory source, such as the combination of the sound of a bike wheel rolling across the street, and the clicking sound from the wheel hub. An object can be added multiple times within a scene. This is comparable to inheritance in object-oriented programming, where an object can be instantiated based on the properties of another object. This concept is similar to perceptual symbol systems [1], where a simulator is the class or template of an object, and simulations are the instances of the simulator. In the example of 1a, the tree object has been instantiated six times in the scene, and includes variations of rotations and tree crown structure.

2.3 Mark

The mark is the foundational element and primitive of the scene, serving as the building block of objects, which themselves can be described as multiple marks within the same region [12]. Regarding 1, marks are the parts that make up a bicycle (the wheels and the frame) and the individual sounds they make. They can be seen as the undividable parts of the scene, or at least within the context of the scene analysis itself. Theoretically, a scene can be divided into as many levels as possible. For example, a bike wheel can be seen as an object consisting of spokes, a hub, and a tire, and the bike hub sound can be divided into the individual clicks that create the auditory object. Restricting the Audiovisual Scene Graph to three levels of organization (scene, object, mark) allows for a holistic overview of the scene and focuses on its most essential components.

3. DATA PERCEPTUALIZATION SCENE GRAPH

In the context of data perceptualization, the framework can be extended to include data mappings. The terminology of this aspect of the framework is based on the theory of visual marks and channels [3], which is also used in more recent visualization models [12]. Furthermore, these visual definitions have been adopted to also function in the auditory domain [6]. Through these definitions, the Audiovisual Scene Graph can be used for visual and aural representations of data (visualization and sonification respectively). This version of the Audiovisual Scene Graph is demonstrated by extending the scene found in 1 to include data mappings through the marks (see 2). See the second half of the previously linked video for a demonstration of the images and sounds.

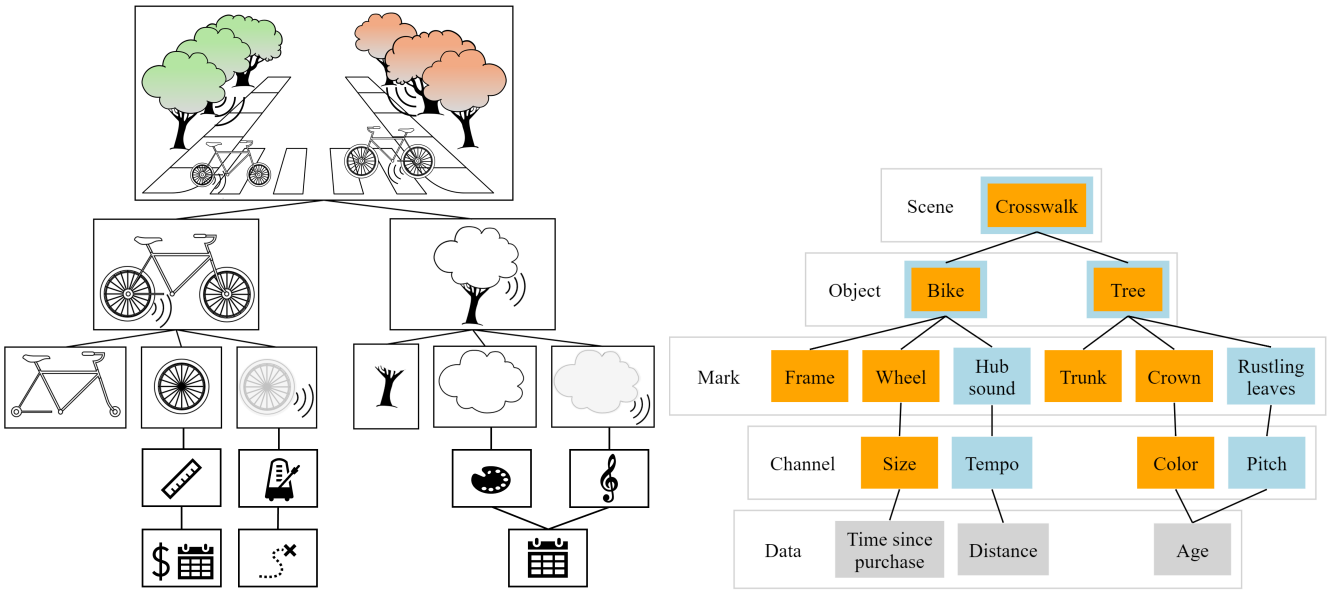
3.1 Channel

The manifestation of a mark is controlled by a set of channels, which can be parameterized to convey information. Visual channels can include size, color hue, and shape, while auditory channels can include pitch, loudness, and timbre. In 2, the size of the wheels of the bike is changed according to the data, as well as the tempo of the bike hub sound. For the tree object, the color of its crown and the pitch of the sound of the rustling leaves is changed according to the data. This is done separately for each object to reflect how each data object contains different values. All mappings have a positive polarity, where increasing data value corresponds to an increasing channel value.

3.2 Data

Data is the variable that is encoded through a channel of a mark. Furthermore, an object in the scene can represent a data object which contains several attributes. In this imaginary example, the data variables are 'time since purchase' and 'distance traveled' for the bike, and the age of the tree. In 2a, the bike to the left has a smaller wheel size than the bike to the right, meaning that the bike to the left was more recently purchased. By comparing the sounds, the bike to the left has a higher tempo of the sound, meaning that it has traveled a longer distance. Through the color of the tree crowns and the pitch of the rustling leaves, it can be seen and heard that the trees to the left are younger compared to the trees to the right. Mapping the same data to a visual and auditory mapping is known as a redundant mapping, and can increase the comprehension of a data representation.

4. REFLECTIONS AND FUTURE WORK



(a) Scene graph represented with images, with data mappings.

(b) Scene graph represented with words, with data mappings.

Figure 2: Extended version of the Audiovisual Scene Graph to present data mappings.

Sonification literature mentions that considering auditory scene analysis is important when creating effective sonifications [8]. Applying the Audiovisual Scene Graph to sonification design would promote the consideration of how data mappings are organized into different sounds, and how these sounds are organized into perceptual objects and a unified scene by the listener. Providing a scene analysis framework that displays the intended way to listen and parse a sonification could reduce unwanted ambiguity in the data display. Conversely, the framework can also be used on more artistic audiovisual representations, where the intent of multiple meanings and interpretations of an audiovisual representation can be investigated. Future work aims to further define and validate the framework, and to test its versatility on existing audiovisual data representations. A source for a list of these representations is the state-of-the-art report on the integration of sonification and visualization [5]. With the continuation of the development and validation, the framework has the potential to become a useful tool in the analysis and design of sonification, as well as its integration with visualization.

Acknowledgment

This work has been supported by the Knut and Alice Wallenberg Foundation (grant KAW 2019.0024).

5. REFERENCES

[1] L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660, 8 1999.

[2] M. S. Beauchamp, K. E. Lee, B. D. Argall, and A. Martin. Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5):809–823, 3 2004.

[3] J. Bertin. *Semiology of Graphics*. ESRI Press.

[4] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.

[5] K. Enge, E. Elmquist, V. Caiola, N. Rönnberg, A. Rind, M. Iber, S. Lenzi, F. Lan, R. Höldrich, and W. Aigner. Open Your Ears and Take a Look: A State-of-the-Art Report on the Integration of Sonification and Visualization. *Computer Graphics Forum*, 43(3), 6 2024.

[6] K. Enge, A. Rind, M. Iber, R. Höldrich, and W. Aigner. Towards a unified terminology for sonification and visualization. *Personal and Ubiquitous Computing*, pages 1–15, 8 2023.

[7] T. D. Griffiths and J. D. Warren. What is an auditory object? *Nature Reviews Neuroscience 2004 5:11*, 5(11):887–892, 11 2004.

[8] T. Hermann, A. Hunt, and J. Neuhoff. *The Sonification Handbook*. 2011.

[9] H. M. Kondo, A. M. Van Loon, J. I. Kawahara, and B. C. Moore. Auditory and visual scene analysis: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714), 2 2017.

[10] M. Kubovy and D. Van Valkenburg. Auditory and visual objects. *Cognition*, 80(1-2):97–126, 6 2001.

[11] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S. A. A. Shah, and M. Bennamoun. Scene Graph Generation: A comprehensive survey. *Neurocomputing*, 566:127052, 1 2024.

[12] T. Munzner. *Visualization Analysis & Design*. CRC Press, 1 2014.

[13] M. H. W. Rosli and A. Cabrera. Gestalt Principles in Multimodal Data Representation. *IEEE Computer Graphics and Applications*, 35(2):80–87, 3 2015.

[14] R. F. Tobler. Separating semantics from rendering: a scene graph based architecture for graphics applications. *The Visual Computer*, 27(6-8):687–695, 6 2011.