Audiovisual Resynthesis in an Augmented Reality

Parag Mital
Dartmouth College
Hanover, NH
parag@pkmital.com

ABSTRACT

"Resynthesizing Perception" immserses participants within an audiovisual augmented reality using goggles and headphones while they explore their environment. What they hear and see is a computationally generative synthesis of what they would normally hear and see. By demonstrating the associations and juxtapositions the synthesis creates, the aim is to bring to light questions of the nature of representations supporting perception. Two modes of operation are possible. In the first model, while a participant is immersed, salient auditory events from the surrounding environment are stored and continually aggregated to a database. Similarly, for vision, using a model of exogenous attention, proto-objects of the ongoing dynamic visual scene are continually stored using a camera mounted to goggles on the participant's head. The aggregated salient auditory events and proto-objects form a set of representations which are used to resynthesize the microphone and camera inputs in real-time. In the second model, instead of extracting representations from the real world, an existing database of representations already extracted from scenes such as images of paintings and natural auditory scenes are used for synthesizing the real world. This work was previously exhibited at the Victoria and Albert Museum in London. Of the 1373 people to participate in the installation, 21 participants agreed to be filmed and fill out a questionnaire. We report their feedback and show that "Resynthesizing Perception" is an engaging and thought-provoking experience questioning the nature of perceptual representations.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities

1. INTRODUCTION

Many theories of perception suggest that our perception is derived from internal representations of the sensory information entering our senses but that we are unaware of their details [1, 9, 8]. These representations, denoted by salient auditory events and visual proto-objects, are theorized to be the latest stage of pre-attentive processing and the earliest stage of representation acted upon by attentional machinery. They also do not require semantics or language in order to be represented, but rather provide a basis for understanding objects and events in the world. As we do not have access to them, what are the representations supporting these processes? How are they modeled, what do they look or sound like, what can they explain, and what can they not explain? Rather than attempt to answer these questions through the traditional sciences, we attempt to open a dialogue around them through an arts practice.

This practice is defined by scene synthesis: a computationally generative collage process making use of a computational model of salient auditory events and visual proto-object representations. We place participants within a real-time audiovisual scene synthesis using virtual reality goggles and headphones. Representations of the auditory and visual scenes as measured by a microphone and a head-mounted camera are continually learned while the participant experiences an ongoing scene synthesis process. In one scenario, no explicitly predefined database is used for synthesizing the ongoing scene. Rather, the scene is recreated using sonic and visual representations the software learns over time. By continually associating the incoming input with its aggregated stored representations, we attempt to create syntheses questioning how associations in similar representations may be perceived and re-contextualized.

Ideally, the stored representations will allow for a synthesis indistinguishable from its target. However, what happens when they are entierly unlike the stimuli? For instance, what if we have only learned representations of the sonic world of 'trees' and 'birds'. How would we then synthesize an acoustic scene full of voices? Or in the visual case, what if we had only formed representations composed by paintings of Hieryonomous Bosch, and then had to synthesize the natural visual world as we are used to it? In a second scenario, we allow participants to ask these questions by choosing pre-built models which have been trained on databases of various visual scenes such as those depicted in paintings by Bosch, Van Gogh, and Monet, for the visual synthesis, and scenes such as different musical genres or natural auditory scenes, for the auditory synthesis.

We first contextualize our work within related arts practices. We then briefly motivate our computational models of auditory and visual representations based on literature in auditory streaming and proto-objects, respectively, and

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655617.

then describe our methods. We then show a few example syntheses and report feedback from an exhibition in London at the Victoria and Albert Museum in 2012.

2. RELATED PRACTICES

Our synthesis process can be seen as a form of a computationally generative collage. Historically, collage is an arts practice which appropriates fragments of culture for its materials. Depending on its medium, it juxtaposes, often chaotically, fragments such as photographs or clips of sound, removing them from their original context. By doing so, it is capable of communicating new interpretations which the original fragments alone could not have provided.

The juxtaposition of fragments of sound as an arts practice has roots at least as early as musique concrète, a compositional technique assembling various natural found sounds in order to produce a collage of sound. Digital Sampling came in the 1970's allowing sound segments to be triggered using an interface such as a keyboard or pad. More recent techniques have focused on granuar or concatenative synthesis, where a target sound is matched to a stored database of segments or sounds (for a more in-depth review of these practices, see [3]).

Visual collage practices making use of computation are abundant within graphics communities, and have taken various forms from compositing, texture synthesis, example-based synthesis, and various methods for artistic stylization. Indeed our own synthesis engine is a repurposing of a previous artistic stylization framework known as corpus based visual synthesis [5, 3].

3. METHODS

3.1 Augmented Reality

The exhibition in 2012 had participants wear the Vuzix Wrap920AR goggles. These goggles house two small CRT screens (640 x 480 @ 30Hz with 31 degree field of view), with two front-facing cameras. Due to procesing limitations, only one camera feed was synthesized and displayed on both CRT screens. Since then, this work has been migrated for the Oculus Rift with a camera mounted to the front, as this setup offers an increased field of view and a design which masks all light except for the display. The processing occurs on a laptop nearby, meaning participants are only able to explore as far as the length of the cable (2 meters).

Participants were also invited to wear a pair of headphones, a Beyerdyanmic DT 770 Pro. These were chosen as being comfortable to wear, sanitary (i.e. as opposed to in-ear monitors), and having excellent acoustic isolation due to its closed ear design. The processing for audio occurs on an iPhone, as the application is self-contained in the iOS app, "Memory Mosaic", freely available on the app store[1] [3].

3.2 Computational Synthesis Engine

3.2.1 Visual

Our visual model attempts to actively describe a visual scene, simulating the movements of the eye using an exogenous attention model [6], and representing a scene by its proto-objects [8, 5, 3]. This model effectively describes the shape and color of finer detailed regions at points of likely fixations, and coarser details in places less likely to be fixated. To do so, we first use a previously motivated model of exogenous attention to define the dynamic visual salience of the scene built using contrasts in a dense optical flow map [6, 3]. The point of highest saliency along with the entire saliency map's entropy is then fed to visual acuity filter simulating the logarithmic drop in spatial resolution outside of the point of fixation. An example in shown in Figure 1. As the entropy is low, the blurring is quite substantial, removing many of the details more likely to be unattended such as the high frequency edges in the bricks, grass, and leaves.

(a) Original Image (b) Visual Acuity Filter

Figure 1: 1a: Original frame; 1b: Examples of how the exogenous attention map (inlayed in the image) is used to simulate the point of fixation (drawn as a black/white circle as seen over the man potting).

The output of this image is then used for corpus based visual synthesis (CBVS) [5]. The underlying algorithm of CBVS, maximally stable color regions [2], affords us the ability to sort coarse to fine blobs based on their level sets. We can further define the coarse to fine precision using simple parameters such as timesteps and placing a threshold on the minimum region size. For real-time aggregation of content, we only aggregate proto-objects every 300 ms, as this temporal resolution is also the average time of a fixation [6]. Only proto-objects whose description falls beyond a threshold of a metric computing color and shape similarity are stored. As space is limited, we refer the reader to the paper on CBVS for more details, including examples of visual synthesis using corpora of Van Gogh, Monet, Klimt, and others [5, 3].

In Figures 2 and 3, we show two examples of visual scene synthesis. We also show an example of visual scene synthesis using an alternate database of Hieryonymous Bosch, producing the scene in Figure 2d. In the latter example, we also reveal more of the attention map (3b), its effect on the visual acuity map (3c), and how this distributes the spatial scale of proto-objects (3d).

3.2.2 Audio

Our auditory model is inspired by auditory streaming [1] and evidence of event related potentials demonstrating the brains remarkable ability to detect temporally incoherent auditory events even without our attention to the task (e.g. mismatch negativity) [9]. We model event detection, store them, and then use the same events for resynthesizing any newly detected salient auditory events. The saliency of an event is denoted by its temporal incoherence, which is described by the event's ability to explain the current auditory model (described in [4, 3]). This model is represented as

[1] `https://itunes.apple.com/us/app/memory-mosaic/id475759669?mt=8`

(a) Original video feed

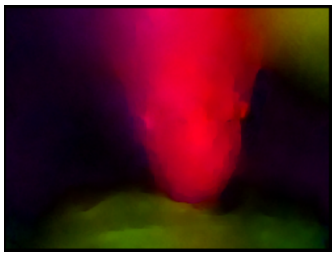
(b) Proto-objects

(c) Visual synthesis

(d) Visual synthesis using Bosch

Figure 2: An example of the real-time visual synthesis. 2a Original image. 2b Proto-objects. 2c Visual synthesis using an aggregated database of the same scene. 2d Visual synthesis using a database of Hieryonomous Bosch paintings.

(a) Input image



(b) Attention map

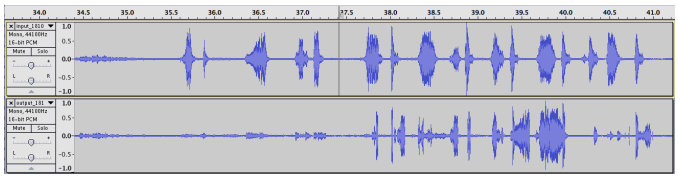
(c) Visual acuity map

(d) Proto-object image

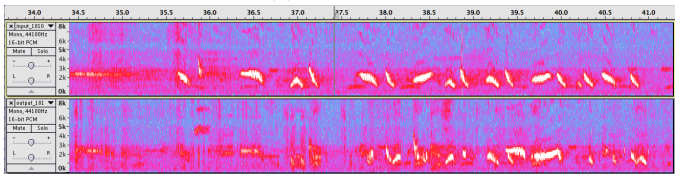
(e) Synthesized image

(f) Another synthesized image

Figure 3: Example processing for producing visual syntheses. The input video 3a is analyzed for the most likely focal point in 3b. This point is used in the synthetic acuity map shown in 3c. This image is used in finding the proto-objects shown in 3d. And finally matched to the existing database to produce the example syntheses in 3e and 3f.



(a) Waveforms



(b) Spectrums

Figure 4: Here we visually represent auditory scene synthesis as its signal and spectrogram. The clip is of David Attenborough documenting the Lyre bird. The original is displayed on top of the synthesis in each case.

a Gaussian Mixture of Mel-Frequency Cepstral Coefficients, their deltas, and their delta deltas. Due to space constraints, we refer the reader to the full implementation details for our event detection model in [4, 3].

Once events have been segmented, their features are concatenated, and matching is done using dynamic time warping. We optimize for real-time performance using recent boundary constraints [7] and for Apple hardware using the Accelerate framework (source code freely available online at `github.com/pkmital`. Finally, synthesis is performed using the Dirac3 LE Time Stretching algorithm[2]. This allows for sequence matches with different lengths to be played back without changing the pitch of the original sample.

In Figure 4, we show an example of audio scene synthesis represented as its original and synthesized waveforms and spectrograms. Notice how the first few large jumps in the original signal's amplitude are matched quite well for their spectra, though not in amplitude. In the later parts of the audio file, the synthesis appears much closer in amplitude, as it has learned more loud sounds, while the spectra is still matched fairly well, though with significant discontinuities.

4. FEEDBACK

Feedback was collected during an exhibition called "Augmented Reality Hallucinations" held at the Victoria and Albert Museum in London in 2012 during the Digital Design Weekend (part of the London Design Festival). The exhibition ran for 2 days from 10 a.m. to 5 p.m. The total number of people attending my exhibit alone as reported by the V&A Museum on Saturday was 445 adults and 128 under 18s and on Sunday, 590 adults and 210 under 18s for a total of 1373 people. Of these, 21 participants agreed to be filmed, photographed, and fill out a questionnaire.

On the feedback form (see [3] for the form and full details of the participant's responses), participants made many qualitative comments regarding the visual aesthetics. For instance, when participants were asked, "Did this experience make you think of anything you had seen or heard before?", three participants made references to their experiences on hallucinogens and two to dreams. Also of note in

[2] `dirac.dspdimension.com`

the qualitative feedback were references to art styles such as, "It reminded me of Francis Bacon's Figurative style" and "The movement was Impressionistic, almost painterly". When asked, "What did you dislike most about the experience?", of note were the responses, "Would have liked more depth in colour", "Not sure what I was seeing at first with the goggles", and "Hard to understand how it works."

Quantitative feedback is summarized in Figure 6. The feedback demonstrates that participants overall rated the visual experience higher than the auditory experience. One possible reason for this may be expressed in the qualititative feedback, with one participant writing, "The granular-like sound is a bit too extreme with headphones." Though another participant also found that "The sound was definitely an enhancement. Not quite the same without."

Participants also made a number of comments regarding perception. When asked what they liked about the experience, participants wrote, "the changing of the perception of reality", "it made me think of how our brain deconstructs and reconstructs elements of perception, particularly in the goggles", "the effect on my perception about the environment and myself as well".

Augmented Reality Hallucinations was also featured in the article, "See Like A Bug On Acid", on the media blog `createdigitalmotion.com`. Editor in-chief Peter Kirn says of the work, "What's great about this project is the way in which it alters reality - with the aid of video input".

Figure 5: A few participants of the installation wearing the AR goggles. Headphones (shown in the top picture to the left) were worn at times, though are not shown being worn here. Photos by the author. Participants gave written consent to be photographed.

5. CONCLUSION

We have described "Resynthesizing Perception", an augmented reality experience employing a computational model of visual and auditory perception to synthesize the world around us. It functions in two possible modes: (1) by aggregating representations learned from the world, and (2), by using a predefined database of representations learned from different scenes such as those of painters and auditory scenes of nature or different musical recordings. By demonstrating the associations and juxtapositions the synthesis creates, its aim is to bring to light questions of the nature of representations supporting perception. One failing of the model has been that we consider either modality separately. Certainly many multisensory and crossmodal effects have been demonstrated in the literature (e.g. the auditory override of visual perception as demonstrated in the double flash illusion and the visual override of auditory perception as demonstrated in the McGurk effect). It would be interesting to explore such interactions in the future.

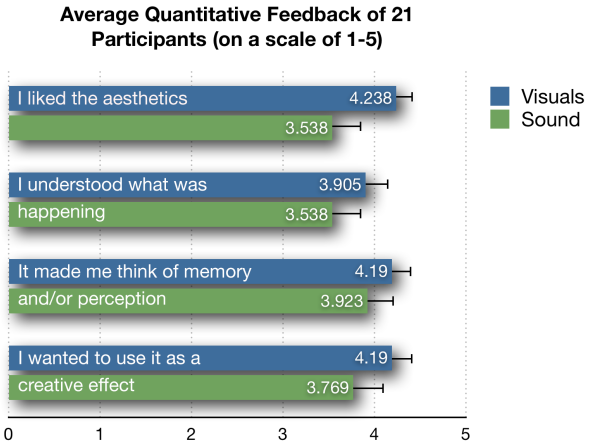


Figure 6: Feedback where 21 participants were asked to rate different aspects of the auditory and visual synthesis. Error bars depict +/- 1 S.E.

6. REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. A Bradford Book (September 29, 1994), 1990.

[2] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, Minneapolis, MN, 2007. IEEE.

[3] P. K. Mital. *Audiovisual Scene Synthesis*. Ph.d. dissertation, Goldsmiths, University of London, 2014.

[4] P. K. Mital and M. Grierson. Mining Unlabeled Electronic Music Databases through 3D Interactive Visualization of Latent Component Relationships. In *NIME 2013: New Interfaces for Musical Expression*, Seoul, Korea, 2013.

[5] P. K. Mital, M. Grierson, and T. J. Smith. Corpus-based visual synthesis. In *Proceedings of the ACM Symposium on Applied Perception - SAP '13*, pages 51–58, New York, New York, USA, Aug. 2013. ACM Press.

[6] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. *Cognitive Computation*, 3(1):5–24, Oct. 2010.

[7] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 262, 2012.

[8] R. a. Rensink. The Dynamic Representation of Scenes. *Visual Cognition*, 7(1-3):17–42, Jan. 2000.

[9] I. Winkler, S. L. Denham, and I. Nelken. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences*, 13(12):532–40, Dec. 2009.