Goldsmiths
UNIVERSITY OF LONDON

Last update on 2013/09/20 at 21:44:07

Audiovisual Scene Synthesis

Parag Mital

thesis submitted in partial fulfillment for the title of

PhD of Arts and Computational Technologies

from GOLDSMITHS - UNIVERSITY OF LONDON

Thesis Advisors:
Michael Grierson
Timothy Smith

Member of the Embodied Audio-Visual Interaction Lab

defended on January 14, 2014

Jury :

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Keywords: synthesis, scene analysis, encoding, decoding

Contents

List of Figures

Chapter 1

Introduction

Contents

These fragments I have shored against my ruins.

– T.S. Eliot, excerpt from *The Waste Land*

1.1 Motivation

The world in front of us is measured by the various sensory mechanisms we have available to us. These mechanisms enable us to convert the physical phenomena into electrical signals which we use to construct a perception of the world. However, we are often taught that the world we perceive is exactly as it is in front of us. Philosopher Alan Watts describes the situation:

> Most of us are brought up to feel that what we see out in front of us is something that lies beyond our eyes, out there. That the colors and the shapes that you see in this room are out there. In fact, that is not so. In fact, all that you see is a state of affairs inside your head. All these colors,

> all these lights, are conditions of the optical nervous system. There are, outside the eyes, quanta, electronic phenomena, vibrations, but these are not light, they are not colors until they are translated into states of the human nervous system. So if you want to know how the inside of your head feels, open your eyes and look. That is how the inside of your head feels
>
> Alan Watts

Though perhaps a bit lyrical, Watts describes the feeling of the inside of our heads as the one that we often mistakenly describe as what is "out there". Many theories of perception suggest that we encode the sensations entering our sensory mechanisms into a set of internal representations inaccessible to our consciousness[1]. Across numerous literature these representations are theorized to be the latest stage of pre-attentive processing and the earliest stage of representation acted upon by attentional machinery. As a result, 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 conscious access to them, what are the representations supporting these processes? How are they modeled, what do they look like, what can they explain, and what can they not explain?

It is the aim of this thesis to develop a better understanding of questions such as these through a computational arts practice defined by a process called *scene synthesis*. Scene synthesis is a computationally generated collage which encodes scenes such as images, videos, or sounds, into a set of internal unconscious representations using computational processes modeled by our own early perceptual machinery. Once a set of units for the collage have been encoded, scene synthesis attempts to decode a target scene such as an existing sound clip, image, video, or for interactive experiences, a real-time video feed and/or microphone using the set of previously encoded internal representations.

Scene synthesis, as it is developed in this thesis, is capable of building interactive, dynamic, and computationally generated collages. A defining aspect of the interaction within this framework is to place a participant within the collage-generation process

[1] This thesis uses consciousness and awareness interchangeably. Awareness denotes the ability to report something. Consequently, unconsciousness denotes things of which we are not aware, and things that we cannot report.

such that the generated collage is specific to that participant. As a result, the encoded representations are based on the participant's previous experiences or their own relationships to the stimulus material. For instance, in one incarnation available as an iOS app, "Memory Mosaic", a participant experiences a real-time stream of sonic collages created automatically by the software. The collages are generated using sounds the app has heard since the participant started the app. By continually associating the incoming input with its stored memories, the app attempts to reveal a basic perceptual process of auditory listening through the decoded collages. In an analogous fashion, visual scene synthesis aggregates representations learned from visual scenes, such as a live stream of a webcamera, in order to build a set of representations used for decoding any other visual scenes. This framework is applied in an iOS application called "PhotoSynthesizer" where a user can encode any number of source photos in their Photo Library. They can then select a target image to decode, which is then presented as a painting, starting from the background and eventually reaching the finer details of the foreground layers.

Overall, scene synthesis can be summarized by coding processes, where an *encoding* stores representations of an audiovisual scene using a model of representation based on human perception, and the *decoding* attempts to describe the stimuli again. The terms encoding/decoding are certainly not without their history. In the 1940's, Shannon mathematical formulation of a generalized communication system described the communication between a source and receiver through a transmitter and receiver, respectively. The source encoded the message, and the transmitter would convert the message to physical phenomena, thus introducing possible noise sources effecting the message. Finally, the message could be received and decoded finally by any destination that was capable of recovering the message. His developments formed the basis of Information Theory, and along with developments occurring in military technology would later be incorporated by Norbert Wiener in a seminal text on communication theory entitled, Cybernetics.

Norbert Wiener's seminal text, "Cybernetics; or control and communication in the animal and the machine" (Wiener 1948), describes a field of research denoted, cybernetics, which evolved out of military experiments in the 1940's describing feedback processes which repositioned actuators (e.g. missiles/guns) based on updated predictions (e.g. sonar/radar/filters). As an example, a matched filter provided a template response of a radar signal (e.g. of a known target) and was matched at each radar sweep providing a new orientation towards a target. This new orientation could then be fed

back into the system which would then again perform a radar sweep, a matched filter response, and a new heading, *ad infinitum*. This process forms the basic idea of homing in missile guidance, and is a defining principle of reducing noise in the field of signal processing (i.e. it maximizes the signal-to-noise ratio of a known signal corrupted by additive noise).

It has also been likened to processes of active perception by Wiener himself. Wiener suggests that each movement of the eye and neuronal levels of processing leads towards an invariant object representation (discussed in greater detail in Chapter 5). This entire process, in Wiener's own words, "brings [visual] information one step nearer to the form in which it is used and is preserved in the memory." Wiener's early insights into the function of the eye, neuronal processing stages, and internal representations supporting perception will be expanded upon more in this thesis, as I attempt to motivate and model a similar perceptual process in both auditory and visual domain and use the model within a collage-based practice.

Cybernetics provided a vehicle for mathematicians and biologists in the coming years (e.g. Ross Ashby, Gregory Bateson). It treated individual humans, animals, machines, or in cybernetic terms, *systems*, as individual entities whose inner functions were not measured by their own capacities, but only in relation to what could be communicated by its own outputs to other systems. It offered a new way of thinking about complex systems in terms of the information that was measurable outside of the contained systems.

Early cyberneticians, information theorists, and others at the time would eventually form the basis for modern day Information Sciences, where encoding and decoding are still very much used. For instance, it is used to describe the processes of information compression, where a very large file such as an image, document, or video must be transmitted or stored more efficiently. To achieve this process, it is first encoded to a different representation by an encoder, only to later be decoded back into its uncompressed original form when it is retrieved or accessed. Decoding occurs every time you listen to an (encoded) MP3 file, where the MP3-codec describes the representation for compression and the procedures for encoding/decoding to/from it. These terms are also commonly used within the field of Neural Coding, where encoding attempts to describe the neural populations representing brain-based representations of an external stimuli and decoding attempts to work from the neural populations back to describing the stim-

uli as a whole again. The field has grown significantly in the last 10 years due to its successes with decoding images, video, and sound, as well as the recent developments within brain imaging and pattern recognition technologies.

Ideally, the representations motivated and built in this thesis will allow for a synthesis indistinguishable from its target. However, what happens when representations used by the encoder come from a set of representations that are unlike the stimuli? For instance, what if my encoder only knows what the world of 'trees' and 'birds' sounds like? How would it then decode a Michael Jackson recording? Or in the visual case, what if my encoder only ever saw the animation "The Family Guy", and had to decode "The Simpsons"? This thesis asks these very same questions when discussing audio scene synthesis in Chapter 4 and visual scene synthesis in Chapter 7. More generally, these questions are framed as "how does experience shape perception?", suggesting that perception is not a one-to-one relationship with the world but one experienced based on our prior notions of it.

We will discuss these and more experiments along the way to our final aim, the combination of both auditory and visual modalities in two outputs: (1) "Augmented Reality Hallucinations", exhibited during the London Design Festival at the V&A Digital Design Weekend, which uses an augmented reality headset originally built for immersive gaming environments delivering the visual experience and headphones to produce the sonic experience; and (2), "YouTube Smash Up", which synthesizes the number 1 video on YouTube of the week using the top 2 - 10 videos of the week.

We first discuss background in collage-based practices in Section 1.2 and computational arts-based practices in Section 1.3. These fields serve to build a better context of the process of scene synthesis within associated practice-based domains. Ideally, we would also include background specific to the theoretical aspects of perception and representation as well as their computational modeling inside this introductory chapter. However, for fear of overloading the reader too soon, we save these background for the later chapters. After this introductory review, we discuss the goals and the outline of the remainder of this thesis in Sections 1.4-1.5.

1.2 Background in Collage-based Practices

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, text, 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. It is inherently a process that invokes new ways of seeing creating a meaning-making process between the artist producing the chaos and the audience that must unify its cut-up percepts into order. More than an artistic technique, it has also been described as a "philosophical attitude" that can be applied to virtually any medium (McLeod 2011). As collage-based practices have progressed over the last 100 years, media too has evolved, affording practitioners new tools and technologies for manipulating and accessing media. Specifically, within computational practices, the procedural manipulation of media content has enabled artists to create dynamic and interactive experiences through algorithmic processing of the media. These advances have enabled us to move collage beyond a static medium and into a dynamically generative one. As we will see, using computational methods in information retrieval, this thesis aims to build a computational framework for generating a particular type of collage called a *scene synthesis*.

This section situates the thesis within a lineage of collage-based practices starting with seminal developments in collage, montage, cut-up, and musique concrète, before moving to an overview of modern developments in technology that radically changed collage practice in terms of its practice, publishing, distribution, and even legality. Finally, computational methods developing procedural manipulation of content for producing collages are discussed.

1.2.1 Beginnings

Though early practices of collage such as the invention of paper in China around 200 BC, calligraphers in Japan in the 10th century, the 15th and 16th century practices of adding gold leafs or other gemstones to canvases, and Giuseppe Arcimboldo's 16th century paintings of portraits composed of fruits, vegetables, and books, it was not until the first-half of the 20th century that an explosion of collage practices spanning visual,

textual, and sonic mediums was exhibited.

In 1912, Pablo Picasso transformed a still-life by gluing an oilcloth to the canvas in *Still Live with Chair Caning*. Along with Georges Braque, they experimented with gluing visual fragments of culture represented by stamps, newspaper clippings, and photos to their canvases. Shortly after in the 1920's, collage-based practice would define a major tenant of the Paris and Berlin-Dada scenes as seen in the work of George Grosz, John Heartfield, and Kurt Schwitters who would often work entirely with found objects meant to serve as representations of the city (e.g. *Irgendsowas*, 1922). Collage also found its way to purely cut-up photographic material, a technique known as photomontage, which can be seen in the work of Max Ernst's *Murdering Airplane* (1920) and Hannah Höch's *Pretty Maiden* (1920). Ernst would also later release the first collage novel in 1929, *La femme 100 tête*, which comprised of 9 plates and included an introduction by the founder of Surrealism, Andre Breton.

Breton notably describes the practice of collage in the *Manifeste du Surréalisme* published in 1924 in terms of an image created by the mind that could not be born from a comparison, but only from a juxtaposition of two distant realities (Breton 1924). He also wrote of a popular parlor game amongst Surrealists in the 1920's called "Exquisite Corpse" which had participants collectively assemble text or images often using simple rules such as, "adjective then noun". Breton later describes the game: "they bore the mark of something which could not be created by one brain alone...fully liberating the mind's metaphorical activity" (Breton 1948).

In a similar vein to Exquisite Corpse, Tristan Tzara while at a Dadaist rally in the 1920's performed a poem by taking cut-up fragments of text-based media such as newspapers or brochures out of a hat and reading them aloud (eventually leading to a riot destroying the theatre on location and the expulsion of Tzara from the movement by Andre Breton). The "cut-up technique", as it is now known, can be found in some of modern literatures greatest works, such as T. S. Eliot's *The Waste Land* and James Joyce's *Ulysses*, both published in 1922. Cut-up is later rediscovered by Brion Gysin who supposedly accidentally rediscovered the technique and consequently shared the technique to William Burroughs in the 1950's. Burroughs made extensive use of cut-up in his writings, most notably in *Naked Lunch* where each chapter was to be read in any order, and in his collaboration with Gysin, *The Third Mind*. Burrough's notes

the power of the cut-ups to conjure up the associations of dreams, stating that he has "quite deliberately addressing [himself] to the whole area of what we call dreams" and has been "interested in precisely how word and image get around on very, very complex association lines" (as quoted in (**?**)). The technique would also eventually find its way into mainstream music in the lyrics of David Bowie, Kurt Cobain, and Radiohead.

Collage practices did not stop with static media, however. In 1925, the Russian film director Sergei Eisenstein demonstrated the power of film montage in *Battleship Potemkin* as he juxtaposed image sequences such as a crowd's flight down a staircase with the image sequence of a baby carriage for 7 minutes, creating viscerally new experiences and emotions than either sequence alone could have. Walter Ruttmann would later explore the use of this practice with purely sound, in a composition in 1928 entitled, "Wochenende", German for weekend, in what would later be known as the first image-less film (Kahn 1994). Comprised of sounds of a moving train, the wind through the trees, voices of whispering lovers, and the sounds of a crowd, Ruttmann produced a montage through abstracting the fragments relating to the start of a weekend into what Hans Richter would describe as, "a symphony of sound, speech-fragments and silence woven into a poem" (Ltd. 2013). Such practices, while not strictly collage as it had been known as, shared the technique of producing new meanings from the juxtaposition of individual fragments. However, these fragments are temporally related, and thus the juxtaposition operates on an imposed temporal dimension rather than a spatial one to create a distinct image in the mind's eye.

By the end of the 1940's, radiophonic art, or the practice of producing sound for radio broadcast, had been well established. Words, music, and noises were combined to produce radio productions of literary stories and news broadcasts. It is no surprise then that in one studio in Paris, France, Pierre Schaeffer was also experimenting with splicing and recombining magnetic tape recordings of sound in a practice later called musique concrète (). Schaeffer argued that classical music practice begins with a set of abstractions which are notated on sheet music and then later reproduced as audible music. Musique concrète on the other hand begins with recordings of the sound outputs themselves, attempting to work from the raw sonic outputs towards any abstraction that may describe them. (Augoyard 2006)

Collage would also find its way into performance art. Most notably, the works of

choreographer Merce Cunningham were heavily influenced by collage-based practice. In his 1964 work in Vienna, *Events*, he takes "splices" of his existing works and recombines them to produce new configurations, further heavily influenced by the chance operations demonstrated in works by Marcel Duchamp and John Cage (Copeland 2002). One may also notice that Duchamp's notion of readymades (e.g., the urinal or bicycle wheel) is also apparent within his collage-based practice. Copeland describes the influential nature context has on the interpretation of the individual fragments within *Events*, suggesting that "when segments of *Winterbranch* (1964) were incorporated into subsequent Events, the emotional tone of the work was no longer nightmarish or apocalyptic [...] what the audience saw was (merely) the act of dancers falling [...] Without other factors to color the emotional texture [...] spectators tended to laugh rather than to recoil in horror" (Copeland 2002). Nam June Paik as well explored collage within performance in *Random Access/Paper TV* (1978-1981), which payed homage to both Cage and Cunningham through silk-screened images of their performances printed on two decks of playing cards. Paik recombined and shuffled the playing cards incorporating chance operations in a similar vein to the early Dadaists and created collages through the dealing of the cards out onto a flat surface (Copeland 2002).

Perhaps the most sophisticated use of visual collage-practice comes from Czech surrealist animator Jan Švankmajer, who combines stop-motion and collage-based techniques within his intricately detailed and surreallist animations. Film theorist Jan Udha describes his work as rapidmontage, "offering amazing and original associations, a kind of kinetic collage" (Uhde 1994). For example, in "Možnosti Dialogu" or, "Dimensions of Dialogue" (1982), Švankmajer creates a collage of two heads composed of a variety of coarsely fragmented objects in the style of Giuseppe Arcimboldo, one composed of various kitchenware and another composed of various fruits and vegetables. Their "dialogue" ensues which effectively entails the one head eating the other one, chewing the head momentarily, then regurgitating it as a finer representation of objects. The process continues in turns with one head eating the other until they resemble smooth clay-like sculpted heads.

1.2.2 Technological Developments

Many developments in technology significantly altered the practice of collage, allowing it to explore entirely new processes for composition, new mediums and new methods of presentations, and new audiences.

1.2.2.1 Visual Culture

For instance, the Xerox Corporations development of the photocopier led to a surge of do-it-yourself small-publishing houses. As well, it led to the development of xerox-based collage artists, most notably featured in a bi-monthly xerox-printed zine called *PhotoStatic* which had a peak circulation of 750 copies in the 1980's finally ending with its 41st issue in January 1993 (McLeod 2011).

As well, Polaroid's introduction of instant film allowed photographers to instantly see their developed film without the need for mixing chemicals. David Hockney explored their use creating patchwork collages (for instance, using as many as 63 Polaroids in *Still Life Blue Guitar*, 1982) which he called "joiners". The individual patches comprised of individual photos that were taken at different times effectively creating different exposures, lighting conditions, and vantage points. Hockney's interest in the joiner's were intimately tied to human perception, stating that they made "things closer to the truth of the way we see things" (Joyce 1988), i.e. perception constructed through the fragments of many perspectives.

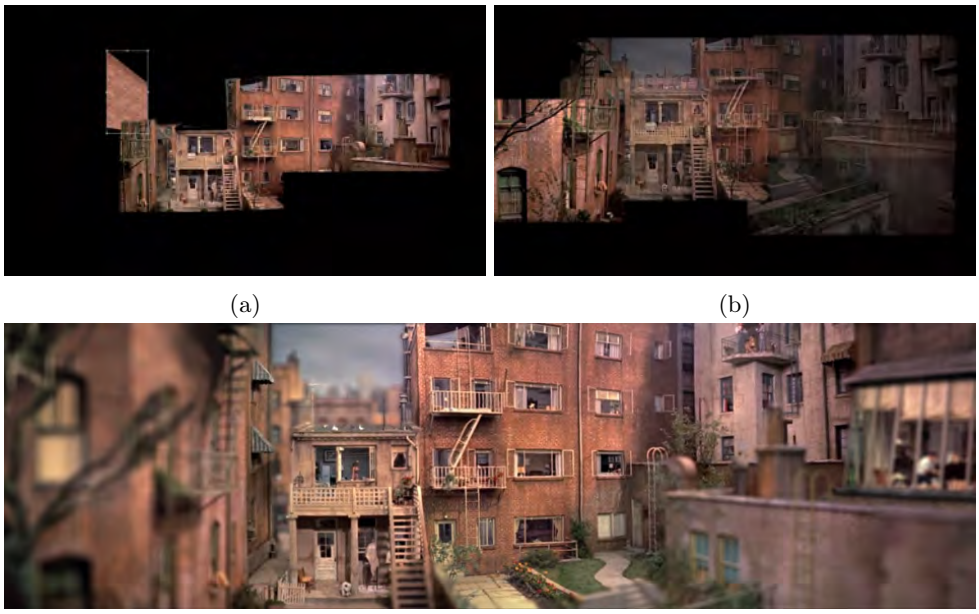
1963 saw pivotal developments for the field of interaction with computers and in particular with computer graphics: Douglas Englebart's computer mouse and Ivan Sutherland's sketchpad (an extension of Robert Everett's light pen for the TX-2). Both allowed a user to isolate and drag pixels on a screen, moving them from one data-space to another. DJ Spooky remarks on the transformations these and other pioneers in human-computer interfacing technologies would have on the arts: "Douglas Engelbert and Ivan Sutherland pioneered graphical user interfaces...but what they accomplished was even more profound than that, their work let us move into the screen world itself" (quoted in **??**).

It is in the 'screen world' that a wealth of new technologies would be developed in the coming years to allow the manipulation of data representing media. As a result,

perhaps the most significant deviation from early collage practice comes from the advent of digital media. As media was no longer contained to the physical world, but a virtual screen-based one, where only software could manipulate the media, collage practice was intimately tied to the capabilities of software.

The notion of collage in the world of digital media may require pinning down, as a variety of new techniques can afford new kinds of compositions. One theorist claims that collage will proliferate so long as "cut" and "paste" operations persist, as described by Taylor in *Collage* (Taylor 2006). Cut-and-paste, in the digital realm, implies an operation of copying a digitally stored representation (i.e. 1's and 0's in machine code) from a source to a target. However, for most computers, a copy operation is carried out whenever information is simply accessed. That is, a digital piece of information is first kept stored in a permanent storage location such as a hard disk, before being copied by software. For example, when a sound or image is required for display or decoding (e.g. from an MP3-file to one that can be played as -1 to +1 values through the speakers, or a JPEG-file to one that can be displayed as RGB values in a monitor), the permanent file must be copied into a temporary high-speed memory storage, and used within the appropriate software making use of that information. Even for lossless file formats, such as a WAVE audio file or RAW image file, most software will require the file to first be copied into high-speed memory to allow it to be manipulated, displayed (e.g. as a waveform or image), and possibly copied again in order to interface with hardware. As a result, copyright laws have likened the pure access and interfacing of digitally stored information as falling within the domain of copyright, leading to several attempts to digitally manage content (e.g. Digital Rights Management, or DRM). The implications of the interpretation of copyright within digital media will be explored more in Chapter 8, when discussing YouTube Smash Up, a project that explores an audiovisual collage within short films through the use of existing copyrighted material hosted on YouTube.

Software for page layout, illustration, graphics, and photo/sound/ and video editing have afforded practitioners with faster and more sophisticated methods of collage. For instance, Adobe Photoshop supports the automatic parsing of an image into object regions that can be individually manipulated enabling artists to finely segment and compose visual media. For video, Adobe After Effects can similarly segment objects and track features across frames, making dynamic visual collages much easier to do.



(a) (b)

(c)

Figure 1.1: A few frames from Jeff Desom's *Rear Window Timelapse* showing the compositing process in Figures 1.1a and 1.1b to produce the seamless composite in 1.1c (reproduced with permission from the artist)

Further, blending and compositing effects allow for seamless compositing in a variety of interesting scenarios.

As one example, design agency leftchannel likely made use of both Photoshop and After Effects in producing a 3D video cutout effect for the musician RJD2 in their music video for *1976*. Though the agency does not describe their technique, a similar process is described in a tutorial on videocopilot.com [2]. Taking still images, they could have used Photoshop for removing objects within the still frames. The voids left from removing these objects are filled in using contextual and image replacement tools within Photoshop. Finally, within After Effects, individual layers are placed at different z-coordinates, where each object cutout in Photoshop represents a new layer, and the filled in image representing the background. Using this interesting reconstruction of the image within a virtual 3D environment, they then use a virtual camera to move around the scene, creating interesting dynamics that seem as though the image comes alive. The virtual 3D scene afforded by After Effects allows the artist to create collages in a 3D space from using only 2D material.

In another variety of After Effects collaging, Jeff Desom takes individual shots within

[2] http://www.videocopilot.net/tutorials/virtual_3d_photos/

Alfred Hitchcock's *Rear Window* taken by the main character in the story, Professional photographer L.B. "Jeff" Jeffries who is wheelchair bound, and reassembles them as a giant panoramic time-lapse view of his backyard in *Rear Window Timelapse*[3]. The compositing process is shown in Figures 1.1a-1.1b to produce the seamless composite in Figure 1.1c. As can be seen, there are no rough borders or edges where the individual frames were blended together. The composite image plays as though the film had been shot in widescreen in the first place, though all the content came from only much smaller shots from the actual film.

JK Kellers' *Realigning my thoughts on Jasper Johns*[4] creates vector-based collages of the introduction for the television animation series, *The Simpsons*, using a variety of features within Adobe Illustrator to vectorize the original image. The artist describes his process with the accompanying video as first ripping each frame, then turning the rasterized images into vectors and using a variety of alignment and distribution tools to sort the vectors spatially. He further describes a similar process for the audio where he takes images of the spectrogram and applies similar manipulations in Illustrator. The self-described result is a "frenetic mess of sight and sound".

Various other software such as Adobe Director, Adobe Flash, VJamm, VDMX, Apple Quartz Composer, Cycling '74's Jitter, Resolume's Avenue, VVVV, Modul8, Algoriddim's VJay, or Processing have created greater possibilities for video-based media than standard non-linear editors such as Adobe Premiere or Apple Final Cut Pro could offer (Jaeger 2006). These tools opened video-based practices to real-time and performative practices within Visual Music and VJ cultures. Many Visual Music and VJ artists even created their own software, releasing them to the public such as Matt Black of Coldcut, who created VJamm, Netochka Nezvanova who wrote Nato.0+55, or Miller Puckette, author of pd (Jaeger 2006) (we discuss more computational approaches in Section 1.3). Though these practitioners did not employ collage solely, it was certainly common of VJ practice to use fragments of existing material in the form of found video footage and mix them.

Situationists detournment

Culture Jamming

[3] https://vimeo.com/37120554
[4] https://vimeo.com/37127916

William Burrough's Electronic Revolution

Scratch Video of British 1980's

Collage-based video practitioners in the 1990's would often use clips of appropriated television footage such as the news footage of the motorcade leading to John F. Kennedy's assassination in Steinski's "The Motorcade Sped On", George H. W. Bush's news broadcasts on the Gulf War in EBN's "We Will Rock You", or Coldcut's use of nature videos. However, these early examples as well as some recent artists such as Eclectic Method did not employ collage in the traditional spatial-sense, i.e. by juxtaposing fragments spatially across a single frame, and generally kept any source frames intact. They did however experiment with the use of "blending", a technique with takes 2 or more images and blends the two together using a variety of functions such as: add, multiply, subtract, or difference.

James Kerr appropriates images "found here/there on the internet (a lot taken from northern and early renaissance paintings)" within his animated digital collages (**?**) created using the Graphics Interchange Format (GIF), a common image format created in 1987 supporting simple frame-based animations supported by virtually all browsers (**?**). The GIF format makes it easy to see animations anywhere in a web page as they are often embeddable in other common social networks such as Facebook and Twitter.

1.2.2.2 Sound Culture

For sound, collage-based practices exploded with the birth of digital sampler hardware. Similar to the practice of collage or montage, sampling refers to taking portions of existing media and using it within a performance or composition. Digital sampling refers to sampling after an analog-to-digital process, where physical vibrations of sound are digitally sampled at equally spaced intervals, creating a discretely sampled representation of the continuous real-world phenomena. This discrete sampling can easily be encoded by 1's and 0's in computer storage, and easily decoded back to the physical world, via an digital-to-analog process, such as through a speaker.

Early digital samplers such as the Fairlight CMI to more recent non-linear editors such as Logic and Ableton Live have made multi-track and cut-and-paste operations trivial to accomplish, while visualizing sound waves has made finding relevant parts

of an audio file relatively easier than listening to an entire tape reel. Early adopters of digital sampler technology include Herbie Hancock and Public Enemy. Founders of Dub, Lee "Scratch" Perry and King Tubby, also made use of existing recorded material. Though not strictly sampling, they created the famous Dub sound by infinitely collaging the same sound, creating intense reverberations that echoed with greater magnitude on each bounce until the sound was cut.

In 1987, the KLF produced "What the Fuck Is Going On?", which made extensive use of samples from The Monkees, The Beatles, Dave Brubeck, Led Zeppelin, Whitney Houston, and ABBA, amongst many others, citing on the album liner notes that the samples had been freed "from all copyright restrictions". Despite their claims, their independent release was ordered to be destroyed by the Mechanical-Copyright Protection Society, leading to a re-release of the album with periods of protracted silence in place of the unauthorized samples. They also released a guide including a detailed construction of how to reproduce the sound of the album, including the hardware used: an Apple II computer, a Greengate DS3 digital sampler peripheral card, and a Roland TR-808 drum machine.

Two years later in 1989, John Oswald released an amalgamation of sampled music in his album, "Plunderphonics", including a visual reference to Michael Jackson's "Bad" on its album, which featured a derivative image of Jackson's original album cover for "Bad" edited to make it look like Jackson was a naked woman wearing a leather coat. On the album, a song by the name, "Dab" was collaged to create the essence of a Michael Jackson track, using samples from Jackson's "Bad". Oswald describes his process of sampling as using "plunderphones", describing them as, "a recognizable sonic quote, using the actual sound of something familiar which has already been recorded", satisfying the essence of being a plunderphone "as long as you can reasonably recognize the source" (Oswald 1985; Oswald 2013a; Oswald 2013b). In his writings available online, he further describes his motivations: "'Plunderphonics' is a term I've coined to cover the counter-covert world of converted sound and retrofitted music where collective melodic memories of the familiar are minced and rehabilitated to a new life" (Oswald 1985; Oswald 2013a; Steenhuisen 2005). Oswald's collage-based practice also extended to text, taking cut-up fragments from existing authors without citation, even including un-cited quotes to his own previous text "Plunderphonics, or Audio Piracy as a Compositional Prerogative" in "Creatigality" and "Bettered by the Borrower" (Tholl 2010).

Unfortunately for Oswald, the Canadian Recording Industry Association ordered him to cease-and-desist production and to destroy all remaining copies of "Plunderphonics". In a similar fate as Oswald, sound collage artists Negativland in 1991 were sued by the pop band U2's label, Island Records, for their use of the trademark "U2" on the album cover and their sampling of U2's 1987 song, "I Still Haven't Found What I'm Looking For".

Perhaps the most pervasive and popular use of sampling, or creative plagiarism as Kembrew McLeod cites it (McLeod 2011), however, came in the form of Hip-Hop music. Public Enemy's song, "Caught, Can I Get a Witness?", released in 1988, remarks on the practice of digital sampling:

> Caught, now in court 'cause I stole a beat
> This is a sampling sport

Why did I bring up Hip-hop?

1.3 Background in Computational Arts Practice

Until now, we've seen how technology has reshaped the practice of collage in interesting ways. We now focus on collage-based practitioners that do not just make use of existing technology, but also write their own software. Such practitioners are often the creators of algorithms and offer new ways of procedurally manipulating media.

1.3.1 Synthesis

Synthesis, across a range of computational practices, describes a technique for creating media. In sound, this can mean generating soundscapes (e.g. (**?**)), entire phrases of music (e.g. ()), vowels (e.g. ()), or even entire phrases of speech (e.g. (**?**)). For visual applications, similarly, synthesis can describe a variety of approaches for computationally generating images describing types of textures or patterns (e.g. (**?**)), artistic renderings (e.g. ()), or for video, dynamic textures such as clouds or fire (e.g. (**?**)). In this section, we limit our review of this expansive field to approaches within arts practices specifically described by collage-based techniques which therefore also require the use of existing content.

Within musical applications, a wide-number of practitioners have adopted the technique of granular synthesis. This approach to generating sound works with an existing sound file or entire library of sounds and segments the files into "grains", or tiny fragments of sound generally lasting up to 50 ms. The grains are collected and depending on the application, can be "scrubbed", meaning an interactive controller, typically a mouse, can select the grain to be played back by moving a mouse cursor over it. Granular synthesis has seen wide-adoption within electro-acoustic and contemporary music practices since its first digital implementation by Curtis Roads in 1980-something

For instance, Joseph Nechvatal's computer-robotic assisted paintings such as *Birth of the Viractual* (2001) contains elements of a portrait being "eaten" away by computational processes. The individual generative "viruses" create distinct boundaries that could easily have been individual pasted elements, though are only created through the dynamic time course of the individual processes. Thus, more often than collage, the notion of generative or viral painting applies to Nechvatal's work. However, one may be able to argue that the paintings are a collage of procedures (i.e. the generative processes, or individual viruses), modeled through their own collage-based set of rules (i.e. each individual virus will "eat" another pixel by copy-and-pasting from other ones).

Ben Bogart and Philippe Pasquier's Context Machines, Ben Bogart's Dreaming Machine

1.3.3 Glitch Art

In glitch art, artefacts exposed by procedural errors are exploited in creative and interesting ways. For instance, datamoshing refers to the practice of intentionally altering information within encoded file formats so as to produce distorted interpretations of that file when it is decoded. The technique itself can be applied in many different ways depending on the codec used. As an example, when using Motion-JPEG, a popular video codec, artists often remove a particular kind of "keyframe" which tells the decoder when it should update all of its pixels. In between these keyframes, the decoder will not use the actual content of the video, but instead track the individual pixels from the previous keyframe so that new image content does not need to be created. This effectively helps keep the file size down while maintaining the important parts of the content. When an artist removes these keyframes intentionally, the image content is

Frame 1 Frame 2 Frame 3

Figure 1.2: 3 example frames demonstrating datamoshing from Yung Jake's "Datamosh" Music Video located at https://www.youtube.com/watch?v=nS7QvOX8LVk. In Frame 1, Justin Bieber is displayed. By removing a keyframe, Yung Jake is able to move Bieber with his own motions in the later frames, as shown in Frames 2 and 3.

tracked into a new video frame where much of the content should have been updated, effectively tricking the decoder into thinking it does not need to update all of its pixels. This has the effect of one frame bleeding into the other in what Peter Kirn describes as causing "frames to melt into one another like wax" (Kirn 2009). Perhaps one interesting feature of the effect is that many people are already familiar with compression artifacts (e.g. from poor bandwidth streaming where network packets encoding the keyframes may have been lost). When the artefacts are appropriated for creative use, however, they can create subtle and targeted effects. Artist Yung Jake demonstrates a variety of these in his music video for "Datamosh". See Figure 1.2 for an example where he moves Justin Bieber's face with the motion of the next few frames. He further explains the technique in his lyrics for the song:

> Datamoshing cool, datamoshing great
>
> Justin Bieber move him with my face[5]
>
> Then use it for an art show, use it for a piece
>
> ...
>
> You thought it was an accident, a video glitch
>
> I did it on purpose though... it's nothing
>
> You don't have bad Internet, I'm just datamoshing[6]
>
> ...
>
> I'm on my datamosh, digital effects, man
>
> I'm moshing data, geeked up can't stand[7]

[5] Jake is describing the effect from removing a keyframe where before there was a picture of Justin Bieber, and after there is Jake's own face. The effect of removing the keyframe allows him to move Bieber's face with his own movements

[6] Jake is describing how the effect resembles streaming artefacts from slow connections

[7] Jake makes mention to being "on datamosh" and that he's "geeked up" and "can't stand", suggesting the effect of data mosh has a drug-like power.

Yung Jake – Datamosh

Other techniques in glitch-based art include opening JPEG or other similarly encoded files in a normal text-pad and rewriting parts of it using normal keyboard letters. This immediate representation (i.e. the one viewed in the textpad) is not meant to be decoded by a text parser. However, the interaction of writing text within this format is then stored back into the file, and when read by a decoder, viewed as an image. The parts that were changed are likely unknown by the one manipulating the file, so there may also be a trial-and-error process.

Certainly, the methods built in this thesis when viewed in terms of encoding and decoding rather than typical collage-based practices can be seen to reflect similar processes. The encoding process, leading to a particular representation, is capable of expressing the original information, though within a different method of representation (e.g. for image, typically pixels will be used; however, in this thesis, we will build an image using a set of psychologically-motivated representations resembling regions). When distorting the immediate representation in some ways, the decoding is likely to have unexpected results. We will see how this process can be used to a variety of effects including artistic stylization of image/video or for producing hallucinations in Chapter 7.

in order to trick the decoder into thinking a new refresh of the image content isn't necessary.

one's notions of the original compression with Jacques Perconte: http://www.electronfestival.ch/en/artist/159

Dimensionality reduction...

Composing an artwork through a collage of processes is certainly not a new one.

Mick's thesis talking about software being composed of collage processes, visual programming environments vuo, maxmsp, jitter, qc, 4v, ...

Computational creativity, Boden, defining a successful work/evaluation...

Other cut-and-paste generative algorithms?

Lev Manovich's Soft Cinema see https://en.wikipedia.org/wiki/Database_cinema

Computer Graphics -> boundaries of collage become less clear: https://www.youtube.com/watch?v=fw3XyOyl47Q

Space of collage: Andruid Kerne's Collage Machine; I/O/D's WebStalker; Mark Napier's Riot;

Aaron Hertzmann; SIGGRAPH community;

Ben Bogart's Dreaming Machine

Diemo Schwarz's CataRT; Nick Collins's BBCut and klippAV

1.4 Goals

This thesis attempts to work towards a better understanding on the representations we build to support our ongoing scene perception. Working in both auditory and visual modalities, the approach of this thesis is to first motivate a theory, computationally model it, and then work through an arts practice in order to better understand the shortcomings and successes of the model. As designing a computational model of our perceptual systems within even one modality is a significant task, the goal of this model is not to accurately represent all possible levels of processing, but rather to focus on earlier levels thought to be unconscious or inaccessible to our conscious awareness. In the following chapters, numerous literature will be reviewed that has motivated such a representation. Within audition, we will see this representation is defined by *streams*, and within vision, *proto-objects*. Despite their theorization, little attempt has been made to understand what an entire auditory or visual scene would sound and look like to our conscious awareness if represented in such a way.

Encode only parts of a scene that are likely to attract attention... motivate attentional model for dynamic content...

Develop representation of audio and visual corpus that affords simple interaction to produce different styles...

Fragments of a collage require precarious balance between what is identifiable, i.e. how discernible it is as the original source, and what can be composited, i.e. how it can fit within the greater context.

Difficulty of evaluating something subjective; not impossible; can measure performance of speed; art critics...

Model is beyond scope; more of a framework that can be expanded.

1.5 Overview

This thesis investigates a computational method for audiovisual scene synthesis, an automated collage generation where the units of the collage are based on psychologically-motivated representations in perception. To produce such a collage, two basic computational processes will need to be developed: encoding, or storing learned representations into memory; and decoding, or matching stored representations to the ongoing sensorial world. The resulting outputs created by these models attempt to demonstrate how the current stream of audiovisual information coming to a participant could be interpreted by our own perceptual systems. After developing the model for both sound and visual mediums, two artistic mediums will be explored in the final practice chapter: an augmented reality and a series of short films.

This thesis will work through a modular approach. At first, it may seem that a single architecture may suffice for both audio and vision. However, these are fundamentally different domains that require individual treatment. This means the thesis will first focus on developing scene synthesis for audio and then move on to developing another framework for vision. Finally, the outputs of the two modalities will be combined, though still as separate modular frameworks in the hopes that their deeper connections (e.g. cross-modal and multisensory influences) could be explored in the future.

To begin either module (i.e. audio and vision), a conceptual framework for unconscious representations in perception will be developed based on theory in behavioral psychology and electrophysiology (Chapter 2 and 5). This conceptual framework will then be modeled computationally (Chapter 3 and 6) and developed into a scene synthesis framework for the purposes of developing automated collages (Chapter 4 and 7). The two modules will later be combined within an augmented reality and film (Chapter 8). Finally, concluding thoughts on the contribution of this work, limitations, and future will be discussed (Chapter 9).

Chapter 2

Conceptual Framework for Building Unconscious Audio Representations

Contents

2.1 Introduction

Our perception of the world is comprised of objects, events, and meaningful entities. Yet, the sensory information we use to constitute these percepts are built from physical signals that have no meaning by themselves. The challenge our perceptual machinery must face is associating this noisy incoming stimuli to meaningful entities we have already learned so that we may identify and utilize them. This challenge was perhaps first described in the work of 19th-century empiricist, Hermann von Helmholtz, who suggested that the mind evaluates sensations through "unconscious inferences" which combine the stimulus of the world with prior notions of it to form our final perception. This process can also

be summarized by two basic coding procedures: *encoding*, or sensorial input as it is represented in our brains, and *decoding*, or the perceptual experience as we interpret it.

To understand how coding processes may occur, numerous theories have proposed a representational framework for perception. Such theories generally described the encoding of the world within our brains using representations that are inaccessible to our conscious awareness, similar to the prior notions in Helmholtz's inferences. As we do not have direct access to them, what might a representation supporting perception look like? What can they explain, and what do they lack the ability to explain? And how are they formed? It is the aim of this thesis to build a better understanding of questions such as these through an arts practice focused on computational approaches to collage processes. Within this collage-based practice, the units being assembled will be modeled based on the theoretical foundations laid in this chapter.

This chapter is thus dedicated to (1) reviewing what is known about how representations in auditory perception may be coded and (2) describe a basic conceptual framework for their encoding and decoding using computation. As the number of approaches to understanding perception are incredibly vast with many non-overlapping research areas, this review will sacrifice brevity for more depth in fewer areas of research, primarily describing research making use of electrophysiological techniques within cognitive neuroscience. It should be noted that this thesis does not aim to replicate or extend these findings, as it does not use electrophyiosiological methods in its practice. Rather, the intention is to only review this literature to motivate a few key concepts that will inform the implementation of a computational model.

In this chapter, we first begin describing the problem of describing objects within auditory perception in order to understand what is entailed by a model of its processing. From here, we will review a presiding theoretical approach for understanding objects: auditory scene analysis. Within this theory, objects are known as streams, and do not necessarily relate to the objects we perceive visually. We then discuss some key literature using electrophysiology which provide supporting evidence for brain-based representations of stream formations. After reviewing this literature, we will be in a position to outline 4 key components describing a general conceptual framework for the implementation a computational scene analysis model. These will include: event detection, segregation, integration, and template matching. We will develop these components

conceptually in this chapter, model them computationally in Chapter 3, and use them in practice in Chapter 4.

2.2 Background Review on Auditory Perception

2.2.1 Defining the Auditory Object

We are capable of selectively listening to one of multiple sounds in an environment cluttered with simultaneous events, a feat known more colloquially as the cocktail-party problem (McDermott 2009). Though we do not have eyes to move to different parts of an auditory scene, the cochlea of the inner ear can effectively break down the set of pressure waves coming to the ear which we perceive as sound. These waves initially form a complex time-varying set of frequencies and are broken down into a set of narrow, "critical bands" (Fletcher 1940). From the energy in these bands, understanding how we represent and understand scenes within the auditory modality has been the main challenge of neuroscientists and psychologists investigating auditory perception.

Should probably discuss the physiology of the cochlea to auditory cortex a bit more.

Certainly one challenge has been to define what the entities that compose the auditory scene may be described by within the auditory modality. As Winkler points out (Winkler 2010), the notion of an object is highly guided by our visual experiences, and even the Merriam-Webster Dictionary defines object as, "something that may be seen or felt" (as quoted in (Winkler 2010)) or as the Oxford English Dictionary puts it, "something placed before or presented to the eyes or other senses" (as quoted in (Griffiths 2004)).

Griffiths notes that objects perceived must originate in the world, or else they are labeled as hallucinations or "errors" in processing in the brain. Considering this fact alone might lead one to consider the auditory object as information leading to the *source* of the object in the external world. However, as Griffith continues to elaborate, the auditory object may also characterize information of an *event* in the external world, and not necessarily provide information leading to its source or discrimination from other aspects in the current environment. As an example, consider a voice in a crowd. The source may be recognized as a particular speaker, lending information to the source

of the sound and even where the sound occurred in space by matching to one's visual knowledge. However, instead of the speaker, the vowels of the speaker may also be perceived, thus representing not the source of the speaker, but the auditory patterns of changes produced by different possible sources and in different possible environments (Griffiths 2004).

Perhaps a more useful definition of an object or of object-ness entails the capability of the brain to represent, attend, and understand complex and dynamic stimuli across what Marr in visual terms understood as *variances* in a representation (Marr 1982). In other words, the perceivable object must be represented by some separable aspect of the environment. It is perhaps due to the complex nature of defining objects in audition that the majority of the research in the last 40 years attempting to discover these separable aspects has focused on simple un-naturalistic acoustic scenes. These scenes are generally composed of sine tones where simple physical parameters such as the tone's pitch or its temporal frequency may be altered. Across modalities, this method of research has been denoted as "psychophysical", as the aim is to investigate the relationship of varying properties of a stimulus along one or more physical dimensions to a subject's experience or behavior.

2.2.2 Auditory Scene Analysis

A seminal starting point into the study of auditory representation comes from psychophysical research in "streaming". Van Noorden in 1975 demonstrated the bistability of perception when listening to alternating tones composed of sine waves (van Noorden 1975). Depending on the tone's temporal frequency of onset, ΔT, and difference in pitch, ΔF, participants would perceive one of three possible scenarios: two separate streams (A tones and B tones, the *segregated* percept); one stream (A and B tones, the *integrated* percept); or a mixture of streams (the *ambiguous* percept). The integrated case is mostly demonstrated for a low ΔF, the segregated case for high ΔF and/or ΔT, and the ambiguous case for values in between. Van Noorden describes the border between ambiguous and segregated percepts as the *temporal coherence* boundary.

Defining this boundary in greater detail has been the focus of research in the field of Auditory Scene Analysis for the last 30 years, a term coined by Albert Bregman (Bregman 1990). Approaches in Auditory Scene Analysis argue that the perceptual

organization of an auditory scene is represented by a decomposition into pre-attentive auditory streams. These streams are thought to be pre-attentive as we do not have conscious awareness of them. However, once a stream is selected by attention, or pops out demanding our attention, it enters our awareness. For instance, if we are able to segregate sounds within a complex natural environment into one where we can describe a person's voice and what they are saying, then we will not be able to report information about other sounds that occurred at the same time in the environment. In other words, the segregated stream that we can report has been *foregrounded* by attention, bringing it to conscious awareness, while the stream we are unaware of has been *backgrounded*.

Bregman argues that before streams are foregrounded, they are encoded by one of two formations: (1) rapidly available cues from primitive, low-level characteristics such as pitch, intensity, and spatial location; and (2), schema-driven integration of sensory evidence where schemas are defined by Gestalt-like regularities such as similarities, differences, common-fate, or continuity in frequency information from a continuous signal. In the literature, these two formations have also been denoted, respectively, as *simultaneous* and *sequential* grouping strategies (Winkler 2009). Simultaneous cues suggest that sounds are likely to be segregated if their physical characteristics despite their context are significantly varied. As a result, these cues are independent of the attention of a listener and are purely based on the cues in the environment. In contrast, sequential cues require a listener to have formed the notion of a representation, as the notion of regularity is based on previous listening. These cues are more favorable in environments where very noisy conditions require precise knowledge of the source to attend to, thus resolving any ambiguities by using existing representations to segregate sources (e.g. the classic "cocktail party problem").

Bregman's theory provides a basis for understanding the formation of auditory objects as they may appear within our perception as unconscious representations. Rather than the notion of an explicit auditory object that must be defined by its physical characteristics, Bregman's streams provide a perceptual basis for representation as determined during auditory scene analysis. Objects in the scene analysis sense are therefore based on the active analysis of a scene, i.e. what can be segregated at the given time based on the aforementioned cues, rather than any fixed description of an entity in the world.

2.2.3 Electrophysiological Evidence for Auditory Scene Analysis

Research supporting the encoding of both of Bregman's theoretical formations of streams has seen great support through investigations making use of *electroencephalograph* (EEG) recordings. Before discussing research in electrophysiology, it is important to have a basic background in EEG techniques.

EEG recordings, first described by Hans Berger in 1929, attempt to non-invasively infer brain activity by measuring a time varying signal of the electrical activity on the scalp of the head. One of the benefits of EEG comes with its high temporal resolution often on the order of milliseconds. To localize or disassociate changes in electrical activity in specific brain regions, more electrodes may also be used (upwards to 128). As the raw electrical signal coming to electrodes are very low amplitude, the signal is first amplified. However, the incredibly complex nature of the brain means that looking at the amplified signal may not be very useful by itself. Thus filtering methods are often performed in order to limit the bands of frequencies within the recorded signal. From here, understanding the relationship between the recorded signal and the external world can still be very tricky, as the world is full of events, and their relationship to the signal may only be present at very different times.

Thus, a presiding technique of investigating brain activity using EEG in relationship to an external stimuli in cognitive studies is to timecode the recorded EEG in relationship to a specific event, a technique also known as the *event-related potential* (ERP). ERPs are simply the measured neural activity as measured in any electrode time-locked to the presentation of a stimulus. By having many presentations, often denoted as trials, of the same stimulus, the average wave can be computed across trials for each electrode and then studied in greater detail.

ERPs are generally described in terms of specific components that contribute to the averaged signal. The nomenclature of some of the more basic components are based on their polarity (positive deflection, P; negative deflection, N) and how long after a stimulus they occur (e.g. N1 or N100 means a negative deflection around 100 ms after stimulus presentation).

2.2.3.1 N1: Unconscious change detection

One of the first well studied components within auditory processing is the N1 component which has been shown to be elicited during the onset of a stimulus (usually from silence to sound), providing evidence of a brain-based representation for unconscious change detection in auditory scenes. Perhaps first described in 1939, Davis described a sound-evoked change in the EEG recordings of the waking human brain (Davis 1939). The onset of a tone evoked a negative wave 100-150 ms after onset near in the frontal-central locations of the scalp lasting for nearly 100 ms. This was again described in a study in 1965 in which a flash of light or sound would proceed another flash of light or sound (Sutton 1965). Sutton et al. describes a negative potential peaking at 110 ms (N1) at the vertex of the scalp and a late positive potential at about 300 ms for sound and 340 ms for light (P3).

Hillyard et al.'s 1973 study expanded on these previous results, showing that the N1 component is of even greater magnitude when a subject's selective attention is required in detecting a target tone (Hillyard 1973). In their study, a subject was presented random tones in either ear with very short irregular inter-stimulus intervals (ISIs, referring to the ΔT), with higher pitch tones in the left than the right ear, and with randomly placed target tones with a slightly higher pitch in both ears. The subject's task was to count the randomly placed tones in a target ear, meaning attention should have diverted to the target ear ignoring the other ear. As a result, Hillyard et al. was able to demonstrate that the N1 component measured at the vertex (around 60-70 ms in most subjects) was of higher amplitude for events in the attended ear. They suggested their results reflected an enhancement of the N1 component. They also demonstrated an evoked late positive component peaking at 250 to 400 ms (P3) in the attended ear only when the target tone was presented, indicating the selective recognition of the target tone from which any subsequent cognitive or motor activities could follow (e.g. counting of the tones).

2.2.3.2 PN: Conscious template-match

However, their result was later reinterpreted in another study which used longer ISIs of 800 ms (Näätänen 1978). Their study demonstrated that a new component, which they denoted the *processing negativity* (PN), emerged exhibiting the effect of selective attention, rather than any enhancement of the N1. This effect was found as a slow negative

shift appearing at 150 ms and lasting 500 ms, rather than an earlier enhancement of the N1. Their results indicate that the increased amplitude of the N1 in the attended ear could have been due to the short ISIs causing an overlap between the PN and N1 components, though even this explanation is still open to interpretation as a genuine enhancement of the N1 component can occur for strongly focused selective attention (Näätänen 1978; Hillyard 1983; Näätänen 2011). In any case, it seems the N1 is well described by the literature as an onset detector for auditory stimuli thus providing the awareness of a stimulus. The PN may then provide an attentional-trace function, and be evoked during a template match during an attentional task requiring stimulus selection. In other words, the PN shows supporting evidence for the brain's ability to retain representations and match them to incoming stimuli during a search task in a very short time.

2.2.3.3 MMN: Unconscious segregation via sequential cues

In 1976, another now well-known combination of components, the N2-P3a (Snyder 1976), was demonstrated. Unlike the N1 component which is based on the simple onset of a stimulus, the N2 was generated whenever a *deviant*, or infrequent stimuli, was presented. This difference became even greater with larger differences in Hz between the deviant and the *standard*, or the stimuli that was presented more frequently. For example, a tone would be presented with an 80% chance of a standard 1000 Hz tone versus deviants of 1020 or 1040 Hz at 20%). This entailed the notion that some temporal regularity had been represented in order to compare it to the deviant stimuli. Unlike the PN, the deviant did not need to be known or recognized based on prior training, or require selective attention, though its characteristics could effectively demand attention, hence evoking a P3a component peaking around 258 ms. In contrast, when the subject was actively attending, a later P3 component denoted the P3b was evoked, peaking around 378 ms, still with the N2 component preceding it.

The N2 component has since been very well studied, as it entails that we are able to represent the regularities of sounds and compare them to the ongoing environment evoking N2s whenever the regularity is violated even outside of a given task. In particular, a re-interpretation of the 1976 study in 1978 showed that when subtracting the deviant trials from the standard ones, the remaining N2 spike (known as "N2a", thus

making the original component combination now more commonly referred to as "N2b-P3a") demonstrates the brains capability to detect irregular changes in what the authors originally called "template mismatch" (Näätänen 1978). Importantly, these changes are not dependent on the deviant stimuli's features itself, but only in relation to the temporal regularities represented in the standard tones. In other words, the template being matched to is likely an unconscious one, as the participant was not searching for it, nor were they aware of the deviant beforehand. Rather, the standard set a predictive or inferential basis for any future irregularities.

The remaining subtraction originally defined as the N2a was later denoted as the *mismatch negativity* (MMN), and has been shown to appear anywhere from 100 - 250 ms after onset of a deviating stimuli (Näätänen 1978; Näätänen 1987; Näätänen 2007a; Campbell 2007; Garrido 2009; Näätänen 2011). Such relational negativities were also capable of appearing within the time-frame of the N1, originally leading it to be classified as a part of the N1 (Näätänen 1987). Though only recently has it has been discriminated from the N1 (Campbell 2007). As well, the MMN's counterparts in other techniques for studying brain activity has also been demonstrated, including equivalents for magnetoencephalograph (MEG) (Hari 1984), optical-imaging (OI) (Rinne 1999), positron emission tomography (PET) (Tervaniemi 2000), and functional magnetic resonance imaging (fMRI) (Celsis 1999).

Research demonstrating the window size of temporal integration has shown even for ISIs up to 350 ms, deviants in sequential tones can elicit a MMN during passive listening exercises (e.g. reading a book while headphones deliver the tone sequences) (Tervaniemi 1994; Tervaniemi 1997a). However, this window of integration can be prolonged to up to 500 ms during active listening, meaning attention is capable of prolonging the window of integration (Kanoh 2001).

The standard tone must also be established through repetition before a MMN can be elicited by a deviant (Cowan 1988). Interestingly, the amplitude of the MMN grows larger as the number of repetitions preceding the deviant gets higher (Sams 1983). This entails that the MMN represents a measure of the deviance established by the variances of the standard. Further, it provides evidence suggesting that temporal regularities can be represented by the brain. Research expanding on the MMN further has shown that more than low-level attributes such as pitch, temporal frequency, intensity, or duration,

the same pattern of behavior is exhibited for a wide-range of stimuli even including higher-order or more complex violations such as timbre changes (Tervaniemi 1997b), grammar in mother-tongue sentences (Näätänen 2001), or temporal sequences such as violations in patterns of ascending/descending tones (Näätänen 2007b; Garrido 2009; Shamma 2010).

2.2.3.4 ORN: Unconscious segregation via simultaneous cues

Evidence for simultaneous grouping cues has been evidenced by research looking at the neural responses during listening of complex tones composed of inharmonicities (Alain 2001; Alain 2002). Participants listened to a complex tone where all of the tones were either harmonic (i.e. all tones were multiples of the fundamental tone) or inharmonic tones (i.e. one tone was not a multiple of the fundamental). The tones were also matched for perceptual loudness to ensure any differences were based on the complexities of the tone rather than a separate loudness cue. Participants were given both an active and passive task. For the active task, participants had to press a buzzer if they heard two separate tones. For the passive task, participants read a book. Participant's were at about 20% chance of reporting that they heard 2 sounds when the mistuning was at 2%. This increased to 50% at a 4% mistuning, and 80% for an 8% mistuning. The ERP data for active listening trials revealed an N1-P2 complex peaking at 110 and 200 ms over the central and frontal electrodes, followed by a late positive P3b component in the parietal regions. Subtracting the ERPs to harmonic trials from inharmonic trials revealed that participants elicited a negative ERP component at 150-180 ms deemed the *object related negativity* (ORN) in both active and passive listening. This was followed by a late positive peak at 350-400 ms (P400) during active listening only. The finding of an ORN in both active and passive listening conditions suggests that the detection of inharmonicities is an automatic process, though can be influenced by attention, as its amplitude in the active case was significantly larger than the passive one (Alain 2001; Alain 2002).

2.3 Conceptual Framework

Research behind the neuronal basis for memory representations, both conscious and unconscious, is greatly evidenced by research in electrophysiology. This body of research suggests that the ORN and MMN provide a neural trace of an unconscious memory representation, the PN as a trace of a conscious one acting through selective attention, and the N1 as evidence for the brain's capacity for an unconscious stimulus onset detector. In particular, research demonstrating ORNs provides evidence for the brain's capability for segregation via a simultaneous cue such as through inharmonicities, while the MMNs provides evidence for segregation via a sequential or schema-based grouping cue. The schema for MMNs are defined by the temporal regularities of the standard, leading any deviances from this regularity to be segregated (with enough deviance). With such deviants, the entire stream is not integrated into one concept, but segregated into two through an explicit measure of its deviance.

Extending this research into a basic computational model requires at least:

1. **Event Detection**, such as evidenced by the N1, where an ongoing auditory stream is marked by an event boundary, allowing it to be encoded into memory. This component is only be necessary in models that do not have explicit event boundaries labeled for them;
2. **Segregation**, as evidenced by the MMN and ORN, where the auditory stream is separated into foreground and background streams based on temporal irregularities;
3. **Integration**, the encoding process which stores the result of event detection, or if presented with enough deviance, the segregated auditory stream, into memory (whether or not segregation can occur before integration is certainly debatable, e.g. (Sussman 2005), though for the purposes of this simplified framework, allowing segregation to occur before integration will have to suffice);
4. **Template Matching**, functionality as evidenced by the PN, where selective attention is capable of recognizing a stream as one that is the target of attention.

Evidence demonstrating the elicitation of an N1 component suggests that as little as 60-70 ms are required before the brain is capable of detecting the onset of a stimulus event, even when engaged in another task. This component demonstrates the brains remarkable ability to unconsciously detect a change in the environment. It takes almost another 250-300 ms before we can become conscious aware of the stimulus onset, i.e. report the actual stimulus onset, as demonstrated by the P3 wave. As a result, it seems that for developing an online analysis of an auditory scene, it is first essential to organize a real-time auditory stream into one chunked by discrete event boundaries.

However, the literature discussing N1 elicitation does not seem to be clear on exactly how an onset may be modeled by the brain. Each of the experiments reviewed above were designed with a streaming paradigm where short evoked tones were presented to a participant, meaning each tone was itself an event likely to be an onset. Thus, each tone evoked an N1 response, as was demonstrated by averaged event-related potential measurements. However, the real world is not aligned by events for us, making the notion of an event evoking an N1 outside of the psychophysical world of simple tones an unclear one. How then does the N1 represent an event and determine when one has occurred? Current approaches based on averaging ERP trials may not be able to answer such a question, as the notion of a change in the environment implies a violation of expectation. By presenting the same auditory scene to a participant (e.g. in trials), the violation of expectation or the evocation of an N1 component, may become less likely. In other words, the repeated presentation of the same stimulus may entail a more regular representation of the entire scene, making any deviations within it less "irregular".

Despite these issues, considering the groundwork laid by Bregman's Auditory Scene Analysis, one may be able to begin to define an onset detector based on the strategies of simultaneous and sequential grouping cues. Though these cues up until now have been provided as evidence of cues for segregation (i.e. as evidenced by the ORN and MMN) rather than eliciting an N1 component, a re-interpretation of this theoretical foundation with the presented literature may allow us to apply the same framework for defining events in a continuous, real-world setting, i.e. defining onsets. As a result, the ORN, MMN and N1 may share very similar strategies. The major difference seems to be that the ORN/MMN are time locked to an onset, providing a basis for foreground segregation, i.e. the sonic environment before and after the N1 response, whereas the N1 may likely

be based on an automatic time window. It is unclear whether the N1 component is also activated for changes in the segregated foreground as may be represented by regularities required for the ORN/MMN, or if the N1 is only defined for changes in the entire acoustic scene. It is likely though that the N1 may be able to consider any irregularity as a result of cues from either simultaneous or sequential grouping strategies, as it must act before both the ORN and MMN. As well, given the research demonstrated thus far, it is also likely that this window is very short, e.g. within the timespan of evoking P3s, given the capability for detecting ISIs of up to 350 ms, though it is likely that this window size is also based on task/attentional demands.

Another fascinating feature of the brain is its rapid capability of detecting a known sound, or a template sound, as illustrated by the PN. What representation a template takes, or the time required before evoking a PN is still an open question, as the literature presented here did not look into complex environmental sounds. However, the notion of a template match functionality does give us some evidence in proceeding towards a basic computational model. Namely, it is likely that during selective attention tasks, the evocation of an N1 component must precede any detection processing of the target. In other words, before knowing if a known sound has occurred in the sonic world, the onset of that sound must be detected via an N1 component. Once the N1 is evoked, a PN component may be elicited if that particular event is detected as a match.

2.4 Discussion

The literature presented here focused on evidence from cognitive neuroscience demonstrating some evidence towards the representations supporting auditory perception. Specifically, this research primarily made use of electrophysiological recordings of subjects listening to psychophysical acoustic scenes composed of simple tones, and made use of repeated trials of the same scenes in order to measure ERPs, the average of simultaneous EEG recordings event-locked to a stimulus presentation (i.e. the individual tones).

Taking this literature towards a scene synthesis, a collage where the units of the collage are based on similar units of representation processed by our perceptual systems, requires a computational implementation capable of representing an arbitrary complex

acoustic scene into a set of units that may be manipulated. However, the literature in developing an understanding of auditory scene analysis does not deal with complex acoustic scenes but is focused on unnatural psychophysical scenes composed of regular alternating tone sequences. Only a few recent attempts have been made towards understanding brain bases for complex scenes (Teki 2011; Teki 2013), where a "stochastic figure-ground stimulus" (SFGS) was created. However, these studies are still modeled by unnatural sequences of randomly evoked collections of pure tones, and not modeled by natural scenes that have defined the literature as a whole, i.e. the cocktail party. As a result, many assumptions are still made about the nature of what constitutes the auditory object.

In these recent studies, the notion of figure is often placed on any aspect of the environment that is temporally coherent (Shamma 2011; Teki 2011; Teki 2013), assuming that these are the "foregrounded" aspects that we are likely to encode in an environment, as demonstrated by MMNs. The authors assume that given the complex and unrealistic nature of their scenes, it is likely that a listener perceives an object consisting of integrated streams rather than segregated ones, as there is no prior notion to help segregate the stream. That is, until a temporally coherent figure appears, providing a cue for a sequential segregation. However, one important factor seems to be missing when considering any temporal regularities within a stream as being attentional "foreground", and that is the intricate relationship with the temporal *ir*regularities in the stream. As the conceptual framework I've laid out in this chapter demonstrates, any temporal regularity is only defined within the onset of a stimulus (e.g. the N1). The onset itself must also be capable of also understanding irregularities in an environment. As a result, any temporal regularity is only defined within the ability to define an event.

Therefore, there is a multiplicity of regularities that must be considered: the first is defined by the onset detector (or N1), and is likely based on whole spectrum irregularities; the second is a segregation into foreground/background, using the time before and after an onset to define the basis for segregation. As a result, a temporal irregularity within the whole spectrum must first be evoked before any temporal regularity within the segregated streams can occur, as there must be some basis for defining the event boundary. Further, the notion of the temporal regularity past the irregularity must be sufficiently deviant from the preceding regularities before it is segregated, otherwise the entire stream is integrated.

Considering this framework within SFGS stimuli, it is likely that many onsets are created during the time course of listening, even before the regular figure appears, since, by definition, the stochastic nature of the stimulus means there is a high chance of hearing something irregular. Previous research has demonstrated that a standard must be set before a deviant can segregate the stream, albeit in simpler stimuli (Cowan 1988). As a result, it is likely that the statistics of the complex tones in SFGS are of high entropy, or high uncertainty, meaning many onsets could be evoked. Within the onset then, it is likely that the time frame before and after the onset are composed of radically different tones, as they are randomly created. It is likely then that many segregations can occur during the time course of listening, and not just one, during the highly regular set of tones, as the authors suggest.

2.5 Conclusion

The evidence presented thus far demonstrates the brains remarkable capability for detecting events in the sonic world and discretizing the continuous amalgamation of sound into units of processing where updates in representations may occur, or selective attention may discover a match. Much of the literature to date has focused on either simple alternating tone sequences or random complex tone sequences, meaning many inferences must be made about how such evidence can be applied within complex natural scenes. Nevertheless, some understanding can be gained towards developing a practical computational model capable of producing either segregated or integrated auditory streams.

The conceptual framework presented here aims to build event-related units through a simple computational model beginning with an onset detection modeled after the N1 component. From here, one of three processes may occur: matching, integration or segregation. Matching occurs when a search task is involved and is modeled by evidence of the PN. Otherwise, either integration or segregation occurs, defining the irregularities of the environment as it relates to the ongoing regularities. This is in contrast to the N1, which likely requires some notion of regularities (as it must be evoked by an event that the sonic world has caused), but likely does not require segregation. WIth enough irregularity, a segregated stream may be defined, otherwise the stream is integrated.

Future models should consider, at the very least: the numerous influences of stream

formation by attention (e.g. (Shamma 2011)), whether integration should occur before or after segregation (e.g. (Sussman 2005)), categorical effects on low-level neuronal representations (e.g. (Samson 2010)), how encoded summary statistics from contextual influences may effect stream formation (e.g. (Piazza 2013)), and how this research may possibly translate to real-world natural sounds (e.g. (Moerel 2013)). However, any one of these questions are all very active areas of research. As a result, the basic set of components laid down in this Chapter can at least provide a plausible model from which a representation of auditory objects, or streams, may be built.

The presented framework has given us the notion of the auditory unit of perception that can be used within a collage-based practice: namely, the auditory stream determined by auditory scene analysis as an integrated or segregated stream. Further, during selective attention tasks, it has given us a notion of matching through a processing negativity, and that this matching process should occur within the timeframe of an onset. We will develop these concepts into a computational framework in the next chapter, Chapter 3, demonstrating its use as a method for online source separation, browsing of a large corpus of audio, and for classifying auditory scenes. This computational framework will then be developed for use within a collage-based practice in Chapter 4, and combined with a visual collage-based practice producing computational audiovisual collages in Chapter 8.

Chapter 3

Computational Auditory Scene Analysis

Contents

3.1 Introduction

In this chapter, we develop the conceptual framework laid down in Chapter 2 into a computational model capable of two processes: *encoding*, or the sensorial input as it is represented in our brains, and *decoding*, or the perceptual experience as we interpret it. These two processes will be necessary for developing an auditory scene synthesis,

a collage-based synthesis where the units of the collage process are based on psychologically motivated representations thought to support perception. To model these two processes, we need to develop our conceptual framework as described in Chapter 2:

1. **Event detection**: Temporally segments the ongoing auditory stream in order to define the boundaries of processing. This component is only necessary in cases where explicit events are not already defined;
2. **Segregation**: In case an auditory stream is composed of a multiple streams, the mixed stream is split into a foreground and background stream;
3. **Integration**: Encodes the segment by building an internal representation of the stream;
4. **Matching**: Decodes the encoded stream by matching to a known template.

For now, we only work with fixed-length units of sound, and therefore already have explicit event boundaries. In a real-time setting, we would not have this luxury, and would instead need to use event detection to discover the segmented units of sound. In Chapter 4, we will see how combining this framework with event detection affords a real-time experience for encoding and decoding when discussing Memory Mosaic.

In any case, our units of sound must still be encoded, represented, and decoded. If the entire sound stream is to be integrated, then the unit defined by the event detection is encoded without further manipulations. However, if the unit requires segregation, then the unit is separated into a foreground and background stream, where the foreground stream denotes the stream that is integrated as a result of being "selected" by attention. Finally, matching determines whether the current stream matches a previously learned target encoding, effectively allowing us to decode a stream based on prior learning.

In this chapter, we implement three computational methods for encoding and decoding. These are described as (1) a full-frequency and (2) reduced frequency representation both built using a probabilistic variant of non-negative matrix factorization, Probabilistic Latent Component Analysis (PLCA), which describes audio by the latent components that represent the signal, and (3) a Gaussian Mixture Model of one of the most common acoustic feature representations, MFCCs. As we will see, the first two

models attempt to compute a basis decomposition of a time-frequency matrix into a set of components modeled by a set of frequencies with a time-varying amplitude, ideal for cases when segregation is required. The last model on the other hand describes a set of frequencies by its overall spectral shape, taking into account the perceptual bandwidth given to different frequencies. This spectral shape does not vary with time like the first two models, and is composed of only one set of values describing the global shape of the spectrum, making it ideal for cases when segregation is not required.

In order to evaluate each model's performance, we consider a common use-case outside of the context of scene synthesis which requires encoding and decoding routines: classification. In particular, classification is often employed within information retrieval frameworks in order to attempt to aid the browsing and retrieval of content within large archives, a need that has grown due to the growth of digital audio archives. As the notion of classification entails providing a semantic description of an auditory stimulus, it is important to be clear about what this thesis means by a *class*, *scene*, *event*, and *stream*. A class semantically describes either a scene or event. A scene defines a collection of acoustic events associated with that particular physical place. As an example, one would expect to find events classified as "honking", "engine noise", and "talking", in an acoustic scene described as "city street". A likely distinction between the term acoustic scene and acoustic event is that a scene is composed of many events, many of which could be of different classes, while an event is composed of only one particular class. As a result, there is a many-to-one definition entailed by scenes, where many possible definitions can give rise to the same scene, and a one-to-one definition for events, where only one possible class exists. The terminology of a stream enters when discussing a particular perception of that scene. For instance, when listening to an acoustic scene, a stream would by described by what is encoded into memory (i.e. the entire scene, which may be segregated during particular events into foreground and background streams) and the resulting perception would describe what is decoded.

An ideal information retrieval engine should be capable of classifying acoustic scenes, events, and when presented with complex scenes, also determine if simultaneously occurring events can be described. For example, an auditory scene of "party" may also be composed of a "hair-dryer" in the background. This additional class may make it difficult to describe the scene, unless it is capable of also segregating the scene into two classes. An analogous model in vision is one that detects objects in a visual scene rather

than the entire scene itself. The difficulty of classifying the multiple events comprising an acoustic scene is well noted in acoustic event detection literature where classification performance breaks down from 70% for classification of a single event in isolation to 25-40% during mixtures of events or events presented with noise (Temko 2007).

With these distinctions in mind, after developing our models, we turn to evaluating them within the context of auditory classification by using a large audio corpus of labeled acoustic events in a series of experiments: (1) classifying isolated acoustic events, (2), classifying an acoustic scene comprised of one known acoustic event and an additional unknown acoustic event, effectively creating noise, and (3), classifying an acoustic scene comprised of two simultaneous events that are both known. By known and unknown, we simply mean to say that the classifier has or has not encoded knowledge of this class in some way. For testing purposes, this entails performing cross-fold validation on 10 folds, where 9 folds are encoded, and the additional 1 fold must be classified. We will see that both PLCA models outperform the MFCC-based model in cases (2) and (3), correctly classifying the mixture's sources 94% of the time in comparison to 74% of the time for the model built with MFCCs. However, for case (1), i.e. when a scene does not require segregation, all of the models perform with excellent performance. This evaluation serves to highlight how each model may perform within complex auditory scenes composed of multiple events when it is later coupled with event detection in Chapter 4 to create a real-time auditory scene synthesis and finally in Chapter 8 when it is combined with visual scene synthesis.

3.2 Related Work

Due to the amount of information contained in archives and the complexity in pre-processing so much information, the first step in a content-based solution to information retrieval is often to reduce the dimensionality of the data while keeping as many of the perceptually relevant dimensions as possible. In audio, this often equates to looking at the distribution of frequencies that describe a signal (e.g. by taking the Fast Fourier Transform (FFT) of an audio signal) and computing features or a fingerprint which could be used to train models/classifiers. Most previous work in acoustic classification and retrieval makes use of a combination of features described by Mel-Frequency Cepstral Coefficients (MFCCs) and low-level psychoacoustic descriptors (Temko 2007; Guo 2003;

McKinney 2003; Allamanche 2001) such as spectrum power, centroid, zero-crossing rate, brightness, and pitch. MFCCs were first described in a seminal study on automatic speech recognition (Davis 1980) as a perceptually motivated grouping and smoothing of power spectrum bins according to the Mel-frequency scaling. MFCCs can be thought of as a perceptually motivated, reduced, and de-correlated representation of a frequency transform, and are approximations to the overall texture of an acoustic signal. Hence, though MFCCs were originally applied to speech recognition problems, they are also widely used as audio features in the domains of general acoustic events (Temko 2007) and music (Pampalk 2006; McKinney 2003) analysis.

Approaches building MFCC and low-level based descriptors into a large feature vector attempt to depict an auditory scene by a vector of global parameters. Thus distance measures attempting to perform matching and that act on the MFCC feature vector are generally unsuited for describing acoustic scenes requiring segregatation. This is because any measures acting on the MFCC will penalize any deviation from the global feature vector's approximation. In other words, approaches to auditory classifiers using k-means (Harma 2005; Eronen 2006; Allamanche 2001), hidden Markov models (Eronen 2006; Mesaros 2010), support vector machines (Guo 2003), or Gaussian mixture models (Wang 2011; Aucouturier 2007; Pampalk 2006) aim to model the distribution of possible variants of a feature vector rather than the many possible subspaces that may define them.

The MPEG-7 standard (Casey 2001; Manjunath 2002), however, describes a modular approach to understanding the subspaces of such feature vectors by looking at their basis decomposition. In this manner, our approach to modeling segregation most resembles models employing spectral basis decompositions which describe de-correlated features of an acoustic signal using principal component analysis and independent component analysis (Casey 2001; Xiong 2003; Kim 2004), local discriminant bases (Su 2011), matching pursuits (Chu 2009), or non-negative matrix factorization (Raj 2010). However, our approach differs from the MPEG-7 standard's spectral basis decomposition (Casey 2001) as we instead investigate a full-frequency and Mel-frequency decomposition, rather than decibel-power scale or de-correlated features, and further use a recently developed machine learning algorithm for discovering latent components rather than any of the aforementioned models.

In order to compute the basis decomposition of an audio signal's frequency transform, we focus on a recently developed method for latent component analysis based on probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999) called probabilistic Latent Component Analysis (PLCA) (Smaragdis 2006). PLCA has shown great promise for a variety of use cases including source separation and de-noising (Smaragdis 2007b; Smaragdis 2007a), online dictionary learning for source separation (Duan 2012), riff-identification (Weiss 2011), polyphonic music transcription (Benetos 2011), and classification during mixtures (Nam 2012). However, no detailed investigation of PLCA into the performance and applicability of classification for isolated or mixtures of classes in comparison to a standard MFCC model exists. Further, previous investigations of PLCA for source separation and classification only make use of the full-frequency spectrum in a limited number of classes, and PLCA's applicability for reduced frequency representations is still unknown. We therefore investigate the performance of 2 models built using PLCA (full-frequency and reduced) through 3 experiments in acoustic classification while comparing it to a classifier based on the well known Mel-Frequency Cepstral Coefficients (MFCCs): (1) classifying isolated acoustic events, (2), classifying acoustic events in the presence of an untrained event, effectively creating additional noise, and (3), classifying acoustic scenes composed of two simultaneous acoustic events.

3.3 PLCA

The underlying basis of the standard PLCA model was first proposed in (Hofmann 1999) as a probabilistic extension to Latent Semantic Analysis (LSA) called probabilistic Latent Semantic Analysis (pLSA). Singluar Value Decomposition (SVD) based LSA methods and their non-negative counterpart, Nonnegative Matrix Factorization (NMF), both aim to describe a matrix using orthogonal projections with a standard Frobenius-norm. This assumption penalizes the true density of data in cases where the l2- or Frobenius-Norm are unable to describe the data (i.e. non-Gaussian data).

PLSA instead describes a factorization in terms of a mixture of the latent components that give rise to an observed multinomial distribution. Recovering the latent structure using iterations of Expectation-Maximization (EM) in order to estimate the maximum likelihood gives a number of benefits on the latent components describing the data. First, being a probabilistic model, the component weights and likelihoods

are easily interpretable in terms of the amount of data they describe, whereas in SVD based methods, the number of singular values needed to describe the data have to be analyzed ad-hoc. Second, by using the data's own distributions in performing the maximum likelihood updates, the assumption of additive-Gaussian data is no longer made, and instead the Kullback-Leibler divergence between the empirical data and the model is minimized. Third, employing model selection allows one to iteratively determine the appropriate number of components required to explain the data (Mital 2012), whereas in LSA and NMF based methods, no measure of likelihood is obtained. Lastly, though we do not make use of this advantage we mention it here for completeness sake, the symmetric nature of the probabilistic model allows for factorizations in higher dimensions leading to a probabilistic variant of non-negative tensor factorization.

Though (Hofmann 1999; Hofmann 2001) did not describe the model in terms of audio, it was not long before it was applied to audio and demonstrated as a source separation algorithm (Smaragdis 2006). It was later greatly enhanced to include a number of extensions including shift-invariance and sparsity using an entropic prior (Smaragdis 2007a). We simply make use of the basic formulation of a probabilistic latent semantic/component analysis described in (Hofmann 1999; Smaragdis 2006) and describe it in terms of an input frequency versus time matrix $\mathbf{X}$ as:

$$X_{f,t} = p(f,t) \approx \sum_{i}^{N} p(k_i)p(f|k_i)p(t|k_i) \tag{3.1}$$

where $p(f,t)$ describes the frequency $f = 1,...,R$ versus time $t = 1,...,C$ matrix as a probabilistic function, k_i is the i^{th} latent component up to N components, $p(k_i)$ the probability of observing the latent component k_i, $p(f|k_i)$, the spectral basis vector, and $p(t|k_i)$, the vector of weights over time. Thus, the spectral basis vectors and temporal weights are described as a multinomial distribution, where the actual density of the data describes the frequency and time marginals. The spectral basis vector is intuitively understood as the distribution of frequencies describing a particular source and the temporal weights as the envelope of sound of the source across time. When multiplied together with their mixing weight, $p(k_i)$, they produce a 2D matrix of the source over time, while adding all N components produces the approximation to the original matrix X.

Formally discovering the marginals requires computing their maximum likelihood estimate (MLE). This can be done iteratively through a variant of the Expectation-Maximization (EM) algorithm, a standard technique for estimating the MLE in latent variable models. The E-step estimates the posterior contribution of the latent variable k:

$$p^{(t)}(k_i|f,t) = \frac{p(k_i)p(f|k_i)p(t|k_i)}{\sum_j^N p(k_j)p(f|k_j)p(t|k_j)} \tag{3.2}$$

The M-step then re-estimates the marginals using the posterior distribution computed in the E-step:

$$\begin{aligned}
labeleq : plca - mstepp^{(t+1)}(k_i) &= \sum_{f,t} p(k_i,f,t) && (3.3)\\
&= \sum_{f,t} \left(p^{(t)}(k_i|f,t) \frac{p(f,t)}{\sum_{f,t} p(f,t)} \right) && (3.4)\\
p^{(t+1)}(f|k_i) &= \sum_t p(f,t|k_i) && (3.5)\\
&= \frac{\sum_t p^{(t)}(k_i|f,t)p(f,t)}{p^{(t)}(k_i)} && (3.6)\\
p^{(t+1)}(t|k_i) &= \sum_f p(f,t|k_i) && (3.7)\\
&= \frac{\sum_f p^{(t)}(k_i|f,t)p(f,t)}{p^{(t)}(k_i)} && (3.8)
\end{aligned}$$

Practically, one can use a fixed number of iterations of EM and assume convergence, though testing for the change in performance avoids the risk of over-fitting (Hofmann 1999) (e.g. using Least-Squares or Kullback-Leibler Divergence).

The basic algorithm is simple to implement and is shown as functional Matlab/Octave code in Program 1:

Program 1 Matlab/Octave code for PLCA

```
function [f,t,k] = plca_basic(X,K)
% Initialize
[M,N] = size(X);
f = col_normalize(rand(M,K));
t = row_normalize(rand(K,N));
k = col_normalize(rand(1,K));
i = 1;
maxiter = 100;
while i < maxiter
   % E-step
   R = X ./ (f * diag(k) * t);

   % M-step
   f_p = f .* (R * (diag(k) * t)');
   t_p = (diag(k) * t) .* (f' * R);
   k_p = sum(t_p, 2);

   % Normalize across components
   f = col_normalize(f_p);
   t = row_normalize(t_p);
   k = col_normalize(k_p);
   i = i + 1;
end

function X = col_normalize(X)
X = X ./ repmat( sum(X, 1), size(X, 1), 1 );
function X = row_normalize(X)
X = X ./ repmat( sum(X, 2), 1, size(X, 2) );
```

3.4 Computational Auditory Scene Analysis Models

In the following section, we describe three approaches for computationally implementing our conceptual framework as described in Chapter 2: (1), a Gaussian Mixture Model of Mel-Frequency Cepstral Coefficients (MFCC Model), (2), a Probabilistic Latent Component Analysis of a short-time Fourier frequency transformation (PLCA Model), and (3), a probabilistic Latent Component Analysis of a short-time Mel frequency transformation (Mel-PLCA Model). These models aim to perform "Segregation", "Integration" and "Matching" of auditory scenes.

3.4.1 MFCC model

The first model we built encodes an auditory scene by first representing any units of sounds as a vector of Mel-frequency Cepstral Coefficients (MFCCs). For the purposes of classification, we define its acoustic class using a generative probabilistic model, a Gaussian Mixture Model (GMM).

MFCCs The basic algorithm for computing MFCCs is summarized below:

1. Apply a Hanning window function to the input audio segment and take the discrete Fourier transform

2. Warp the absolute power spectrum into M triangular sub-bands, spaced equally on the Mel-frequency scale with 50% overlap. The following approximate formula describes a frequency on the Mel-frequency scale given an input linear frequency:

$$mel(f) = 2595 * \log_{10} 1 + \frac{f}{700} \tag{3.9}$$

   Use this mapping to warp the power spectrum to the Mel-scale and compute the energy in each sub-band as follows:

$$S_m = \log \left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \right) \tag{3.10}$$

where H_m are the filter-banks described by the Mel-frequency scale.

3. Finally, after taking the log result, compute the discrete cosine transform to obtain the first C MFCCs:

$$c_n = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} (\log S_m \times \cos[n(m - \frac{1}{2})] \frac{\pi}{M} \tag{3.11}$$

and $n = 1, ..., C$, where C is the number of coefficients to return (discarding high-frequency coefficients), and M is the number of triangular sub-bands.

For our purposes, we use a standard decomposition of $M = 40$ triangular bands and keep $C = 13$ coefficients.

GMM In order to build the encoding of an acoustic class, we assumed the distribution of each class's MFCC vectors could be described by a multivariate Gaussian, i.e., for each class $k = 1...N$,

$$p(\mathbf{x}|k) \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{3.12}$$

where μ_k is the mean vector of MFCCs, and Σ_k is the full covariance matrix. In order to combine each of the N classes into a mixture of Gaussians, we simply assumed a prior equal weighting on the mixture proportions of each Gaussian, i.e., $\pi_k = 1/N$. Finally, we assumed any test vector of MFCCs could be generated by one or more of the N multivariate Gaussian distributions. Classification is therefore calculated using the posterior probabilities $p(k|\mathbf{x})$ of each of the k components in the Gaussian mixture distribution:

$$p(k|\mathbf{x}) = \frac{p(k)p(\mathbf{x}|k)}{\sum_j^N p(j)p(\mathbf{x}|j)} \tag{3.13}$$

The class giving the highest posterior probability is therefore the chosen class.

3.4.2 PLCA model

As PLCA operates on a matrix in order to describe its latent decomposition, we first transposed each unit of sound into a matrix describing the magnitude of frequencies over time. Each audio signal was multiplied by a Hanning window and broken down into a frequency representation using the discrete Fourier transform with a window size of 371.5 ms (16384 samples at 44100 Hz) and hop size of 92.9 ms (4096 samples at 44100 Hz). Composing the absolute power spectrum into a matrix of frequency versus time denotes the Short Time Fourier Transform (STFT).

Using the formulation described in Section 3.3, we encoded an acoustic class by running PLCA on each training example's STFT, allowing each class to be described by a single component. From these results, we formed a dictionary that was used for classification by aggregating into a matrix each class's latent frequency distribution, $p(f|k_i)$, for $i = 1...N$ where N equals the total number of trained classes. Then, using the trained dictionary $p(f|k)$, the latent distribution over weights $p(k)$ and impulses $p(t|k)$ are maximized using the EM update rules described in Equations 3.2 and **??**. As we were testing whether our possible distributions of frequencies (our dictionary) were capable of describing the audio signal, we did not allow updates of $p(f|k)$.

3.4.3 Mel-PLCA Model

The last model we describe was built in the same way as the PLCA model, except it uses as input a Mel-frequency transformed STFT rather than a linear-frequency scale (i.e. $p(f,t) \rightarrow p(S_m(f),t)$. The Mel filter-bank effectively performs a data-reduction from a 16384 point frequency transform in the standard PLCA model to a 40 element vector by summing the energy in the Mel-frequency critical bands. This model most resembles approaches taken in MPEG-7 Spectral Basis Decomposition, however does not take the last step of de-correlating the frequency scale, and further makes use of PLCA instead of PCA or ICA.

3.5 Evaluation

With our computational frameworks developed, we now turn to evaluating them within an information retrieval context. In particular, we look at how they perform classifying acoustic events presented in isolation, with noise, or as simultaneous mixtures.

3.5.1 Material

Sounds were sourced from both the Sound Ideas[1] archive and the BBC Sound Library and selected based on whether the sound file consistently represented a single sound class, thus being consisted with our specific definition of an acoustic event. We removed any beginning or ending silences or envelopes of sound, and constrained examples that were not at least 10 seconds long. In total, we were left with a single example of $N = 37$ classes: *airplane, arcade, booing, bubbles, bus, cheering, chickens, clapping, clock-ticking, conversation, copier, crickets, dirt-drive, fan, fire, fire-gas, geiger, hair-dryer, jet-engine, laughing-audience, laughing-man, motor, race, rain, refrigerator, shouting, sink, spray-can, steam, swamp, sword, train, treads, trees, typing, waterfall, and wooden-gears.*

3.5.2 Experiments

Experiment 1 We train one classifier for each of our possible acoustic event classes, thus building a set of 37 classifiers for experiment 1.

Experiment 2 We also determined whether the MFCC and PLCA models were able to correctly classify the trained class in the presence of an untrained class (noise). As we have 37 classes, this equates to 36 possible mixtures for each class, where each of the 36 classes are trained in isolation, and tested in a mixture of a 37th untrained class. In order to create the $37 * 36 = 1332$ possible mixtures, we used balanced mixing. For this experiment, this means each class is actually represented with 36 possible examples (36 possible mixtures for each class).

Experiment 3 Finally, we added the 37th un-trained class to the set of possible classifiers in order to see if both classes could be correctly classified when presented as

[1] Courtesy of Ianis Lallemand, http://imtr.ircam.fr/imtr/Environmental_Sound_Dataset

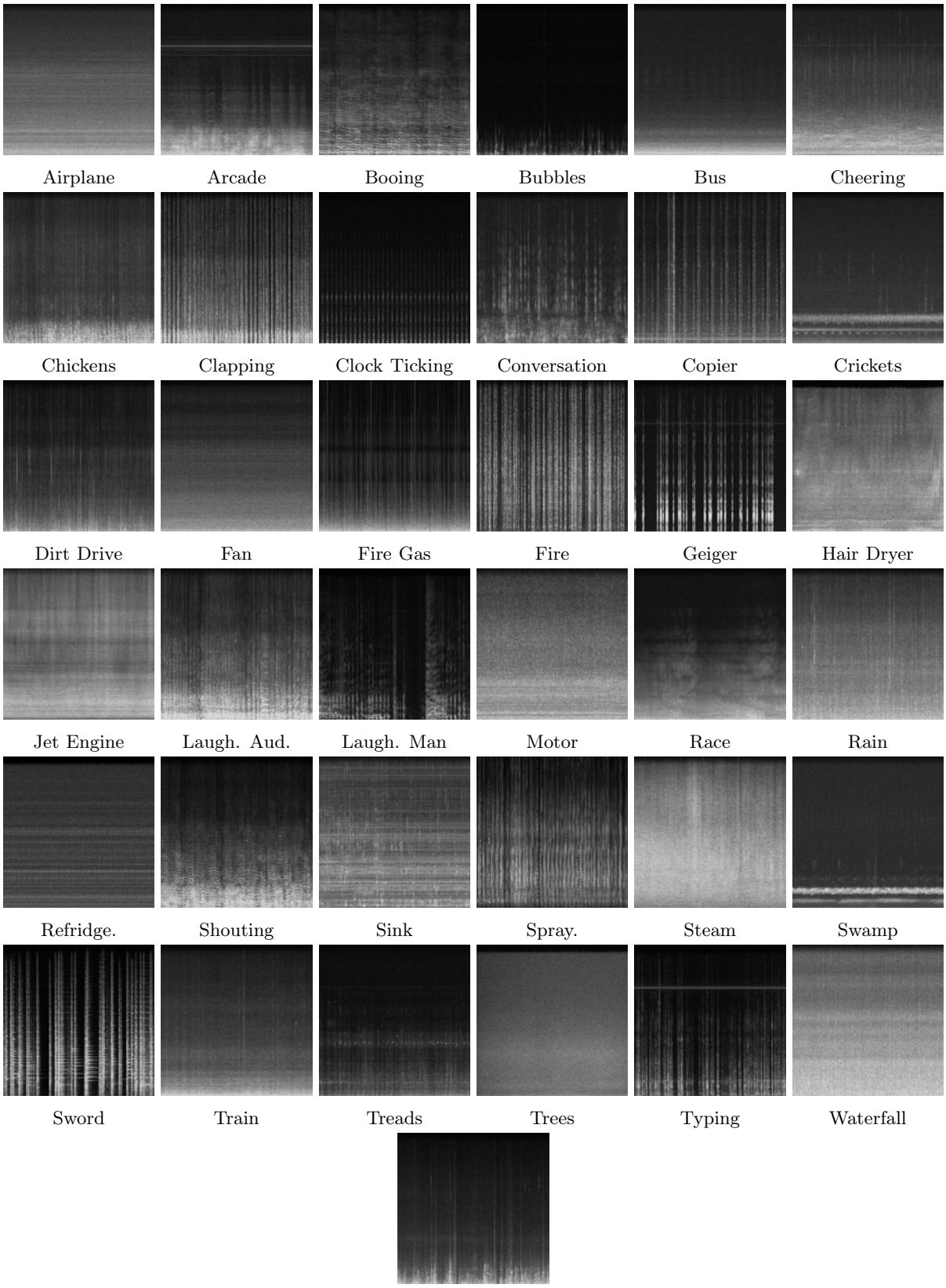


Figure 3.1: Spectrum describing the frequencies (y-axis) over time (x-axis) of our 37 different classes. White corresponds to higher values.

an acoustic mixture. This means we tested on $\binom{37}{2} = 666$ possible mixtures and sought to find out whether the MFCC and PLCA models were capable of classifying either or both of the mixed acoustic classes, even though they were presented as a single acoustic stream.

3.5.3 Validation and Reporting

We performed k-fold cross-validation using 10-folds. With 10 seconds per class (370 seconds total), this equates to 1 second folds per class where training occurs on 9 seconds of material per class, and testing occurs on 1 second of material per example. The results of all folds were then averaged together to produce a single estimation.

In order to assess the estimated results, we made use of a standard technique in describing classification performance, the Receiver Operator Characteristic (ROC) curve. ROC analysis describes ground truth classes as true and false and the predicted measures as positive and negative for a binary classifier. The ROC curve then measures the accuracy of the classifier in separating the actual true class from the non-classes by relating the sensitivity, or the *true positive rate*, against 1−specificity, or the *false positive rate*. In order to build the curve for a continuous classifier, the classifier's response must be converted to a set of binary classifiers by using equally spaced thresholds. We did this by taking equally spaced thresholds on the results of our cross-validation, and calculating the true positive rate of a bin i as:

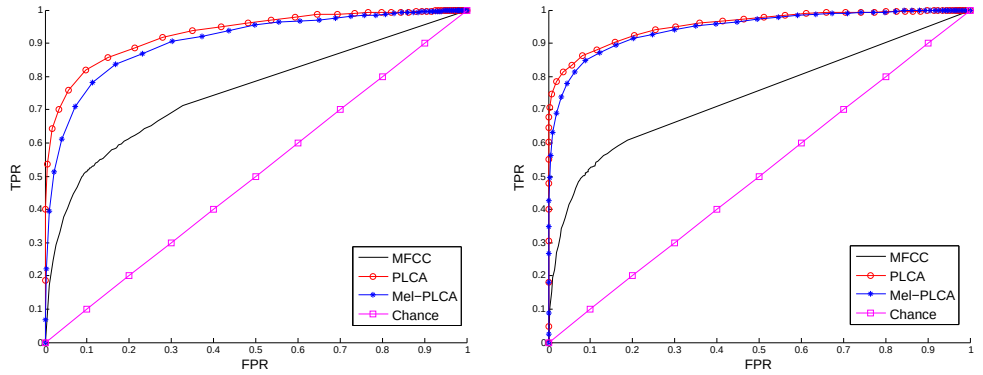
$$\mathrm{TPR}_i = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.14}$$

and the false positive rate as:

$$\mathrm{FPR}_i = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{3.15}$$

The resulting (x, y) points relating the false positive rate to the true positive rate are plotted for each classifier.

A perfect score is denoted by 100% sensitivity (no false negatives) and 100% specificity (no false positives) and corresponds to a point in the top-left corner, (0,1). A



(a) Experiment 2: Classification masked by noise.

(b) Experiment 3: Classification of acoustic mixtures.

Figure 3.2: Results of Experiment 2 and 3 as depicted by the ROC curve.

classifier that performs at chance lies along the diagonal going from the bottom-left to the top-right corner.

As well, the area under the ROC curve (AUC) neatly summarizes the performance of the curve with 1.0 being a perfect score, and 0.5 being a classifier that performs at chance. We can also understand the AUC as the probability of classifying a randomly chosen positive instance with higher likelihood than a negative one.

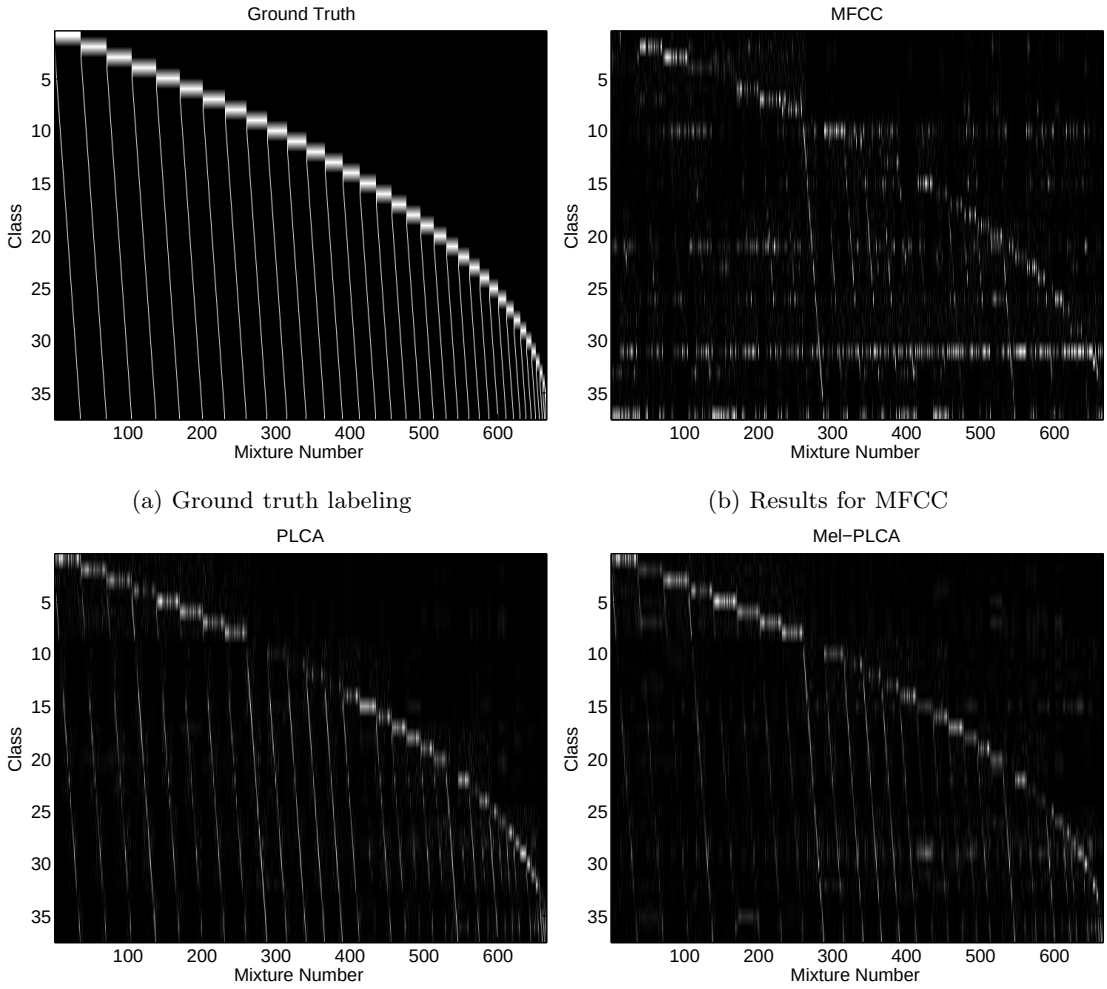
3.6 Results

3.6.1 Experiment 1: Classifying isolated acoustic events

We tested the performance of a single event class in isolation. The performance of the MFCC and PLCA models as determined by the ROC analysis are all excellent, with an AUC of within 0.001 of perfect discrimination. These results suggest that all three computational methods are suitable for representing acoustic scenes when they do not need to be segregated.

3.6.2 Experiment 2: Classifying acoustic events in the presence of noise

We tested the performance of both the MFCC and PLCA-based classifiers in the presence of noise by mixing one of 36 trained classes with an untrained class of sound (the 37th class), effectively masking the trained class with noise. The average results of



(a) Ground truth labeling

(b) Results for MFCC

(c) Results for PLCA

(d) Results for Mel-PLCA

Figure 3.3: Experiment 3: Classification performance of acoustic mixtures depicting the ground truth classes for each of the 666 mixtures and the MFCC model, the PLCA model, and the Mel-PLCA model's classification likelihoods for each of the 666 mixtures. Images represent likelihood of a class in a given mixture, with white being 1.0, and black being 0.0.

1332 mixtures are depicted in Figure 3.2a using ROC curves depicting each model's performance in classifying the correctly masked class. We can see the MFCC model does well above chance, though both of the PLCA models do a much better job. Interestingly, the Mel-PLCA model is very close to the performance of the full-spectrum based PLCA model, even though this model uses only 40 samples versus 8192 samples per frequency frame.

Table 3.1: Area Under the Curve of ROC Analysis

Method	Experiment 1	Experiment 2	Experiment 3
MFCC	**1.0**	0.7388	0.7410
PLCA	**1.0**	**0.9303**	**0.9548**
Mel-PLCA	0.9989	0.9065	0.9443

3.6.3 Experiment 3: Classifying mixtures of acoustic events

The last experiment measures the ability of our 3 models to classify both classes in an acoustic mixture of 2. The results depicted in Figure 3.3 show the ground truth for the 37 possible classes across all 666 mixtures ($\binom{37}{2=666}$ classes) as an image. As well, this figure shows the likelihoods assigned to each of the 37 classes across all 666 mixtures for each of the 3 models. From these figures, we can see the performance of the MFCC model struggles to classify most of the mixtures accurately, and often produces a false positive for classes 10 (conversation), 21 (laughing-man), and 31 (sword). Thresholding columns of this image and storing the TPR and FPR as described in Section 3.5.3 produces 666 ROC curves. The average of these curves are depicted in Figure 3.2b, showing the performance of the MFCC model to be well above chance, though with the PLCA and Mel-PLCA models doing far better. Interestingly again, we find the Mel-PLCA model is able to perform nearly as well as the PLCA model, performing within 0.01 of the PLCA model's AUC.

3.7 Discussion

We developed 3 computational models for the encoding and decoding of auditory scenes and tested their performance in the context of acoustic classification. All 3 models performed with near perfect results when a single acoustic class appeared in isolation. However, when the event was masked by an unknown acoustic class, effectively adding noise, the performance of the MFCC model dropped, though still performed well above chance. The remaining models however performed much better. Our last experiment tested the performance of each model to classify multiple parts of an acoustic scene by mixing 2 classes together, effectively requiring segregation. The MFCC model again performed well above chance with an AUC of 0.74, but the models built with PLCA again performed with much stronger results, exhibiting > 0.9 AUC.

One possible reason for the poor performance of MFCCs during classification of mixtures is the signal model assumes a single excitation source (e.g. vocal tract or instrument). In the presence of multiple sources, ambiguity is created, making it difficult to estimate which source contributes to each of the coefficients, especially since the sources are also combined non-linearly through the step of a log-transformation.

Two disadvantages of using a full and direct spectrum model such as our "PLCA model" noted by (Casey 2001) is their inconsistency and dimensionality. We therefore tested a second model similar to the MPEG-7 spectral basis decomposition described in (Casey 2001), "Mel-PLCA", which reduced the 16384 point Fourier spectrum to a 40 element vector. However, unlike the MPEG-7 spectral basis decomposition, we made 2 significant changes: (1) we made use of the Mel-frequency scale rather than log-decibel scaling and normalization; and (2), as the article in question was written nearly 12 years ago, the only basis methods described were SVD/ICA/ and PCA based methods as PLCA had not yet been published. Incorporating these changes, we found that the Mel-PLCA model performed within 0.03 of the full-spectrum PLCA model. Using the critical bands defined by the Mel-frequency scale ensures the inconsistencies that may be apparent within similar acoustic classes are averaged out, and perceptually relevant frequency dimensions describing the class are retained while keeping dimensionality very low.

3.8 Future Work

A number of viable extensions are possible. First, as we only made use of highly textured atmospheric sounds, it remains to be seen whether the following method alone would suffice in modeling more impulsive sounds, e.g. drums, birds, or less atmospheric sounds. In such cases, an entropic prior on the temporal weights of a PLCA decomposition would very likely greatly improve results (Smaragdis 2007a), ensuring the sparsity of temporal weights in the latent distribution $p(t|k)$, while capturing the bulk of the frequency distribution in the latent factor $p(f|k)$.

Second, 2D patch-based and shift-invariant convolutive pLCA (Smaragdis 2007a) has shown great promise in capturing the structure of music when applied to chromagram features and when using sparsity and shift-invariance in all features (Weiss 2011). Such a

technique has the power not just for classifying the instruments that describe a musical passage, but as well the course of events that describe the musical scene, essentially identifying whole musical passages or riffs.

Third, In real-time scenarios, it is often the case that a dictionary of classes is not readily available. Recent work describing the online-learning of dictionary elements using PLCA has shown great promise in performing real-time speech de-noising (Duan 2012), resulting in components separating noise and speech. Such a distinction has wide applications in fields such as surveillance and tele-presence technologies.

Lastly, in developing this work, it became apparent that no standard publicly and freely available libraries for evaluating acoustic scene classification algorithms exists. Though the problem is well noted in music information retrieval (Casey 2008a; Rhodes 2010), and recently addressed with databases such as the million song dataset (Bertin-Mahieux 2011), no standardized databases have been developed as freely available archives in the general sound-based multimedia communities. As such, testing the scalability of our approach proved very difficult, as we could only obtain 37 classes and a total of 1332 mixtures even though databases such as YouTube and typical multimedia archives are on the order of many millions. Future work must therefore be done to help understand the scalability and performance across different approaches using a standardized database.

3.9 Conclusion

The currently developed models have shown promise at encoding and decoding simple auditory scenes composed of a single auditory event, and even complex auditory scenes composed of multiple events. However, the evaluations presented here do not consider the integration of a very large corpus. In the next chapter, we explore scene synthesis of much larger audio corpora in two practical outputs, Memory Mosaic and The Daphne Oram Browser. Consequently, we also explore methods for integrating a large number of representations, effectively allowing only representations that are more deviant than others to be learned. We also extend the model to incorporate Event Detection within the real-time system of Memory Mosaic.

Chapter 4

Computational Auditory Scene Synthesis

Contents

4.1 Introduction to Auditory Scene Synthesis

The juxtaposition of fragments of sound as an arts practice has roots at least as early as music concrete, 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 corpus-based concatenative synthesis, where a target sound is matched to a stored database of segments or sounds (for a comprehensive review, see (Schwarz 2006)). In this thesis, we focus on a collage-based technique called scene synthesis which assembles units created through the computational modeling of psychologically motivated representations thought to support perception. The aim in building these units and re-assembling them within a collage-based practice is to open a dialogue into the nature of representation within perception.

We have so far derived a conceptual framework for building auditory representations based on evidence in electrophysiology in Chapter 2 and started to develop this framework computationally in Chapter 3. In particular, we demonstrated that our computational model described by Mel-Frequency Cepstral Coefficients (MFCCs) perform very well at encoding auditory scenes in cases where the entire scene must be integrated. Further we demonstrated that when segregation or understanding the multiple sources in a scene is required, models based on Principal Latent Component Analysis (PLCA) perform very well and are much better suited than MFCCs.

We now turn to two practical developments using these computational models as it is within the necessary practical developments that we can better understand the applicability of these models to real-world and interactive settings, and also better understand where there are any missing pieces to our computational model and conceptual framework. The first practical output, the Daphne Oram Browser, compares our MFCC and PLCA models within an interactive scene synthesis. This scene synthesis is presented to a user of an audio archive in the form of a 3D virtual browser where they can select any segments of sound and hear them played back in real-time as an interactive sonic-collage. This output attempts to discover how the representations motivated in Chapter 2 and computationally modeled with PLCA may help researchers understand any inherent content-based relationships within the content of a large audio corpus. In other words, by using the representations built using our computational models, the relationships are not based on the content's filenames or metadata, but based on the way the content sounds. A successful representation will ideally be able to define any inherent relationships within the archive based on purely sonic aspects.

Finally, we describe Memory Mosaic a real-time auditory scene synthesis in the form

of an iOS application. This output will require the development of the last missing component of our conceptual framework, event detection, which explicitly defines events within a real-time stream of audio. As motivated in our conceptual framework in Chapter 2, event detection describes our ability to unconsciously detect change within an auditory scene. The auditory units demarcated by this process are then encoded to produce, over time, a large database of sonic "memories". Within this real-time experience, the current unit of sound is also decoded, effectively matching the segment to the closest previously encoded stream to produce a real-time scene synthesis.

4.2 The Daphne Oram Browser

The Daphne Oram Collection presents a unique case-study for the study of auditory representations. It contains over 120 hours of 1/4" tape dating from 1957 onwards and includes studio compositions, radio plays, sound effects, lectures, and interviews related to the British electronic musician, Daphne Oram (31 December 1925 - 5 January 2003) (Young 2008). Researchers working to understand the archive want to know more about her, her compositional techniques, and her radically innovative instrumental technique of "Oramics" for creating electronic sounds. They have thus recently begun the process of digitizing the archive and at the time of this study had digitized 60 hours of the collection. This process requires converting the analog tape reels to digital lossless formats allowing them to markup the digital files with metadata, or any additional textual descriptions of the audio file. However, as this process has gone on, researchers have realized that the archive contains many duplicates and very little to no metadata. In an effort to aid the researchers understand the digital contents of their archive better, we employed our representations to build a 3D visualization of the archive where the axes of the visualization refer to relationships between similarly represented material.

Using two of the methods developed in Chapter 3, we focus on visualizing the Daphne Oram archive using the encoded representations in order to project the archive as a 3-dimensional visualization. These two methods are: (1) discovering latent distributions of frequencies using a recently developed source separation algorithm, probabilistic latent component analysis (PLCA) (Smaragdis 2006); and (2), using a widely-adopted multi-dimensional feature for speech, music, and general acoustic classification, the Mel-Frequency Cepstral Coefficients (MFCCs). We cluster the data from either descriptor

using Multidimensional Scaling and develop a 3D visualization that allows researchers to project the archive onto multiple dimensions of the data. Finally, we report user-feedback from researchers of the archive using the 3D visualization tool.

Our main contribution is in describing the impact of visualizing segregated streams of a large audio archive through a case-driven exploration of the work of Daphne Oram. We compare the feedback from archivists using a visualization of latent timbre-relationships versus one using a perceptually inspired multi-dimensional feature, MFCCs, and find that PLCA is more effective at producing a meaningful visualization.

4.2.1 Related Work in Visualizing Audio Archives

A number of previous approaches for visualizing large audio corpora have focused on the application of music-based corpora. Some approaches to content-based musical information retrieval solutions require a user to search by example or performance, aiding retrieval when a user is unaware of exactly what they are looking for. However, visualization of such retrieval methods often amounts to viewing lists of the k-most similar results of an explicit query, and thus any exploratory analysis of the corpus as a whole requires further research into approaches for visualization. For instance, SMILE (Melucci 2000) presents a MIDI-keyboard for the user to "perform" a query, and results are presented based on how similar the MIDI sequences are to the performance. Similar approaches built for more generic signal-based audio break a corpus into frequency information and further into fingerprints such as MFCCs or psychacoustic descriptors. audioDB (Casey 2008b; Rhodes 2010) for instance allows a user to input shingles, or segments of an audio track for discovering similarities in an archive. Other solutions such as Query-by-humming allow a user to hum/sing a tune in order to discover similar results (e.g. (Wang 2006; Cartwright 2011)). The previous methods may be suitable for applications where a user has an explicit example query. However, in exploring an archive, it requires the user to have a priori knowledge of what the archive already contains.

Early work in exploratory content-based visualization systems can easily be traced to the 1990's where Starfields were commonly employed. Starfields are interactive scatter-plots that allow for zooming, panning, and selection for greater detail, allowing one to view an archive through interaction. The Informedia Digital Video Library Sys-

tem (1994-1998) (Himmel 1998; Christel 1998) is one such system making use of the Starfield visualization approach, which accesses over a tera-byte of video and presents the user with an interactive scatterplot organized by the user's query. Beginning with the audio signal, Informedia-I uses the Sphinx-II speech recognition system to discover annotations of audiovisual material. Adding these to any existing text-annotations from captions, they create a term-document frequency matrix for each video segment, where segments are determined through the use of motion-based video-cut detection. They are then able to discover latent relationships using PCA for reduction and visualization. Other approaches such as IVEE (Ahlberg 1995) allowed for visualization options such as Tree Maps, Cone Trees, and even 3D scatterplots, though were not rooted in content-based information retrieval and instead relied on explicit relations of existing meta-data. Though these early works were not directed for musically-based archives, their approaches towards visualization and interaction are very similar to ours, as we also look for latent relationships for reduction and visualization.

More recently, CataRT (Schwarz 2008) approaches Starfield style visualizations of large audio corpora by computing low-level psychoacoustic descriptors of grains segmented from a corpus for the purposes of composition, orchestration, and texture synthesis. Visualizing the resulting mappings occurs in a 2D space where each axis is defined by a descriptor chosen by the user. Such a visualization has the benefit of user awareness and control over the mappings that define a parametric spatial mapping. Plumage (Schwarz 2008) extends the CataRT visualization into a 3D space creating a performance and composition environment where grains are colored, textured, and morphed in 3D space based on their psychoacoustic descriptions. nepTUNE (Knees 2006) and (Dominik 2009) are two approaches to visualization which create a 3D terrain-style virtual space. Songs are clustered using a self-organizing map of acoustic similarity in order to create virtual islands and terrain based on their clustering density. The created virtual space thus encourages exploration and navigation of the visualized corpus. (Bartsch 2001)'s approach employs the use of chroma-features for producing audio thumbnails of tracks, or segmented versions of an audio track encoding heavily repeated structures of harmonic relationships. Though their approach is well-suited for popular music archives, they note that it is not suitable for music that does not obey a simple "verse-refrain" form. (Stewart 2008) uses mood words to describe a 3D interactive visualization, though relies on having access to socially tagged music in order to represent

the music archive. (Heise 2012) use MFCCs to describe an unknown corpus of audio and explore the audio using a 2D visualization created with a self-organizing map.

The critical deviation of our approach to feature analysis from the previous approaches is by defining a 3D space using the corpora's own latent frequency distributions. As we make no assumptions to the structure, perceptual relevance, or harmonic nature of the corpus, using probabilistic latent component analysis, we can discover the archive's own predominant distributions of frequencies and are able to use this reduced dimensionality dictionary as a representation of a high-dimensional space. When projecting any 3-dimensions, the user is able to navigate the archive in a manner similar to CataRT (Schwarz 2008). However, the axes are not user-defined psychoacoustic descriptions, but rather are projections of the archive onto "timbres" defined by latent frequency distributions (i.e. the encoded representations). Our work similarly encourages exploration and navigation as in (Knees 2006; Dominik 2009; Heise 2012), though takes a information-centric point of view to analysis and retrieval. We build a second visualization using a model which does not take into account the density of the data but instead uses a perceptual frequency transformation for building decorrelated features, MFCCs, similar to (Heise 2012) and report the user feedback for each visualization.

4.2.2 Methods for Visualization

Currently, the Daphne Oram Archive has over 215 tape reels or 60 hours digitized. As the amount of available memory is a constraint on our approach, we are only able to investigate the first 10 minutes of the first 60 tape reels or 10 hours in total. We describe each half second segment by their frequencies over time, described using the short-time Fourier transform, and describe each time-frequency matrix as a slice. In total we have 1200 slices per tape and 72,000 slices for all 60 tape reels[1]. We aim to visualize this data using a clustering algorithm able to extract the timbre-relationships within the archive. Specifically, we look at two methods for grouping the possible interesting frequencies describing the archive: (1) PLCA, a probabilistic method for discovering latent component relationships of a time-frequency matrix, and (2) MFCC, a widely-adopted approximation of the frequency spectrum inspired by the human auditory system's response properties.

[1] We use all data for building the description of the corpus, though later use a reduced subset for visualization.

4.2.2.1 PLCA Model

We use the basic model for PLCA as described in Section 3.3 for encoding the entire archive. However, one important extension is required, as with the basic PLCA model, the number of components describing a distribution must be known *a priori*. We therefore incorporate model selection, a commonly employed information theoretic approach to determining parameters of a model. In the case of PLCA, the model parameters are described by N, the number of components or encoded representations to use. We could use as many components as we have slices, producing 72,000 components. However, this would likely over fit the database. To appropriately determine the correct value for N, we use *Bayesian Information Criterion* (BIC) model selection. Using the log-likelihood of the optimized parameters, an additional parameter which penalizes model complexity is subtracted from the log-likelihood:

$$\ln p(X) \simeq \ln p(\mathcal{D}|\theta_{\texttt{MAP}}) - \frac{1}{2} M \ln N \tag{4.1}$$

where M is the number of parameters in θ and N is the number of data points. BIC ensures that we do not let the model overfit to a large value of N, while still producing a suitable log likelihood explanation of the observed data.

To begin the model selection, we iterate through every slice of audio. Using model selection, we compare the results of using the current number of components and using an additional component. If the results are better explained with an additional component, we add one to the value of N and continue to the next slice. Iteratively running PLCA across all slices on increasing values of N until finding the maximum BIC results in finding $N = 45$ for 10 hours of audio.

4.2.2.2 MFCC Model

For our second model, we use Mel-frequency Cepstral Coefficients (MFCCs), as described in Section 3.4.1 which approximates a frequency spectrum by a set of de-correlated features.

4.2.2.3 Multi-dimensional Scaling

After running each model, we are left with an $M \times N$ dimensional matrix, where M refers to the number of time slices, and N to the number of dimensions that describe each feature. In the case of PLCA, after running model selection, we are left with $N = 45$ dimensions describing the data. With regards to MFCCs, we specifically choose $N = 13$ cepstral coefficients.

In order to visualize the high-dimensional space created by either model and cluster together similarly weighted features, we make use of Multi-Dimensional Scaling (MDS), a popular technique for multivariate and exploratory data analysis. MDS is a common technique for projecting data in high-dimensional spaces to 2 or 3 dimensional spaces for the purposes of visualization. It aims to preserve the pairwise distances between data points, starting with the notion of distance, and working backwards in order to create the coordinate space. The basic algorithm for calculating the unknown low-dimensional coordinate map $\mathbf{X}$ thus starts with a distance or proximity matrix, $\mathbf{P}$. We aimed to use the full archive of 72,000 slices, however creating a matrix of float values this large requires 20 gigabytes of information which must be held in RAM. Therefore, we reduce our database by taking every 5th slice, effectively looking at 0.5 second slices every 2.5 seconds rather than every 0.5 seconds. However, the description of the data in the case of PLCA is still dependent on all 72,000 slices.

In order to calculate the low-dimensional coordinate matrix, we calculate the largest eigenvalues of the distance matrix after applying a double centering procedure. The basic MDS algorithm is summarized as follows:

1. Compute a $M \times M$ proximity matrix $\mathbf{P}$ by calculating the Euclidean distances between each of the M features
2. Compute the inner product matrix $\mathbf{B}$ by applying double-centring to the proximity matrix $\mathbf{P}$:

$$\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{P}^{(2)}\mathbf{J} \tag{4.2}$$

where $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\mathbf{T}}$ and n is the number of objects.

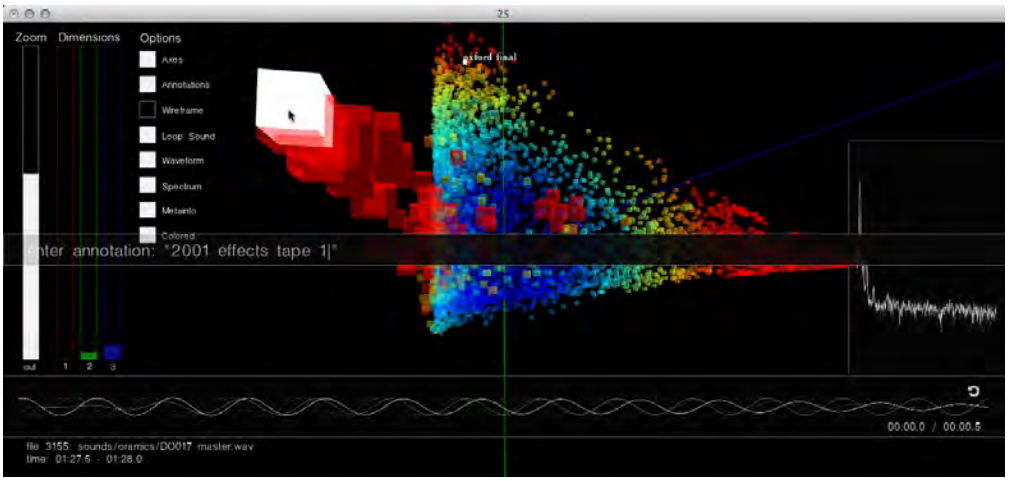


Figure 4.1: Screenshot depicting the GUI of the browser (best viewed in color). Here a user is currently inputing text in order to annotate one of the sound segments. We can see sliders to the left allowing the user to zoom in/out, change the dimensions of the visualization, and control which elements are drawn on screen. With all of the options being drawn, we see the waveform of the currently highlighted sound (depicted with a white cube under the mouse cursor) is drawn on the bottom. As well, the meta-data describing the file name is just below the waveform. To the right, the decibel-scale spectrum is also drawn. All elements are drawn in real-time and are interactively manipulated in 3D space.

3. Compute the eigenvalue decomposition and retain the n largest eigenvectors, $\mathbf{e}_1, ..., \mathbf{e}_n$ in order to compute the n-dimensional coordinate matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{E}_n \mathbf{\Lambda}_n^{\frac{1}{2}} \tag{4.3}$$

using the eigenvectors $\mathbf{E}$ and eigenvalues $\mathbf{\Lambda}$ of $\mathbf{B}$

One may also notice the algorithm is equivalent to a doubly-centered version of PCA in the case where the distances are Euclidean. As both the PLCA and MFCC model's feature dimensions are de-correlated, we would expect to find the number of eigenvalues approach the same dimensionality as either model. Thus, the PLCA model is clustered in 45 dimensions, and the MFCC model in 13.

4.2.3 Graphical User Interface of the Browser

The interface is shown in Figure 4.1 and is built in C/C++ using the creative-coding toolkit openFrameworks[2]. The user is presented with a 3D space (see Figure 4.1) where

[2]http://www.openframeworks.cc

each slice of sound from the archive is represented as a cube projected in 3D space. The coordinates of the cube are determined by which dimensions of the MDS coordinate matrix are selected. To begin, the first three dimensions are displayed. Users can then select any dimension to be displayed on the 3-axes. As a result, the visualization can also be constrained to a 2D visualization by simply choosing the same dimension for 2 axes. A colormap is used to help depict distance from the OpenGL origin (using a "jet" colormap, i.e.: blue-yellow-red), though the user can turn this off. Figures **??** and **??** depict the visualizations of the first three dimensions produced using MDS inside the browser.

While inside the browser, pressing space-bar allows one to annotate the currently selected slice. The annotated text appears in 3D next to the slice's cube. The slice's audio is also visualized as a waveform and its instantaneous Fourier transform. As we used the first ten minutes of every tape-reel, the waveform for any given slice is presented as a looped region within a 10 minute audio file. However, the user can change the loop regions to hear any other portion of the original audio file while selecting a slice, thus allowing the user to listen to the audio before and after the slice.

The user can also move the camera around the OpenGL origin by dragging the left mouse button in the 3D space. Highlighting any of the cubes with the mouse allows the user to inspect the clip in greater detail. Taking a cue from the 2D analog CataRT, any of the cubes can be "scrubbed" for playback by simply moving the mouse over any of the cubes, not requiring any further interaction to listen to the sound sample. Zooming in and out of the 3D space can be done via the mouse scroll wheel or graphical slider. Double-clicking on any of the cubes re-centers the origin to the selected cube, allowing camera interaction to occur with respect to the cube. Cubes can be spaced closer or farther from each other using another graphical slider. This allows more tightly clustered portions of a visualization to be explored in greater detail.

4.2.4 User Feedback

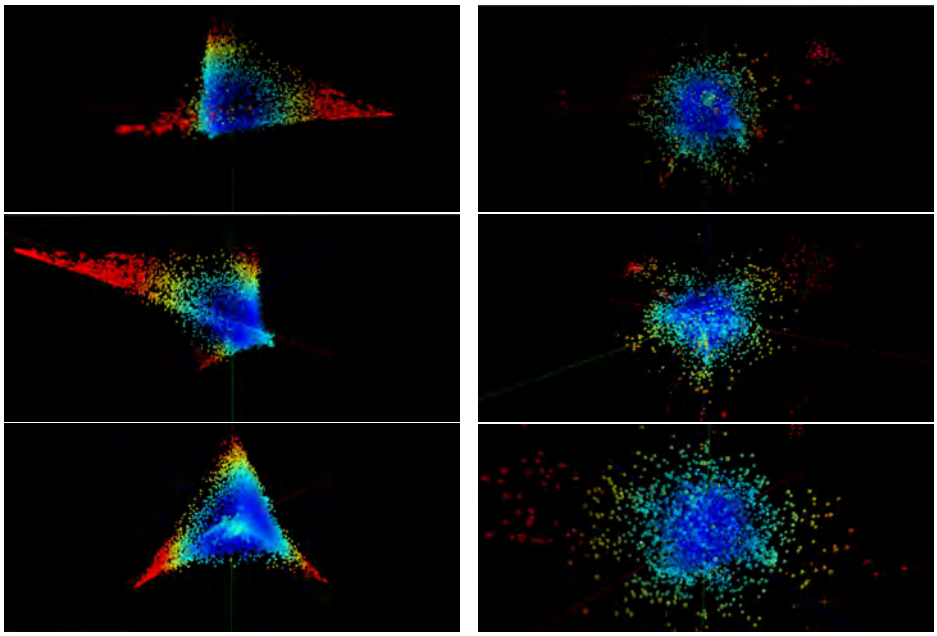
Three researchers of the Oram Archive were invited to navigate the browser and spent 1 hour in total using both the PLCA and MFCC visualizations. They were unaware of how either model was created, were unfamiliar with signal processing and machine learning, and were only told that we are investigating a way to navigate the Oram Archive. Each

user was given 5-10 minutes of explanation of the features of the browser and were then left to explore the browser by themselves. Each user proceeded to explore the archive by using the mouse to listen to the different slices located in 3D space. In addition, each user managed to find particularities of the archive that seemingly would have been very difficult without the browser. For instance, finding a significant portion of one tape reel that was labeled as "POP TRY-OUTS" in another reel labeled as "COPY DONKEY HELL ABC & ITV. BIRDS & PERC" by exploring slices located near each other in the 3D space. Also, one found components relating to Daphne Oram's piece, "Birds of Parallax" during lectures series that were only labeled by their location, indicating she demonstrated these components during her talk.

When asked to compare the two visualizations and remark on their usability as a navigation tool of the Daphne Oram Archive, the three researchers reported on the form of the MFCC model in comparison to the PLCA one, saying (1), "it has a less useful shape in general", (2) "it has less detail", and (3) "this dense mass represents total variety...and I don't quite understand how it is mapped." In response, we asked what if anything made the PLCA model more useful for navigation in comparison to the MFCC model. User 1 reported: "it has a more definite and understandable space. For example, prongs that have specific information in them such as silence." and User 3 reported: "Oh that's really successful, it seems to be matching pitch and you start to see how she was using pitch" and "I had a clear sense of how it was mapped".

Each user also gave many helpful possible extensions to the current functionality of the browser, including the ability to save camera states, only view a particular reel's slices, and auto-zoom and rotation around a particular point. User 1 found the 3D nature of the visualization required more practice saying they "might get used to it" while User 3 commented on navigating around a single slice saying "I understand it as a structure, but I'm working out where in 3D space [the slice] is. You have to move around in 3D before working it out." User 3 also expressed the scope of the browser for new users to see and appreciate Daphne Oram's work, remarking, "Goes to show just how much variety there are in the samples, and this has made that variety accessible."

User 2 additionally remarked on the potential of incorporating other mediums of Daphne's work saying it would be great to "include other mediums than audio, combining with video/letters/images." As well, both User 2 and User 3 commented on the



(a) PLCA model (b) MFCC model

Figure 4.2: Screenshot of the first 3 dimensions of the PLCA and MFCC models visualized in the browser. We show three different views here.

tool's applicability to performance and composition, saying he/she was "fascinated as a compositional tool. Navigating different dimensions, it's a beautiful instrument" and "it is nice to categorize sounds as it is what we do in sampling".

4.2.5 Discussion and Future Work

Each researcher had prior knowledge of many aspects of the Archive and Daphne Oram's composition techniques, and were also familiar with many of the recordings. Their interests in the archive stemmed from her methods in composition to the actual electronics of the Oramics Machine. When given the chance to navigate the archive in a 3D space arranged by acoustic similarity, each user was incredibly pleased by the possibilities and results of just one hour's navigating, and also preferred the PLCA model to the MFCC one generally for 3 reasons: (1) the visual form and structure of the PLCA model was easier to navigate, as knowing where one was in 3D space is easier to notice, (2), navigating within the "glob-like" mass of the MFCC representation in 3D required users to go inside the sphere, making interaction very difficult, and (3), the mapping and clustering in the PLCA model appeared more intuitive, with users reporting they understood how it was mapped and the similarity of sounds along a projection seemed to cluster sounds

better.

Regarding (1) and (2), the form of the PLCA model (see Figure 4.2a) is a result of the probabilistic nature of the component weights needing to sum to 1. In 3D space, this space is defined by a 3-simplex or tetrahedron. In comparison, MFCCs may have energy explained in all bands as there is no normalization procedure. Plotting the first three dimensions of the MFCC model thus produces similar distributions of energy in all dimensions, creating what users called both a "glob like" and "blob like" sphere (see Figure 4.2b). Navigating inside and around a sphere presents unique challenges for a 3D browser, namely, it is difficult to select elements within the sphere and understanding the orientation of the sphere is difficult as there are no identifying features. Thus, exploring visualizations in 3D seems to require landmarks for useful navigation. In regards to point (3), this may be due to the greater classification and recognition performance of PLCA over MFCCs as shown in Chapter 3.

Further work should focus on issues with navigating in 3D space, as some users reported on the 3D nature as requiring practice to navigate. One solution may be to create more intuitive control through the use of other input and display devices such as touch-screens. As well, similar latent-analysis techniques may be applied for additional meta-data from the archive as is done in audiovisual and text corpora, e.g. (Himmel 1998; Christel 1998), to create more informed visualizations. In this case, the input of text annotations as well can create a user-guided visualization, where feedback from the user reshapes the 3D visualization.

4.3 Memory Mosaic

We now turn to a second practical development in auditory scene synthesis: Memory Mosaic. This app attempts to develop a real-time scene synthesis experience of sonic "memories". This app will later be combined with a real-time visual scene synthesis in Chapter 8. Following our conceptual framework described in Chapter 2, we must develop event detection in order to understand when an event of sound has occurred within a real-time stream of audio. Further, another motivation for developing auditory scene synthesis is to be able to take the application and explore different auditory scenes (and not just an office scene), meaning we restricted our development to using a mobile

platform.

Following our conceptual framework, after event detection, we must encode the event using our integral (MFCC) or segregated (PLCA) computational representation, and decode it by matching to any learned representations. In our initial tests, we developed Memory Mosaic with segregation in mind, given its greater performance for encoding as demonstrated in Chapter 3. However, practically, running such a model in real-time and on a mobile platform is not yet possible. We therefore restrict our discussion within this sub-chapter to using the computational model described by MFCCs, though hope this work could be expanded in the future to include segregation as well.

4.3.1 Related Work

Our work very much resembles SoundSpotter (Casey 2007), though works with a real-time audio stream rather than a pre-recorded one. It is also very similar to approaches for real-time event segmentation and classification (Collins 2004**?**) which have also been implemented for a mobile platform, though our approach is not meant for solely musical or drum track events. Instead, our interests are in allowing scene synthesis to occur in any auditory environment, even allowing one to encode representations learned from one auditory space, and decode an entirely different auditory space using the other space's representations. However, it was also not our intention to make these or any other instructions explicit to a user, and we instead allow users to raise their own questions in the process and their understanding.

4.3.2 Methods

The auditory scene analysis model employed in Memory Mosaic is motivated by evidence in literature of auditory perception stressing the importance of temporal regularities of an acoustic scene in providing continuity for maintaining a cognitive model of an acoustic scene. Such research reinforces Bregman's theory of streaming (Bregman 1990), where one phase consists of the formation of primitive based features, and another on the schema-based selection of streams. Our model thus places emphasis on temporal discontinuities of the auditory stream (see Chapter 2 for a more in-depth discussion) using a computational description of the acoustic scene based on the well-known cepstral

coefficients (described in Section 3.4.1) .

As our iPhone-based implementation requires a real-time listening experience, our feature transformation must also be fast enough to be computed on every audio frame and each frame must also be tested for segmented, and possibly decoded. Our approach begins with the Fast Fourier Transform (FFT) of an audio signal using a fixed frame-size at a sample rate of 44100 Hz. A 4096-sample FFT provides high spectral resolution while still being fast enough to perform in real-time. Following the real-FFT, the magnitudes undergo a Constant-Q Transform (CQT) with 40 triangular windows, a real log base-10 operation, and a Discrete Cosine Transform (DCT) in order to produce an 89-element MFCC feature vector. We keep the first 13 of these. For our purposes in defining a real-time scene synthesis, we also store the MFCC delta values, or their change from the previous frame, as well as this delta's delta. The resulting 36-element vector can be thought of in terms of the magnitude, velocity, and acceleration of the global shape of the auditory scene. All math operations, including DCT, CQT, FFT, addition, subtraction, and multiplication are performed using the Apple Accelerate framework in order to achieve real-time performance on an iPhone. The full implementation of the presented framework incorporating these libraries are made freely available by the author here: http://github.com/pkmital.

4.3.2.1 Event Detection Model

Our event detection model is based on temporal irregularities. Specifically, the regularities are assumed to be Gaussian in the representational space of the acoustic scene descriptor. This means that any deviances from the generated Gaussian model will be detected as segments. We test deviances to the current acoustic description, though first apply a commonly employed technique of a low-pass operation in an attempt to remove spontaneous noise which may produce false positives. Following the feature transformation, the model has two states of operation: segmenting or not. When the model is not segmenting, the model of background is built up until a temporal irregularity appears. Similarly, for the foreground model, a separate Gaussian model of the foreground is built up from the start of the segment. Aside from detecting the discontinuity within the foreground model, the model also detects if the current frame returns to the background model by computing distance to the background model. Using a prior

assumption of a normally distributed feature-space, we compute deviations of the background and foreground model past 3 standard deviations. As well, checking whether the currently observed audio frame returns to background, we check if its features are within 3 standard deviations of the background model and stop segmenting.

Each new detected segment, or sonic "memory" in the context of the app, is written to disk using Apple's Extended Audio File Format. Only the audio segment's 36-dimensional descriptor is kept in memory in order to form a matrix of vectors where the index of this vector also provides a lookup to the disk recording.

4.3.2.2 Matching

We decode an input sound's encoding using a polyphonic reconstruction from the nearest similar represented sound segments. At each new onset determined by a temporal irregularity of the acoustic scene, a new set of matches are found to the current encoded acoustic scene. The onset detection for the synthesis engine is much like the one for the event detection model; however, no foreground model is kept, and only the background model may deviate, creating new background models at each onset. Thus, the model does not require knowing what is foreground or background, and only requires deviations in the continuous acoustic space.

Matching can be formulated as a nearest neighbor algorithm which begins by creating a metric space X of known points $P = p_1, p_2, ..., p_n$ for n points. These points are pre-processed in such a way that a neighbors to any query point, $q \in X$, are found quickly. To pre-process the points, we use a distance metric to keep a user-selected number of highest matches to any query point using a simple linear index. Iterating linearly over the dataset of the acoustic description vectors, the best matched vectors' indices are kept using cosine similarity, which measures the angle between two vectors $\mathbf{A}$ and $\mathbf{B}$ like so:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \times \mathbf{B}_i}{\times \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{4.4}$$

Using Apple's Accelerate framework, this metric can be computed using efficient

vector operations that are optimized for the iPhone:

Figure 4.3: Screenshot of the iOS application Memory Mosaic

Program 2 Vectorized code for performing cosine distance

```
float cosineDistance(float *x, float *y, unsigned int length) {
   float dotProd, magX, magY;
   float *tmp = (float*)malloc(length * sizeof(float));

   vDSP_dotpr(x, 1, y, 1, &dotProd, length);

   vDSP_vsq(x, 1, tmp, 1, length);
   vDSP_sve(tmp, 1, &magX, length);
   magX = sqrt(magX);

   vDSP_vsq(y, 1, tmp, 1, length);
   vDSP_sve(tmp, 1, &magY, length);
   magY = sqrt(magY);

   delete tmp;

   return 1.0 - (dotProd / (magX * magY));
}
```

4.3.3 Application

The interface of the iOS application is shown in Figure 4.3. The title appears at the top, with a subtitle prefaced by "size" and a number. This number reveals to the user how many "memories" have been learned. Two sliders allow for interactive control of mixing between the fully decoded scene, and the original input target scene. The fully decoded scene is built as a result of the matching procedure outlined above. The input target scene can be either the microphone or a user selected song from the iTunes Library (by

pressing the middle button). The user can also start/stop learning, defining whether any segments from the ongoing target are encoded. As well, the user can also erase the memory, thus reseting the size of the memory database to 0. Lastly, the second slider allows the user to control the number of retrieved matches to any detected events. The higher the number, the greater the polyphony of the resulting scene synthesis.

4.3.4 Results

4.3.4.1 User Reviews

Since uploading Memory Mosaic to the Apple App Store (henceforth "the app") in August 2012, A total of 235 downloads in 2011, 273 in 2012, and 140 in 2013 for a total of 648 downloads have been recorded on the iTunes App Store. During this time, 3 reviews have been sent to the app store where reviewers had to indicate the number of stars to give the app, and 1 review has been sent to me personally. These are reproduced below:

> This a fascinating audio app that creates beautiful and thought provoking audio landscapes from your the sounds in your immediate world. Even after a 1/2 hour of recording and fiddling with the sliders the results are very interesting and got me thinking and made me more aware about my own personal audio environment. My only quibble is that i would like to able to save the result and have a clearer "recording" button. But the button is a minor thing and its probably just me. Great App!
>
> "Get this app"
>
> 5/5 stars
>
> by Fritzyqnoodleman - (Canada) - Sep 3, 2012

> Great application, have to work with it more to get the hang of it.
>
> "Brilliant app"
>
> 5/5 stars
>
> by Sha. 1 - (UK) - Aug 10, 2013

> You can use this app with speakers too if you place the mic away from the speakers.
>
> "Interesting app"
> 4/5 stars
> by TruthOverFear - (USA) - Jan 20, 2012

> The sound app. I actually quite liked this one in the sense that, very quickly, I had a basic understanding input/output! I am not a sound person at all. After selecting, I think three tracks, I realised what was happening. As someone who is not into apps, it's excellent that it worked quickly and that I didn't have to spend too much time grasping what was happening - once I realised what was happening I thought, this is interesting. If I were really into sound, I can see myself getting into selecting tracks which would "mix" a bit better, it has the sort of vibe similar to my current obsession with Scrabble on the Ipad - you use your brain, there is a sort of strategy involved and an end result. For tracks I used AC/DC, Motorhead and Calibre. For some reason the AC/DC track (black in black) went crazy with the "number of voices". Very weird, but the other tracks didn't.... Yeah, interesting app and why AC/DC went crazy interests me and yeah, good.
>
> by Anonymous - Aug 22, 2013

4.3.4.2 Personal Experiments

As well as releasing the app to the public, I have explored the app in a series of sonic collages. These pay homage to John Oswald's own experiments with sonic collage, as they use Michael Jackson's "Beat It" for the target. The first of these had the app learn a nature recording by Chris Watson. Afterwards, I set the app to stop learning, and had it listen to "Beat It" while I recorded the resulting synthesis. Birds, leaves, squeaks, and other fragments of nature sounds are mashed into a cacophony strangely resembling the original "Beat It" track. Likely, the rhythmic features of "Beat It" are the strongest indication of the original track. Another possible indication may come from a simple

hallucination of hearing a voice which certainly cannot possibly be there. The track has so far amassed only 76 plays. Other experiments in the same series all used "Beat It" for the target, and include "Michael Jackson versus Liszt's Hungarian Rhapsodies" (37 plays), "Michael Jackson versus Rachmoninov's Piano Concerto No 2 and 3" (431 plays), and "Michael Jackson versus Michael Jackson" (1946 plays). These can all be found online at: https://soundcloud.com/pkmital/sets/michael-jackson-experiments.

Waveforms of original and mosaic? The original waveform is actually really clipped...

4.3.5 Discussion

Memory Mosaic presents a significant step towards the final aim of this thesis, an audiovisual scene synthesis, completing the conceptual framework described in Chapter 2 by incorporating Event Detection. One important issue was discovered in moving the framework towards a practical implementation, namely, segregation in real-time could not be achieved due to limitations in CPU power. Ideally, a better implementation or faster processor could afford such an interaction.

4.4 Conclusion

This chapter has discussed two practical outputs employing auditory scene synthesis: (1) The Daphne Oram Browser, which displays The Oram Archive as a 3D interactive visualization allowing for real-time scene synthesis via a simple mouse cursor; and (2) Memory Mosaic, a mobile application for iOS allowing one to create real-time scene synthesis aggregated from a microphone or songs from the iTunes Library.

Chapter 5

Conceptual Framework for Building Unconscious Visual Representations

Contents

5.1 Introduction

This thesis has so far seen how scene synthesis may occur within audition. Certainly, the issue of perception tackled by our brains must functionally find similar outcomes in vision as it does in audition (e.g. discover representable aspects of the world and be able to function with them). Still, these domains require individual treatment as we employ fundamentally different behaviors when using either modality. One obvious

reasoning for this is that movements of the eye allow us to spatially orient within the visual world, whereas within audition, no similar action exists. However, this should not be taken to mean that these domains are independently processed in the brain. Clearly, many multisensory and crossmodal effects have been demonstrated (e.g. the auditory override of visual perception as demonstrated in the double flash illusion (Shams 2002) and the visual override of auditory perception as demonstrated in the McGurk effect (McGurk 1976)). However, only so much can be explored within this thesis, and we therefore work towards a modal solution in the hopes that the interactions within the two modalities could be explored in the future.

Similar to aims in developing the framework for audition in Chapter 2, conceptualizing an overarching framework for visual perception is beyond the scope of this thesis. Even if it were within the scope, we have only really begun to understand the some 1 million retinal ganglion cells (Curcio 1990) or the estimated 140 million neurons in early visual cortex (Leuba 1994), let alone how the remaining 5 higher cortical layers or inferior temporal cortex and fusiform areas may possibly effect early layers of processing. There is a great amount of work to be done before a comprehensive model can be achieved. Therefore, this chapter limits its discussion to seminal research within the physiology and behavioral psychology of visual perception. From these select few readings, we aim to develop a plausible conceptual architecture that will inform the implementation of the computational model developed in Chapter 6 and used in practice for producing visual scene syntheses in Chapter 7.

5.2 Background Literature on Visual Perception

5.2.1 Early Theory

In a seminal thesis on visual perception in 1915, Edgar Rubin describes a fundamental form of experience consisting of a figure standing on a ground (Rubin 1915). The figure describes the focal or fundamental experience of a scene, whereas the ground describes the ambient or marginal portions of a scene. Expanding this point further in the 1920's, the Gestalt psychologists developed a comprehensive perceptual theory employing figure-ground as a fundamental type of perception where the notion of a Gestalt, or totality, is described by a set of rules describing the formation of a figure and ground. These rules

included:

- **Similarity**: Entities sharing physical properties in terms of size, color, or other visual features, are grouped to form the same figure
- **Proximity/Contiguity**: The smaller the distance between two entities, the more likely we are to group them together
- **Symmetry**: Objects in symmetrical alignment can be perceived as the same figure even if they do not have close proximity to one another
- **Good continuation**: Grouping of a pattern of visual features beyond what we can see
- **Common Fate**: Grouping of the same temporal movements
- **Closure**: Filling in missing pieces to complete a figure

Effectively, the "Gestalt" describes the fundamental experience of perception. As a result, any subdivision or interrogation of a part of the Gestalt would alter experience into yet another figure and ground relationship (Koffka 1935; Köhler 1947). Though these rules provide a good logical understanding of how figures may be described, understanding how our brain represents or discovers such relationships is still a very open question.

5.2.2 Visual Physiology

From these early studies describing logical rules for grouping, visual perception research has continued in trying to understand how the brain could possibly encode the formation and understanding of such figures. The physiology and behavior of the eye alone has given researchers an enormous understanding of the processes supporting perception. Starting with the physical wavelengths of light entering our eyes, we rapidly shift our gaze an average of 3-5 times a second, completely disrupting the continuity of light entering our eyes. Visual acuity limitations mean that our eyes require rapid ballistic movements of the eye taking all of 30 ms (a *saccade*) to project the light from the particular point of a visual scene we are interested in onto a 2-degree area of the retina

with the highest spatial resolution (the *fovea*), finally stabilizing our eyes to the region of interest (a *fixation*), a process lasting on average 330 ms. During this time, it is thought that the encoding of details at the point of fixation into memory occurs as well as planning of the next eye-movement (Henderson 2003; Findlay 2003a).

Going away from the fovea (the *parafovea*), resolution for spatial detail drops logarithmically, while resolution for motion detail increases. This relationship is explained (1) by the refractivity of light afforded by the lens of the eye itself, (2), the distribution or density of cells in the fovea is more tightly clustered than in the parafovea, and (3), the response properties of the individual cells (also defined as "receptive fields" (Sherrington 1906)) of photo-receptive cells including retinal ganglion and their connections along the optic tract to lateral geniculate nucleus (LGN) cells (Curcio 1990). Other connections from the ganglion cells in the eye extend along the optic tract to the superior colliculus where it is thought that encoding of the control of eye-movements in retinotopic coordinates occurs (Klier 2001).

Connections from the ganglion cells travel along the optic tract eventually extending all the way to the back of the brain to the occipital lobe where the visual cortex resides. Seminal research demonstrating physiological evidence of the receptive fields and the discovery of 'simple' receptive field cells in the early visual cortex (the first of six cortical layers, denoted as V1) of cats (Hubel 1962) and monkeys (Hubel 1968) has helped to better understand the early processing occurring in the brain (for which David Hubel and Torston Wiesel later won the Nobel Prize in 1982). The authors explain that:

> These fields were termed 'simple' because like retinal and geniculate fields (1) they were subdivided into distinct excitatory and inhibitory regions; (2) there was summation within the separate excitatory and inhibitory parts; (3) there was antagonism between excitatory and inhibitory regions; and (4) it was possible to predict responses to stationary or moving spots of various shapes from a map of the excitatory and inhibitory areas.
>
> (Hubel 1962)

In other words, they demonstrated the ability for such cells to effectively respond with higher magnitudes whenever presented with contrast changes of certain sizes and

orientations, depending on their contrast specificity, and to more complex gratings that respond to the change of a stimulus in all directions (center-surround). Despite these early findings, to date, there is still no generally agreed upon classification of cortical receptive fields within the scientific community. However, one that has been most widely used is the response to sinusoidal gratings at different scales and orientations. These are typically modeled by Gabor filters (2D Gaussian multiplied by an oriented sinusoid, where the positive regions of the sinusoid refer to the excitatory and the negative to the inhibitory regions), however even this widely used approximation has raised some criticism, as it bypasses modulations occurring between the retina and V1 (i.e. within the Lateral Geniculate Nucelus) (Azzopardi 2012). Effectively this literature has demonstrated that cells in V1 can respond in a selective manner to a variety of different features in the visual scene including line orientation, direction and speed of movement, luminance and luminance contrast, color and color contrast, retinal disparity, and spatial frequency (Rao 1999).

Cortical analysis of motion stimuli also shows that the majority of V1 cells have selective response to motion in different orientations before sending their output to the medial temporal cortex (Palmer 1999). It thus does not seem surprising to find that evidence based on search efficiency has shown that our visual system is able to notice moving objects even if we are not looking for them. Such phenomena associated with parts of a scene that eventually attract our attention and gaze are often given the label "pop out". It is further thought that given evidence of the electrophysiology of monkeys, the encoding of an actual map prioritizing locations likely to "pop-out", also denoted as a *saliency* map, is likely to exist in the lateral intraparietal cortex (LIP) and the frontal eye fields (FEF) (Gottlieb 1998; Kusunoki 2000; Moore 2003; Thompson 2005). Saliency representations in LIP are thought to encode either abrupt onsets, stimulus motion, or task-dependent, behaviorally-relevant stimuli (Gottlieb 1998; Kusunoki 2000). Other studies have demonstrated that the electrical stimulation of sties within retinotopic maps in FEF were likely to evoke saccades to the retinotopic site of the target and could even suppress responses coming from visual cortex (Moore 2003). A later study also demonstrated that responses in FEF were independent of any explicit saccade command signals (Thompson 2005), further suggesting that the FEF is likely to encode attentional biases and modulate ongoing visual processing.

As evidenced by the physiology of the eye, LGN, and early visual cortex, our periph-

ery has been engineered for high motion resolution rather than high spatial resolution. Therefore, we cannot encode with high spatial detail an entire visual scene at once, as a camera with a small aperture may be able to do. Instead, we require an active viewing of a scene in order to perceive the details of a scene and use saccades and head-movements in order to perceive with greater detail the parts of a scene required for our ongoing tasks. Many researchers in visual cognition have therefore focused their research in understanding eye-movement behavior, investigating why, where, and under which circumstances we move our eyes. Loosely, this research can be defined by the field of visual attention.

5.2.3 Visual Attention

The earliest studies (Buswell 1935; Yarbus 1967) describe two main influences of a viewer's attention to a visual scene: (1) influences dependent on mental states which focus attention towards contextually and cognitively relevant aspects of the world (*endogenous*), and (2) influences dependent on involuntary capture of attention from the external environment (*exogenous*). As exogenous factors are involuntary, one would expect to find the behavior influenced by these factors to be highly consistent across viewers. In contrast, as endogenous influences are dependent on cognitive factors resulting from emotion, memory, language, task, and previous experiences, the relationship of a scene and one's endogenous influences on the scene are much less consistent across viewers.

5.2.3.1 Exogenous Influences on Attention

In seminal work investigating the speed of visual perception using Gestalt-like primitives, Sziklai demonstrated the human visual system exhibits an attentional bottleneck of 40-bits per second on selected information, suggesting our visual systems require a simplified representation from the many megabytes per second of information coming from exogenous visual information (Sziklai 1956; Merrill 1968). Much research investigating exogenous influences on static visual scenes therefore describe a simplified representation of attentional control in what they originally called *bottom-up* models (Koch 1985; Itti 1998; Wolfe 1989; Itti 2001). These models attempt to explain how we possibly allocate processing or plan a future saccade to certain regions, features, or

objects in a visual scene while also dealing with the many overlapping, occluding, or other difficulties in continually recognizing objects in natural scenes.

Early bottom-up models are built upon theories of feature-integration (Treisman 1980) and are modeled based on the response properties of simple receptive field cells. There are also denoted in the literature as "saliency" models of attention, though this was before the discovery of attentional maps within LIP and FEF. To discover the attentional biases for portions of a scene, bottom-up models start with a camera image and filter it using a series of filter banks tuned to multiple frequency orientations and scales to produce a set of conspicuity maps. Effectively, this process discovers oriented edges and color contrasts within an image at different orientations and scales. Saliency is then defined by a weighted linear summation (*integration*) of the resulting conspicuity maps.

It is further suggested that the final integration within bottom-up models are modulated by what Itti et. al calls "top-down" influences (Itti 1998; Itti 2001). Within their model, they implement an inhibition-of-return measure which is based on evidence of longer reaction times measured with the return to a previously cued or fixated location (Posner 1984; Posner 1985; Klein 2000), though cite that other influences from the current ongoing task (Yarbus 1967; Smith 2011) and the context of a scene in order to reduce processing load (Henderson 2003; Torralba 2006) are likely to effect this map.

If we take these bottom-up models as a literal translation to early cortical filters, and the resulting integrations as those denoted in FEF or LIP, then the level at which top-down influences may actually affect processing is still an open research question. For instance, it may be that top-down effects do reach V1 or more likely at least V2, thus effecting the formation of filter bank maps (e.g. (Rauss 2011)). Further, though these modulations are often described as top-down influences, such a term should not be confused with endogenous influences, as much research has shown that memory, context, and other endogenous factors affect early visual processing (Tatler 2011) which would correlate with initial feature stages thought to be unaffected in a bottom-up model.

Gist-based models (presented later) separating "conceptual" and "perceptual" influences may similarly be looked at in terms of "bottom-up" and "top-down" influences, or even "exogenous" and "endogenous", respectively. Another common distinction is the notion of 'covert' attention, or influences dependent on internal visual attention sys-

tems or the ability to attend to a peripheral location, and 'overt' attention, essentially eye-movements. However, even this distinction has been criticized as they may not be independent given the complicated and interconnected nature of attentional mechanisms (Findlay 2003b). As a result, there appear to be a variety of confounding terms introduced by this literature which were intended to help the understanding of processes in the brain. However, they all vary slightly in their use and meaning across literature, which should be no surprise given the complicated nature of the brain. We therefore attempt to restrict our use for the purposes of this thesis and refer to only endogenous and exogenous processes (as other researchers now also do, e.g. (Soltani 2010a)), as these terms seem to offer the most useful definition separating influences originating from the brain and from the sensorial world, respectively, and other distinctions, e.g. bottom-up/top-down, are not used given their possible ambiguous meanings.

Early saliency models focused their investigation of attentional biases towards static scenes. Rosenholtz however investigated how attention would shift in the context of motion-based stimuli. In her 1999 study, she created a measure of saliency based on the extent to which the motion of a scene differed from the general pattern of the scene (Rosenholtz 1999). She showed that a simple model measuring motion outliers can detect motion pop out phenomena reliably. As well, Itti et al. has since incorporated measures of motion into the most recent versions of the iLab Neuromorphic Vision C++ Toolkit for their saliency computations (Itti 2005b).

Despite significant efforts, some researchers have claimed that the attentional biases within early saliency models would not translate to the real-world (Henderson 2003; Tatler 2009a; Tatler 2011). One reason is suggested by the fact that the stimuli used to motivate the early saliency models were composed of randomly placed letters such as t's and l's. Certainly the responses within V1 may describe a priority map within such simple scenes, as one misplaced letter would create a strong center-surround response in V1. In fact, it has even led some researchers to suggest that V1, V2, or V4 may be the site of the saliency map, despite only ever testing their hypothesis on simple psychophysical stimuli (Li 2002; Soltani 2010b; Zhang 2012). Further, endogenous influences to saliency as modeled in early saliency maps were primarily based on the notion of inhibition of return in static picture arrays (Posner 1984; Posner 1985; Itti 1998; Itti 2001). However recent research suggests this inhibiting effect does not occur in dynamic scenes (Scotia 2010) or given the presence of sudden onsets (Smith 2009). Essentially, the real-

world is not composed of simple statistical features and is in fact much more complex and certainly dynamic, if not even for our own motions within it. As a result, there has been a trend to move away from understanding behavior in simple psychophysical scenes and to use eye-tracking within real-world complex scenes to better understand behavior (Henderson 2003; Tatler 2009a; Tatler 2011).

Certainly, behavioral evidence demonstrating exogenous attentional biases within dynamic scenes is greatly evidenced by eye-movement recordings (Itti 2005a; Carmi 2006; Mital 2011). In a previous study, we investigated the contribution of a set of different visual features (including those commonly implemented in saliency models) to gaze location during free-viewing of dynamic scenes (Mital 2011). We also attempted to measure the correlation of gaze behavior with a measure of optical flow (Horn 1981). Optical flow simply attempts to measure a frame-to-frame registration, discovering how luminance values in a scene shifts thus trying to recover not just the magnitude of motion but the oriented vectors of movements in the recorded image over time. It also controls for consistency in a scene by ensuring smoothness across a scene similar to processes of lateral inhibition, while still affording sharp discontinuities likely to occur at object borders through a data penalty term (often l1-normalization). Our results indicated that foveated locations could be discriminated from control locations (randomly sampled from the distribution of all foveations) by optical flow as the strongest predictor, beating all static features including oriented and differently scale edges and corners, mid-level features thought to be encoded in V2 such as corners, and even simple dynamic features encoded in many saliency models such as flicker (simple frame-to-frame luminance differences).

put in diem investigation roc curves?

5.2.3.2 Endogenous Influences on Attention

Despite the overwhelming influences of certain visual features such as motion to attract our attention, simple task manipulations can easily override these biases. One of the first demonstrations of an endogenous influence on eye-movement behavior investigated eye-movements during static scene viewing (Yarbus 1967), Yarbus tracked the eye-movements of participants viewing a painting entitled, "An Unexpected Visitor." His study showed that when participants viewed the painting and were given a task such

as to determine the ages of the people in the painting, they looked more at the faces of each person. When asked to determine what they were wearing, their eye-movements strayed away from faces, and looked more towards the clothing of people. Yarbus further describes 7 different tasks and shows how the eye-movements of each participant reflects the information required for processing the task at hand. Since then, numerous studies have expanded on their results, suggesting endogenous influence that can override exogenous influences (Tatler 2009b).

In a similar study though with dynamic scenes instead of a static painting, we investigated task-based effects on viewers' eye-movements looking at unedited videos of natural scenes from a camera mounted on a tripod (Smith 2011; Smith 2013). Participants were natives to the city of Edinburgh and viewed a variety of indoor and outdoor scenes from the city. Our study revealed that during free-viewing, i.e. not given any task other than to look at the video, participants looked at mostly moving objects such as people moving across the frame or cars. However, when given the task to identify the location of the presented scene, participants had to concentrate their gaze towards the elements of a scene depicting landmarks such as buildings, signs, and trees and showed a remarkable ability to distract away from moving objects. After viewers pressed a button indicating recognition of the location, their viewing behavior reverted to resembling the free-viewing task, fixating on moving objects such as people and cars again. Our study also re-asserted the findings of Yarbus, though for a dynamic time-course. Further, it also provided evidence of default viewing conditions during the time-course of viewing, as participants were able to "return" to the free-viewing task after having finished the task of recognizing the location of the scene.

attention is indexed by proto-objects, reference roi analysis, ante's study jov 2010 building up representation .. fd/sa binding attention to maintain.

5.2.4 Gist

Given that we are constantly moving our eyes and radically altering the light coming into our retina, how much can we encode in the short amount of time we spend when fixating? Researchers wanting to better understand the same question created a paradigm called rapid serial visual processing (RSVP), where a fixation cross would appear followed by an image that would be presented for a short amount of time. In their original study,

participants were tested for details of a scene presented for only 100 ms and performed very well (Potter 1969). When they were asked to do the same thing given a series of images that were presented rapidly and with the same duration, their performance went to chance levels (Potter 1976). However, when given either a pictorial or verbal a cue to the image to remember details of (*priming*), their performance went to significantly better than chance performances with the pictorial cue performing at about 80% and the verbal one at 60% (Potter 1969; Potter 1976). As well, it should be noted that despite the 100 ms presentation time, it would take participants a longer amount of time before they were able to report the content of the scene. This is likely due to the limits of conscious processing (e.g. as demonstrated by the P3 ERP component).

From these studies, it was understood that we can rapidly encode information about a pictorial scene with a presentation lasting only 45-135 ms (*Gist*) (Potter 1969; Biederman 1974; Potter 1976; Schyns 1994; Henderson 1999). Extensions within this body of research have suggested that we likely encode the general shape and structure of a scene in order to infer its context. Further, a number of theories have suggested that this shape is defined by either volumetric forms (*geons*) (Biederman 1987), spatial arrangement of blobs defined by contrasts in luminance or color (Schyns 1994; Oliva 1997) or by using a scene's spatial frequency content (Oliva 2001; Oliva 2005). Spatial frequency content can be described by oriented band-pass filters: at a low spatial frequency, this content resembles broad edges and the layout and orientations of a scene's largest similarly textured regions, whereas at a high-spatial frequency, the response of the sharpest edges and their directions are encoded.

Endogenous influences on gist processing seems to influence the spectral scale at which gist is selected (Schyns 1994; Oliva 1997). Schyns and Oliva describe an experiment where a low-spatial frequency (*LSF*) and a high spatial frequency (*HSF*) image are created for two separate pairs of images. Creating two new images by combining the LSF of one image and the HSF of the other, and vice-versa, they investigate the scale space of gist recognition with and without a verbal cue to indicate what type of scene will follow. Without priming, subjects are able to recognize the scene described by the LSF content of an image given 45 ms of presentation time, and the HSF one within 135 ms. As well, subjects are unaware of the content in the other scale space (i.e. shown an image with LSF and HSF content for 45 ms, the participants are unaware of there being separate HSF content). However, being primed with either the LSF or HSF content of

the scene, subjects report perceiving the given cue instead.

While gist is thought to be pre-attentive, i.e. before the timescale of acts of selective attention, such research suggests either that (1) the scale at which the early representation of gist operates at is affected by task-demands (i.e. only one scale of gist is encoded for pre-attentively), or (2), attention and further encoding into memory is dependent on endogenous influences on scale selection, (i.e. gist may be encoded at multiple scales, but only the scale selected by attentional machinery is encoded into memory). Though not all scales are necessary for determining a scene's content when given prior cues (*priming*), the neurobiology of early visual cortex gives scope for encoding of multiple visual scales. It thus seems possible to assume (2) is a more likely model for the interaction of gist and attentional machinery.

Gist has been understood in terms of the classic rapid serial visual paradigm (RSVP), or within a screen based presentation where stimulus presentations are preceded by empty or noisy screens. However, the real-world is not preceded by such screens, and rather the notion of gist across saccades in a situated and embodied world becomes an important one to make. How does a dynamic real-world model of gist help us to maintain a coherent perspective of the world, and further guide our attention? Research demonstrating our failure to notice dramatic changes within static and dynamic scenes may help us to decipher this question better.

5.2.5 Change and Inattentional Blindness

Specifically, research demonstrating the failure to report large changes in the visual world (*change blindness*) as well as the failure to report unexpected visible changes due to task requiring attention elsewhere (*inattentional blindness*) (Simons 1999; Rensink 2000; Rensink 2001; Hollingworth 2001) have shown that our visual systems are unaware of changes in visual world outside of the point of fixation. Simons and Chabris demonstrated *Inattentional Blindness* by composing a video of two basketball teams dressed in white and black passing a ball to each other (Simons 1999). Participants were asked to count the number of passes that the white team makes. During the course of the video, a person wearing a gorilla suit walks across the frame of the camera. However it was unnoticed by 75% of participants. The selective attention mechanisms therefore seemed to not be able to encode the gross contextual violation of a gorilla within a

basketball scene. This seems to reveal some indication to the nature of representation outside of selective attention. Namely, the semantic details of the entities outside of fixation are either (1) not encoded, or (2) encoded, but not able to be compared to any other semantic entities (as this would trigger some contextual violations, thus capturing attention).

Another fascinating "failure" of the human visual system to detect large changes in the visual world is demonstrated by the phenomena of *Change Blindness*. As demonstrated in a real-world psychology experiment (Simons 1998), participants arrived at a kiosk to fill in a consent form and handed the completed form to a man behind the counter. The man ducked behind the counter as to pretend to file the paper, while a different man came up from behind the counter, again unnoticed by a majority of the participants. In a similar demonstration, an RSVP paradigm displayed two different images separated by a transient of white-noise. The white-noise effectively removed the capability of discovering the change between the two images. That is, under a normal viewing circumstance, a gross change in a visual scene would likely capture attention due to the onsets of motion and changes in visual features of luminance and color. However, removing these transients by placing a scene of white noise effectively removes the ability to detect transients of the previous scene. Moreover, the memory representations retained from the scene pre-white-noise are such that post-white-noise, they are not challenged in any way (since they do not capture attention).

In both change and inattentional blindness, the failure to detect changes outside of the point of fixation suggests that any peripheral representation of a scene would likely not encode details of object specific features such as color, motion, or orientation gratings. Rather, our visual machinery integrates the detailed aspects of objects across eye-movements, retaining that information as a perceived representation of the visual world. The broad spatial scale afforded by the lens of the eye and the higher motion resolution afforded by the spacing of cones in the periphery give further indication to the lack of highly detailed feature encoding in the periphery. Though, to what form, and to what detail a periphery representation may encode is still an open question.

5.2.6 Visual Object Representation

Despite the numerous occlusions, lighting changes, and possible angles at which we view an object, we continue to be able to recognize objects in all their possible slight variances. We can even extract object representations from clouds, turkish coffee, and even linen cloths (e.g. Shroud of Turn [1]). The incredible feat here is not the grouping strategies for determining boundaries of an object, as suggested by the Gestalts, but rather the binding of these groups into semantic labels.

Rensink takes evidence in change and inattentional blindness in developing a theory of coherence, proposing that object representation depends on focal attention. He suggests that once attention is focused, a "coherence field" is established, linking the details at the point of fixation to a semantic object representation. For objects outside of the point of fixation, Rensink proposes we encode volatile units of *proto-objects* (Rensink 2000; Rensink 2001). Proto-objects are argued to be amorphous and blob-like in nature, representational-less and concept-less lasting only a few hundred milliseconds. It is further argued that attention operates on groupings of proto-objects rather than at the earlier feature levels making it the highest level of early vision, and the earliest operands of selective attention. Rensink also hypothesizes that proto-objects may explain non-attentive processes capable of recognizing the abstract meaning of a scene and the spatial layout of the scene (Rensink 2002). In relation to perceptual influences, implicit behavioral measures suggest that grouping processes can also occur for task-irrelevant visual stimuli, i.e., for stimuli that has not been attended to by a fixation, further supporting theories of proto-object formation (Lamy 2006).

Object files, described by Kahneman and Triesman (Kahneman 1984), suggests that visual objects are hierarchically described, with the higher levels denoting the semantic description of the visual object (e.g. a group of dancers), mid levels describing any possible parts making up the entity (e.g. individual dancer or the hand of the dancer), and the lowest levels describing the spatiotemporal constraints of the object itself. Their theory suggests that attention describes the level that object files are accessed, and that they are kept open for a limited time as attention is withdrawn, effectively creating a limit to the number of object files that are kept open. Coherence theory certainly resembles object files, though with one important distinction that once attention binds

[1] en.wikipedia.org/wiki/Shroud_of_Turin

proto-objects into the coherence field, the field collapses as soon as attention is removed from it, leaving only the volatile proto-objects to collapse shortly after. Thus, in Rensink's model, only one stable semantic representations seems to be active (in the coherence field), whereas in Object Files, there can be many open at any given time.

Though these theories work towards an understanding of how the various processes of perception (e.g. attention, gist, representation), may interact, they do not yet paint a complete picture for describing a truly computationally implementable model. This is because the descriptions of "spatiotemporal" constraints in object-files or the "groupings" of proto-objects are open to interpretation. However, some understanding may be gained when considering Marr's seminal work in describing shape invariances (Marr 1978; Marr 1982). Marr considered that perception must be 3-dimensional, and the goal of perception is therefore to reconstruct the images coming into the retina as a 3-dimensional scene. He describes a transition from lines, to contours, to surfaces, and finally to volumes. At the end of this process, Marr suggests that the representations held in the brain are defined by *view-specific* "sketches" of this volumetric reconstruction. Therefore, to recognize an object, we would have to have already encoded a set of "sketches" that were similar to the current view of that object.

Biederman's Recognition-By-Components (RBC) offers another framework towards understanding representations, suggesting that the recognition of an object should be *view-invariant* so long as one can encode an object's geometric-ions (*geons*) (Biederman 1987). In Biederman's case, the representation of an object entails the encoding of a set of shapes in 3-dimensions, such that their relative groupings are also encoded (e.g. a sphere and a cylinder in some relationship describe a baseball bat). In contrast to Marr's approach, one would only have to have seen one or a few views of an object, encoded their geons, and then any novel view of it would be decoded so long as the same geons could be detected.

The distinctions between Marr and RBC-based approaches to object representation have continued to present day, labeling one camp as view-based, and the other as object-based. It is likely that the human visual system is capable of discovering objects in either case, though discovering this is beyond the scope of this thesis. What can be gained from this discussion, however, is the importance of shape to early object representation in either case.

Shape in cortex...

Object in cortex...

Neural Populations of Object Classes in Inferior Temporal Cortex...

5.3 Conceptual Framework

Taking this literature forward, we are now in a position to motivate a conceptual framework for a computational model of visual scene synthesis. In particular, this literature has demonstrated the importance of an active viewing of a scene. However, as we may not have eye-movement information about a scene, we will need to build a model of where people are most likely to attend, allowing us to describe any realistic visual scene in terms of higher spatial acuity at a point likely to be fixated. In other words, we will restrict encoding details of a scene unlikely to be attended. Following Rensink, we assume a representation for encoding and decoding described by volatile units of proto-objects within the scene. Finally, these units will be described by representations encoding properties of their shape, color, and luminance.

5.3.1 Exogenous Attention Model

How this is a proxy for a very simple estimation of change, but we are very much capable of encoding higher order regularities... An example of how this is not true, a case where motion patters are learned, but some static pattern appears as being very unregular... No sense of semantic or contextual relevance, but this could be added by discovering scene layout... prioritize gaze to horizon... people... text... faces...

The entropy of the motion correlates to entropy of gaze locations... use this value to drive the acuity map then...

5.3.2 Visual Acuity

Behavioral evidence from change blindness studies has indicated an essential gap in our encoding of a scene, namely the precise details of a scene are only encoded at the point of fixation. This is likely due to the computational complexity required for encoding an entire scene's high-spatial resolution information. More importantly, it is

physiologically impossible given the visual acuity limitations of the retina. As discussed in (Kalloniatis 2007), a number of factors effect visual acuity: refractive error, size of the pupil, illumination, time of exposure of the target, area of the retina stimulated, state of adaptation of the eye, and eye-movements. Modeling each of these would be a thesis in itself. We therefore take an overly simplified view of acuity in suggesting that a 2-degree Gaussian window describes the highly detailed regions of foveal processing, from which a logarithmic falloff is a plausible general approximation.

One could design visual acuity as a method of "focusing", if no prior eye-movements onto a scene are known.

5.3.3 Proto-objects

Spatial grouping of color, motion, and luminance cues. Builds a shape representation.

5.4 Discussion

Research in change blindness has indicated that though we experience a rich, detailed visual world, we do not use such rich details in building a stable representation (Simons 1997). Rensink argues that object representation requires focal attention. However, in considering an architecture of visual perception, what is the cause of producing focal attention? The literature presented here suggests that there is either an endogenous explanation or exogenous one.

When considering evidence for gist in relation to Rensink's theory of coherence, it seems viable to consider proto-objects as the same representation that gist may use (Rensink 2002). Though Schyns and Oliva argue for using oriented banded filters, it is not unlikely that collections of blob-like entities which necessarily also respond to the scale of the proto-object could provide a cue for spatial layout. However, when considering evidence in rapid determination of the meaning of scenes, Schyns and Oliva demonstrated that early processing of a scene could be re-organized based on prior experiences (Schyns 1994; Oliva 1997). Thus, it is not clear from their research alone whether the pre-conceptual representation itself can be changed, or if only the attentional machinery acting on a set of possible representations has changed. The latter effect would entail a sort of conceptual prior on a scene, suggesting the organization of a

scene's early representation remains untouched.

Pylyshyn theorizes that the understanding of a concept is not all that is required for visual experience:

> "Vision suited for the control of action will have to provide something more than a system that constructs a conceptual representation from visual stimuli; it will also need to provide a special kind of direct (preconceptual, unmediated) connection between elements of a visual representation and certain elements in the world. Like natural language demonstratives (such as 'this' or 'that') this direct connection allows entities to be referred to without being categorized or conceptualized. (Pylyshyn 2001)"

The preconceptual connections Pylyshyn describes are easily described by the pre-attentive proto-objects Rensink also describes (Rensink 2000; Rensink 2001). What is interesting in Pylyshyn's theory is the notion that this pre-conceptual representation does not need to be categorized or conceptualized in order to be referred to. In other words, the categorization which Pylyshyn theorizes of is part of the attentional machinery which refers to proto-objects, rather than an explicit property of the proto-object themselves. According to Pylyshyn's theory, proto-objects of a visual scene are then described by one particular fate, and attentional mechanisms can only select from the set of possible proto-objects, rather than influence their definition.

Proto-objects act as an indexical reference to a conceptual referent. How proto-objects are defined is based on the features that best give coherence to the scene. They are also scaled with increasing eccentricity, presumably at least because of the shape of the eye would decrease spatial resolution, but as well due to the nature of attention acting at the point of fixation. Functionally, visual material outside of the region is of less importance to the task at hand. Thus, the biological nature of the shape of the eye could also be understood in terms of evolving to act in the world. Proto-objects may also possibly find re-definition from the act of attention. As proto-objects are volatile in nature, the act of attention reshapes ones perspective of a scene, thus reorganizing the boundaries defining proto-objects, even at the point of attention.

5.5 Conclusion

Considering both the implicit, unmediated representation and the attentional and contextual mechanisms, at least three critical layers should be built into any computational model based on the evidence presented here: (1), a pre-conceptual representation which takes into account different possible spatial configurations, composed of either band-passed edge-oriented filters or proto-objects, (2), where this representation is affected by a logarithmic filter around the point of fixation based on the evidence of response properties of photo-receptors (visual acuity); and (3), an attentional and contextual influence supported by the ongoing experiences of the subject such that parafoveal information becomes unstable without ongoing attention and is only inferred by through the context of the scene. The intentions of an agent within this model are still not well-understood, as the variety of possible endogenous influences that may be possible are too great.

Similar computational models have been developed to explain visual perception machinery (Walther 2006; Orabona 2007), however they each suffer from a number of problems: (1) they lack the inclusion of the evidence of the response properties of photoreceptors in the retina as there is no indication of the current or ongoing attention within the visual scene; (2) they infer context based on solely a static image whereas the real-world is dynamic; and (3), they cannot distinguish groupings of proto-objects and instead create discrete maps which are thresholded as attention or saliency maps. Furthermore, the interest in the previously cited models of visual perception is in predicting attention towards a scene, rather than allowing an agent in the world to explicitly define this. In such a case, these models are unsuitable for applications in augmented or virtual reality where the agent already provides attention within a scene.

CHAPTER 6

Computational Visual Scene Analysis

Contents

6.1 Introduction

Explain assumptions... Difficult problem... future work to make it plausible/evaluate it... just setting groundwork for next chapter...

Flicker provides an exogenous attention cue during dynamic scene viewing. However, attention must act on groupings of proto-objects, rather than a pixel-based notion of change. As a result, it is likely that the flicker is modulated by spatial grouping cues.

A. get flicker for a frame B. get objects for a frame C. modulate objects based on flicker D. recompute flicker based on aggregated proto-object flow

Where do I put a background of computational visual models? I haven't done this yet.. .shit.

6.1.1 Exogenous Attention Model

Cortical analysis of motion shows that the majority of V1 cells have selective response to motion in different orientations before sending their output to the medial temporal cortex [68]. It thus does not seem surprising to find that evidence based on search efficiency has shown that our visual system is able to notice moving objects even if we

are not looking for them. Phenomena associated with parts of a scene that attract our attention are often given the label “pop out”. Rosenholtz [75] has investigated these phenomena in the context of motion by creating a measure of saliency based on the extent to which the motion of a scene differed from the general pattern of the scene. She showed that a simple model measuring motion outliers can detect motion pop out phenomena reliably. As well, Itti et al. has incorporated measures of motion into the most recent versions of the iLab Neuromorphic Vision C++ Toolkit for their saliency computations [77].

Horn and Schunck [96] optical flow is a differential method for calculating motion vectors in a brightness image. The flow of an image, I, is defined by:

(6.1)

where Ix, Iy, and It are the image derivatives, and Vx, and Vy are the components of optical flow in the x (U in vector notation) and y (V in vector notation) direction, respectively. This method is often referred to as a global method for calculating the optical flow, as the energy constraint (the first term in Equation 3) assumes gray value constancy and does not depend on local image patches [e.g. 97]. The energy equation includes a second term known as the ÒsmoothingÓ term in order to smooth flow where distortions in the flow occur. This leads to better performance at filling in ÒgapsÓ left from moving objects with similarly textured regions than local methods.

6.1.2 Visual Acuity

6.1.3 Proto-objects

Shape recognition. Skeletonization. Level sets, Boundary value problem. Agglomerative clustering.

Figure 6.1: The output of the visual acuity filter on two example frames from an idealized scene depicting a man standing up are demonstrated. (Left): Original frames; (Right): 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). This point along with the entire map's entropy, is then fed to visual acuity filter. 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.

Chapter 7

Computational Visual Scene Synthesis

Contents

7.1 Introduction

NEED TO REORGANIZE THIS CHAPTER. Technical stuff will go in the previous chapter. Outputs will be separated into different sections: (0) Early Experiments, (1)

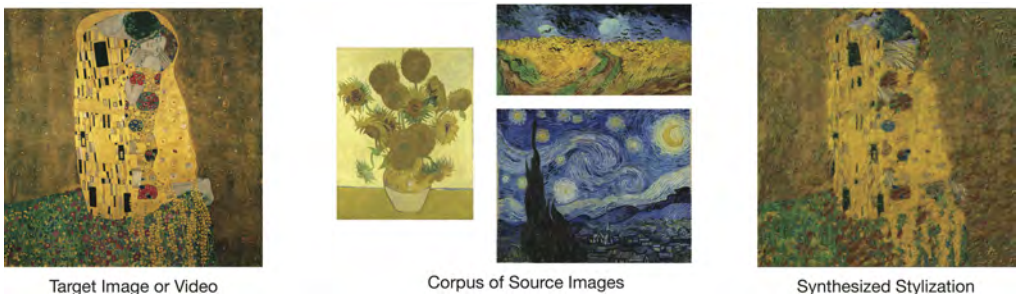


Figure 7.1: Klimt's "The Kiss" is synthesized using 3 images of Van Gogh paintings to produce the result on the right. Best viewed in color at 400%. Images representing faithful reproductions of Gustav Klimt and Van Gogh sourced from Wikimedia Commons are public domain.

Artistic Stylization of Image/Video, (2) PhotoSynthesizer iOS App (e.g. related work of Instagram, Vine, Camera Pro, etc...), (3) Memory Mosaicing

Despite its apparent precision, our perception of reality is not representative of the way that we see. For instance, the light coming to our eyes is distorted, upside-down, and constantly disrupted with each movement of the eye. How can this noisy process ever constitute our experience of the visual world? Numerous theories have argued that in order to perceive the world as a continuous and richly detailed one, our vision system must use abstracted representations of the world (Marr 1982). It is argued that these representations are created by grouping together coherent visual features that resemble abstract forms - such as geometrical primitives. Grouping such primitives together eventually leads to the formation of semantic representations such as objects. Importantly, the representations used in vision are not necessarily what we perceive, but are what we use in order to help us perceive. As a result, these representations are likely to remove details that are unimportant to a person's ongoing task while making other details more explicit.

Artists are well aware of the role of representation in perception. By leaving out particular details from a visual scene and accentuating others, they are able to direct a viewer's attention within a visual medium, influencing their perception (Haeberli 1990; Zimmer 2003). Picasso once famously said, "I paint forms as I think them, not as I see them" (Hughes 1991). As one of the pioneers of Cubism, Picasso wanted to represent the fact that our perception of an object is based on all possible views of it. He did so by compressing all views of an object into a synthesized one built using abstracted shape primitives. Other movements in art can also be characterized as

utilizing representations formed through geometrical primitives. In Impressionist painting, these forms are often described by a dense number of short and visible brush strokes. In Abstract Expressionist painting, the primitives are again dense, though tended to be of much larger strokes in an attempt to abstract away as much detail of a real scene as possible.

In this paper, we investigate an approach to the artistic stylization of photographic images and videos through the use abstracted shape representations. The representations that are built by this method can be varied in size and density using a process that allows the user to manipulate parameters in real-time. Our system first learns a database of representations from a corpus of images. It then synthesizes a target image or video by matching geometric representations in the target to the closest matches in the database. We show how changing the parameters of the synthesis process results in stylizations that represent aesthetics associated with Impressionist, Cubist, and Abstract Expressionist paintings. As the stylization process is fast enough to work in real-time, this approach can also be used to learn and synthesize the same camera image, even aggregating the database with each new video frame in real-time, a process we call "Memory Mosaicing".

7.2 Related Work

Artistic stylization has seen significant advances over the last 14 years. Kyprianidis recently surveyed the field in (Kyprianidis 2012). The field began as filtering and clustering algorithms were applied to images, accentuating regions within an existing image to produce aesthetics associated with different styles (e.g., for Pointillism (Yang 2006; Seo 2010); for cartoonization (Wang 2004b); for oil and watercolor (Meier 1996; Hertzmann 2000; Bousseau 2007; Gooch 2002); for Impressionism (Litwinowicz 1997; Hertzmann 1998)). More recent approaches focused on using user-guided segmentation, where the user manually labels key frames with strokes defining how the frame is stylized (e.g. (O'Donovan 2012)) or uses eye-movements in deciding which aspects of a photo are most salient (DeCarlo 2002).

Hertzmann's seminal work in Image Analogies (Hertzmann 2001) presented a branch from the aforementioned approaches by allowing control of the stylization process

through choosing a pair of example images. By finding the patterns associated with an existing stylization of an image A to another image A', a user could then stylize a target image B by analogy into B' (later extended to include analogies between curved strokes (Hertzmann 2002)). In the same year, (Efros 2001; Liang 2001) also developed methods in texture transfer and patch-based sampling, where existing image material was used to synthesize textures of arbitrary sizes. These methods were later extended in (Wang 2004a), where a user specified small blocks in an example painting that represented the style to recreate. These blocks were then synthesized along computed paint strokes in the target image using an efficient hierarchical texture synthesis method. Though Wang's approach and even more recent methods (e.g., (Guo 2006)) produces impressive results, it also relies on user interaction to select the representative patches expressing an artistic style. Further, the aforementioned work in texture transfer as well as more recent approaches (e.g., (Lee 2010)) all rely on a single source image in order to transfer the style of the texture, meaning the range of stylizations possible are constrained to the information contained in a single image. In this paper, we develop an approach that does not require the user to manually label any regions and that is not confined to a single example image while still affording a range of possible styles.

Our approach, corpus-based visual synthesis (CBVS), synthesizes a target image/video using existing pre-defined visual content. As a result, it is also borrows methods from dictionary-based approaches ((Zeng 2009; Healey 2004)), though our approach does not focus on developing strokes from expert training as we automatically segment a corpus of user chosen images. It also shares methodology with collage/mosaic-based work (e.g. (Kim 2002; Orchard 2008; Huang 2011; Miller 2012)), allowing a user to work with a period of an artist's work or entire videos, for example. Though these approaches are targeted for collage/mosaic-based purposes rather than artistic stylization, (Huang 2011) describes an approach that is also motivated by an artist making use of collage. Their approach produces what they call "Arcimboldo-like" collages in the style of 18th century painter Giuseppe Arcimboldo, relying on user strokes to segment the images used. In contrast, CBVS is aimed towards producing a range of possible artistic stylizations through changing a few simple parameters. Further, as segmentation happens without requiring user-selected patches or strokes, CBVS is also suitable for producing stylization of videos, unlike the very impressive though slow approach (15 minutes for a 300 x 400 pixel image) reported in (Chang 2010).

7.3 Corpus-based Visual Synthesis Framework

CBVS begins by first aggregating all frames from a user chosen corpora of images, $\mathbf{C} = \{C_1, C_2, ..., C_N\}$, containing N total candidate images. We aim to use the content solely from this corpus to artistically stylize a target image or video, $\mathbf{T} = \{T_1, T_2, ..., T_M\}$, containing M total frames. We develop a rendering procedure for image and video-based targets where parameters of the synthesis can be changed interactively. To begin, we describe detection, tracking, description, matching, and synthesis of the abstracted shape representations. We then describe parameters influencing each of these steps before showing our results in Section 7.5.

7.3.1 Detection

For both the candidate and target frames, we aim to detect abstracted shape primitives described by coherent image regions. For this purpose, we make use of maximally stable color regions (MSCR) (Forssén 2007). The algorithm described in (Forssén 2007) successively clusters neighboring pixels with similar colors described by multiple thresholds of a distance measure which takes into account the inherent camera noise and the probability distribution of each RGB color channel. Regions are denoted as maximally stable if they do not grow larger than a minimum margin for certain number of time-steps. Previous techniques employing posterization, filtering, or watershed have had to apply their algorithm at multiple scales in order to discover regions that are superimposed or overlapped, increasing their computational complexity. MSCR has the benefit over these previous techniques as it provides an implicit ordering of superimposed regions discovered through successive time-steps of the clustering algorithm. Further, it allows us to prune regions by restricting their area to a range of minimum and maximum sizes. In Section 7.4.1, we discuss these parameters in greater detail in relation to the styles they can produce. We use MSCR to detect the set of all regions in each candidate and target frame, denoted as $\mathbf{R_C} = \{R_1, R_2, ..., R_{N_C}\}$ and $\mathbf{R_T} = \{R_1, R_2, ..., R_{N_T}\}$ where N_C is the number of regions detected in all candidate frames and N_T is the number of target regions.

7.3.2 Tracking

It is often desirable to produce temporally coherent stylizations, meaning if a region within a target video frame has not moved, it is not re-stylized. This is especially the case in noisy or compressed videos, where artifacts may appear that should not be stylized. One approach would be to track regions using a GPU-based Optical Flow measure. This would likely produce reasonable temporal coherence without sparing real-time interaction. However, we simply follow (Hertzmann 2000) in using the flicker for detecting the change in the original target video, as this approach is fast and easy to compute. Let the flicker for a pixel at location (i, j) be described by:

$$f(i,j) = I_t(i,j) - I_{t-1}(i,j) \tag{7.1}$$

where I is the image luminance at time t. Then, if the flicker at the region's centroid, $f(C_{R_i})$, between the current and previous frame is greater than a threshold, $threshold$, we remove the region from the set of detected regions to synthesize:

$$R_T = \{R_i \mid f(C_{R_i}) > threshold, \forall i = 1...N_T\} \tag{7.2}$$

7.3.3 Description

We form a descriptor comprised of shape and color values. The shape descriptor for each region, d_{R_i}, is composed of the normalized central moments up to order 2. The average color of the region is converted from RGB to the 3-channel CIELAB color space, L, a^*, b^*. These form the final descriptor:

$$d_{R_i} = \left(\mu_{00}, \eta_{11}, \eta_{20}, \eta_{02}, L, a^*, b^*\right) \tag{7.3}$$

where μ_{ij} is the central image moment of order i and j, i.e. μ_{00} is simply the area, and η_{ij} is the normalized central image moment computed as:

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{\left(1+\frac{i+j}{2}\right)}} \tag{7.4}$$

Centralizing the moments allows us to compare regions with translation-invariance, while normalizing the first and second order moment allows us to compare regions with scale-invariance. We include the area as the first term as this ensures regions are not distorted too much when matching. Further, employing CIELAB allows us to define the region in a color space where we can then use perceptual metrics for matching. We describe this metric in greater detail in the next section.

7.3.4 Matching

We match each region in the target to its nearest neighbors in the database using a metric combining distances from each region's shape and color, $d_s(R_t, R_c)$ and $d_c(R_t, R_c)$, respectively:

$$d(R_t, R_c) = d_s(R_t, R_c) + d_c(R_t, R_c) \tag{7.5}$$

The shape distance is simply computed as the absolute difference between the first and second order normalized central image moments of each region (i.e. the first four components of the descriptor). For the color distance, we make use of the official CIE color-difference equation, CIEDE2000, which provides reliable color discrimination with interactive terms for lightness, chromaticity, and hue weighting (Luo 2001). This difference formula has been shown to be more perceptually accurate at determining the difference between colors than previous methods employing linear difference using RGB or LUV color values, as it is based on empirical evidence of perceived color difference. For our tests, we use the default parameters described in (Luo 2001) for the weighting terms.

7.3.5 Synthesis

To ensure regions are drawn from their background to the foreground, we synthesize each target region in order from the largest to smallest area sizes. In contrast to methods

that place brush strokes based on the stroke direction at each pixel on the medial axis (e.g.,(Wang 2004a)), we find the affine geometric transform describing the transformation from R_{C_i} to R_{T_i}. This can be described by a translation, rotation, and scaling. The translation component is simply the difference in each region's centroid. The rotation can be found using the central image moments:

$$\Theta = \frac{1}{2} * \arctan \frac{2 * \frac{\mu_{11}}{\mu_{00}}}{\frac{\mu_{20}}{\mu_{00}} - \frac{\mu_{02}}{\mu_{00}}} \tag{7.6}$$

Finally, scaling is simply the ratio of the target to candidate region's bounding box. This process has the benefit of being very fast using graphics hardware as it can be computed by a single matrix multiplication. Each region is then layered above the previous one before creating a synthesized image. In image-based stylization, multiple syntheses created with changing parameters can be blended together to create more detailed and expressive styles which may require many "layers" of "paint". We discuss these parameters in greater detail in the next section.

7.4 Parameters

Parameters influencing the region detection algorithm are set independently for the corpus and the target, as their function differs.

7.4.1 Corpus Parameters

For the corpus, we define the *timesteps*, *minimum region area*, and *maximum region area* of the detected regions. We use a set of parameters that learns the widest range of possible regions covering both small and large regions. In some cases, as in more abstract styles, it may be desirable to learn a very small number of regions, limiting the range of expressiveness to a few possible primitives. As the timesteps parameter influences the number of evolutions allowed in the MSCR algorithm, the higher this number, the more regions will be discovered. Similarly, lowering the minimum region size and increasing the maximum region size reduces the number of region that are pruned. In our tests, we found a single set of parameters to be sufficient for defining a varied corpus: 100 for

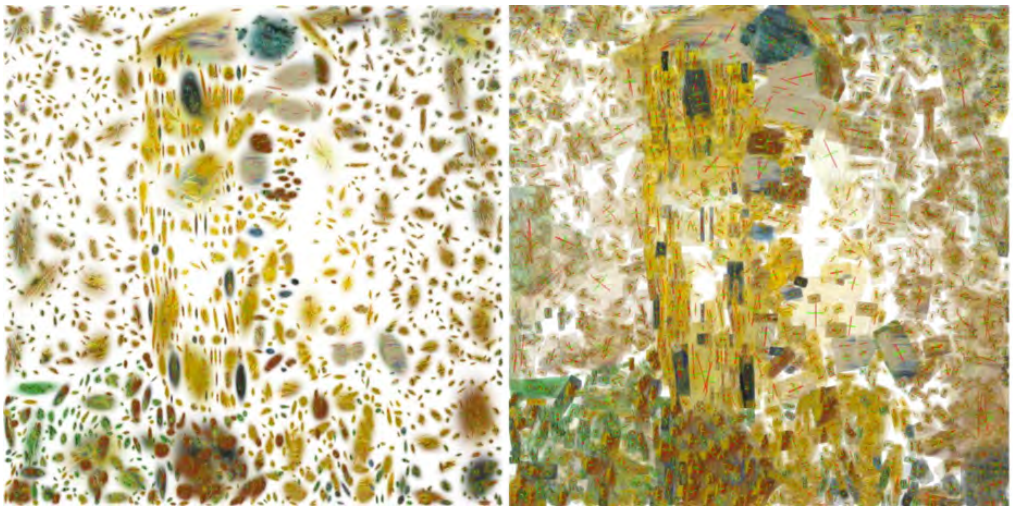


Figure 7.2: Using the target image and database shown in Figure-7.1, we show an example stylization with (first image) and without (second image) spatial blending. We also draw the region's orientation depicted by red/green axes in order to better show the regions (best viewed in the color manuscript at 200%).

the timesteps, 35 pixels for the minimum region area, and 50% of the image's size for maximum region area.

When learning a corpus from many images, we restrict learning regions that are within a distance threshold (using Equation 7.3.4) of all regions in the existing database. For our examples, we set this parameter to 50. This value is low enough to include many regions, though high enough to avoid detecting duplicate regions. A higher number for this parameter will lead to very discriminative regions. In our tests, when setting this number higher, we found that our corpus had less variety of regions to synthesize from, leading to stronger shape or color mismatches.

7.4.2 Target Parameters

For the target, we allow the user to interactively define a few parameters affecting the output stylization.

- *Spatial blending*: Allows the user to use feathered elliptical regions instead of rectangular ones (see Figure-7.2). When stylizing finer details of an image, this parameter is very useful for removing hard edges produced by rectangular regions.
- *Timesteps*: Increasing this produces more regions, making the image denser

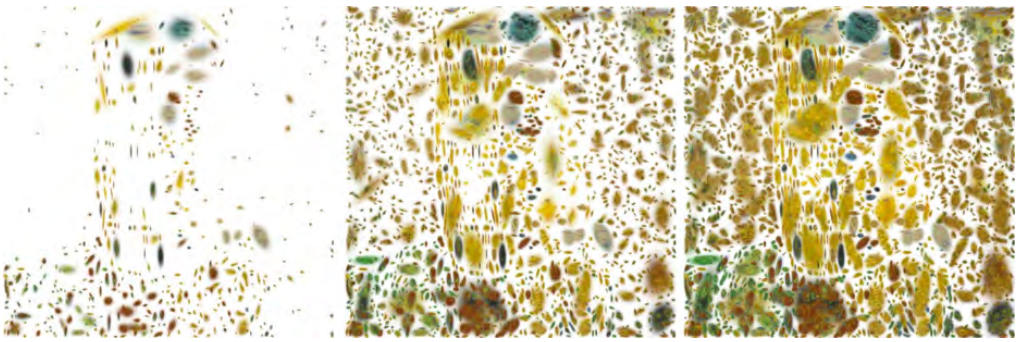


Figure 7.3: Using the target image and database shown in Figure-7.1, the timesteps are increased over time. This allows the user to detect more regions and develop a denser and higher contrast stylization.

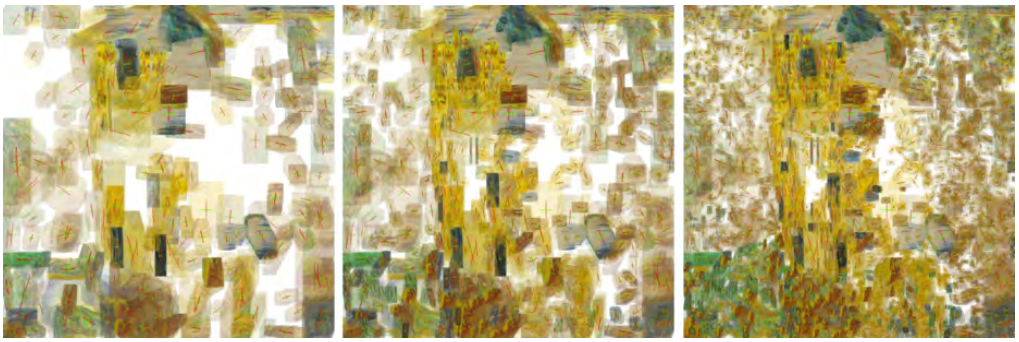


Figure 7.4: Using the target image and database shown in Figure-7.1, the minimum region size is decreased over time, allowing the user to detect smaller regions and produce finer detailed stylizations.

Figure 7.5: Using the target image and database shown in Figure-7.1, the blending radius is increased over time. This parameter influences the overall size of the drawn regions. Setting this number smaller can help to produce finer details on top of existing layers, often associated with both Impressionist and Abstract Expressionist styles.

Figure 7.6: Using the target image and database shown in Figure-7.1, we increase the temporal blending factor. This influences the opacity of every region drawn.

Figure 7.7: Using the target image and database shown in Figure-7.1, we use temporal blending as well as decreasing minimum region size and increased timesteps to begin to produce the final synthesis.

(see Figure-7.3). As well, this will also produce more regions that coincide with each other. As a result, when synthesizing with a high number for the timesteps, the result resembles an overpainting effect. For styles that require many "layers" of "paint", we use a higher number for the timesteps. When used in combination with blending, increasing this can also increase the contrast.

- *Minimum region size*: This parameter determines the minimum allowed region size for synthesis. Setting this number very low (e.g. below 100 pixels) produces styles more similar to Impressionism, as many small regions are detected (see Figure-7.4).

- *Maximum region size*: Similar to the minimum region size parameter, this parameter determines the largest allowed region size. Generally setting this number as high as possible will be sufficient. However, it may be desirable to interactively change this parameter over time, allowing for large regions to be drawn at first, then only allowing smaller ones.

- *Temporal blending*: Uses alpha blending to composite regions over time (see Figure-7.6). Together with an increased number of timesteps, this parameter can be used to change the contrast of the overall image (as shown in Figure-7.7).

- *Motion tracking*: Allows regions to be drawn only if their detected motion is higher than a fixed threshold. For our experiments, we set this number to 5.

- *Blending radius*: Influences the feathering radius of the detected region (see Figure-7.5). Normally, each detected region is matched to one in the database and then through an affine transformation placed where the detected region was using the same scale and rotation. However, it may be desirable to change the scale of this region using the blending radius to produce different effects. When scaling this region down, a user confines drawing to only small regions being painted, often produces styles associated with Abstract Expressionism.

For image-based targets, the aforementioned parameters effect the frame-to-frame compositing, meaning the same image is rendered over itself. For video-based targets, however, only a single iteration is used for each frame, as much of the information required for building styles requiring more detailed composites can be extracted over

the first 1 or 2 frames. We demonstrate how these parameters can influence a wide range of stylizations in the next section.

7.5 Results

We use the presented framework to produce artistic stylizations of photo-realistic images and videos. In this section, we show our results in image-based stylization using a landscape, abstract, and painterly scene. We then show how the same framework can be used with video targets, including an abstract and portrait video. As well, we show a particular case where the source material is aggregated from a live-stream of the target, i.e. the source and target are the same, a process we call "Memory Mosaicing".

7.5.1 Image: Landscape

In Figure-7.8, we synthesize a landscape photo of cows grazing using Expressionist painter Paul Klee. We turn off spatial blending and use a small value for the minimum region size. We also allow the maximum region size to be very large. This results in a relatively smaller region being matched to the sky and stretched to fill the top-half of the image. The synthesized region happens to look like a rainbow, though the original region itself was very abstract (see the first image in the second row of the Klee corpus).

7.5.2 Image: Abstract

In Figure-7.9, we synthesize a close-up picture of a blanket using Klimt's The Kiss. The target this time is very abstract and we will not need to synthesize parameters that force an abstract quality rendering such as large region sizes. As such, we allow the minimum region size to be very small producing more details, though retaining a style associated with Abstract Expressionism.

7.5.3 Image: Painterly

We demonstrate how CBVS can stylize existing painterly images into other styles. In the teaser graphic in Figure-7.1, we use three paintings by Van Gogh to stylize Klimt's The Kiss. Here, we set the minimum region size to be small, allowing finer details and

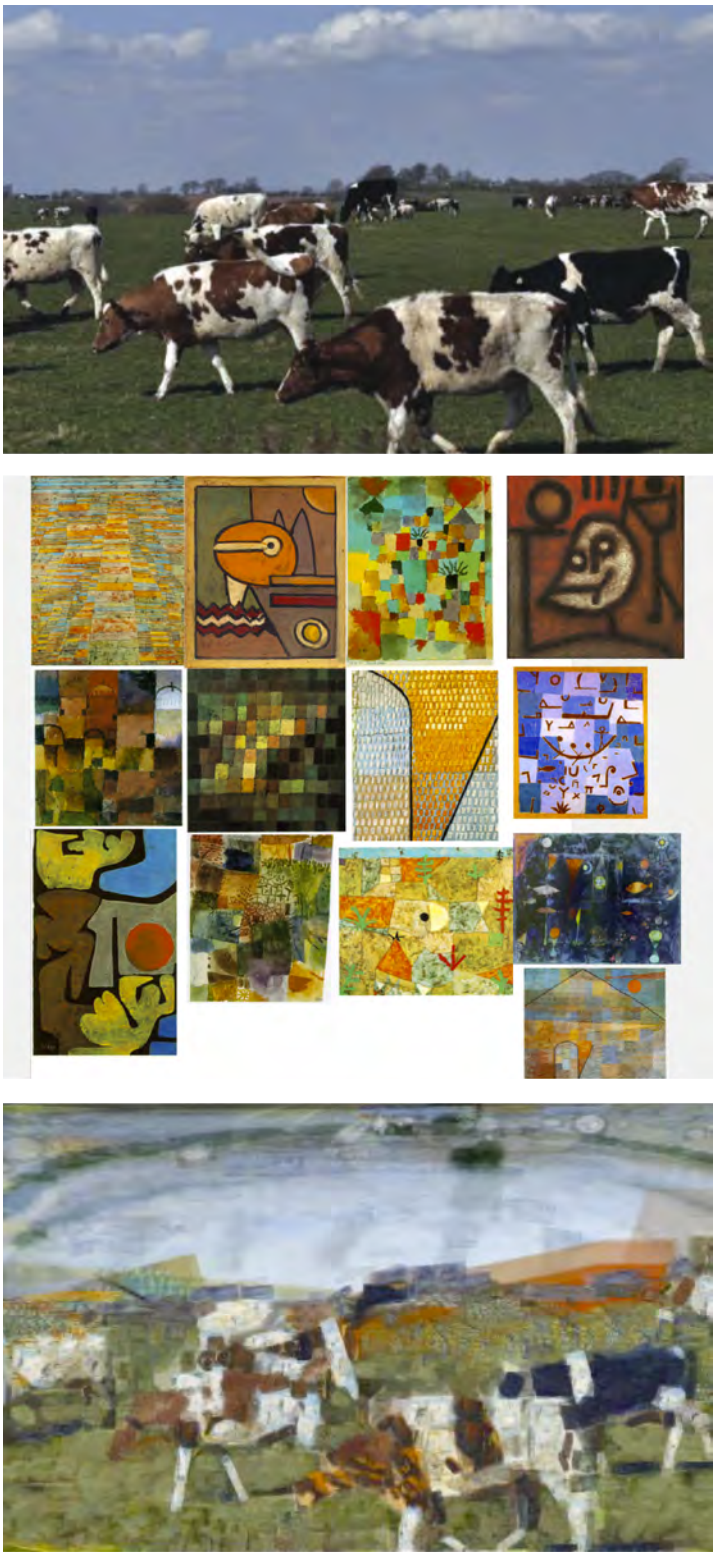


Figure 7.8: A landscape picture of cows grazing is synthesized using 13 images of Expressionism painter Paul Klee to produce the image on the bottom. Images representing faithful reproductions of Paul Klee sourced from Mark Harden's Artchive are public domain. Photo of cows taken by the author.

Figure 7.9: A close-up picture of a blanket is synthesized using Klimt's The Kiss to produce the image on the right. Best viewed in the color manuscript at 200%. Images representing faithful reproductions of Gustav Klimt sourced from Wikimedia Commons are public domain. Photorealistic scene of blanket taken by the author.

smaller brush strokes, and allow the timesteps to be high as we want to bring out as much contrast as possible.

In Figure-7.10, we try synthesizing Van Gogh's The Bedroom using 3 images of Monet's Water Lilies series. Here, we ensure we detect many small regions by increasing the timesteps and setting the minimum region size to be very small. Further, we turn on spatial blending as we decrease the minimum region size, as we want to avoid rendering any strong edges, retaining an Impressionist quality.

7.5.4 Video: Portrait

Two examples in video-based stylization are presented: one of a subject rowing a boat and another of abstract imagery. In Figure-7.11, we can see 4 frames taken from a video stylization. We use the same corpus as in Figure-7.8 and allow the minimum region size to be very small, resulting in a more Expressionist style. The first frame is not as composed as the later frames, as there will have only been 1 frame of compositing. As a result, the first frame in video-based Expressionist stylization may not be a consistent style with its later frames.

7.5.5 Video: Abstract

In Figure-7.12, we stylize a video using the same corpus as in Figure-7.8 and set the minimum region size to be very large. Thus, instead of producing an Expressionist style as in Figure-7.11, less details are synthesized resulting in a more abstract style. The first frame in this video does not necessarily require more than 1 iteration as it is synthesizing very large regions that often also overlap.

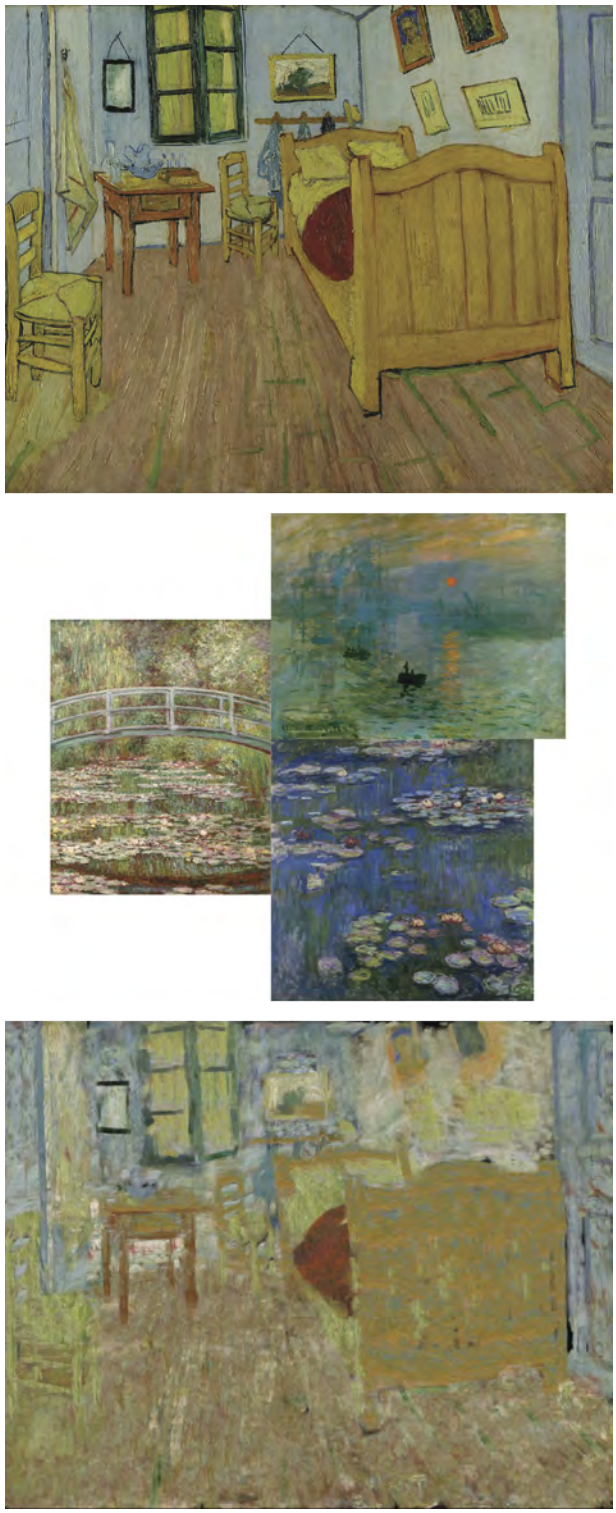


Figure 7.10: Van Gogh's "The Bedroom" is synthesized using 3 images of Monet paintings to produce the image on the bottom. Images representing faithful reproductions of Van Gogh and Claude Monet sourced from Wikimedia Commons are public domain.

Figure 7.11: Left: 4 frames from a target video; Right: Stylization using Paul Klee's corpus in Figure-7.8. We aim to synthesize with greater expression and less abstraction, and allow the minimum region size to be very small. Best viewed in the color manuscript at 200% or in the video online. Photos by the author.

Figure 7.12: Left: 4 frames from a target video; Right: Stylization using Paul Klee's corpus in Figure-7.8. Here we aim to stylize with greater abstraction than in Figure-7.11, and set the minimum region size to be fairly large. Best viewed in the color manuscript at 200% or in the video online. Photos by the author.

7.5.6 Video: The Simpsons vs. Family Guy

70k views and Vimeo Staff Pick, and selected for exhibition twice, once at the Tin Shed Gallery in London, UK, and again at the Media Art Histories conference during the ART+COMMUNICATION: SAVE AS Exhibition in Riga, Latvia.

7.6 Extensions

7.6.1 Memory Mosaicing

The artistic stylization process can be used in a real-time context without an explicit corpus. In this case, we aggregate representations learned from the ongoing stream of target frames. Parameters are generally set by the user interacting with the process, or contained to a single preset. In particular, restricting the total number of representations as first-in-first-out queue allows the process to continue in real-time with a simple linear search index. In the examples shown in Figure 7.13, we show two example outputs from the same camera stream. In the left image, we aim for large region sizes and low timesteps, resulting in a more abstract style, reminiscent of Cubist style paintings. In the right example, we allow higher timesteps and only small region sizes, resulting in a more expressive style similar to paintings in Abstract Expressionism.

7.6.2 Photosynthesizer iOS Application

Since the release of Photosynthesizer (henceforth "the app") in August 2012, the iTunes App Store has recorded 6,640 downloads for 2012, and 1,843 downloads for 2013 up to August 30th for a total of 8483 downloads in 13 months. During this time, it also reached the iTunes Top 50 for Photo and Video Apps in 8 different countries (including US and UK) and the Top 500 in 26 different countries. A large number of reviews are also recorded on the App Store, with an average rating of 3/5 for both the US (of 32 reviews) and UK (of 15 reviews). As there are nearly 100 in 26 different countries, only a few of the English ones representing different views are reproduced here:

> Like the way you can watch art materialize before you're eyes. First image important. Colors are fun to try and predict where the images go.

> Can't remove pics when added to pallet. Other than that. Awesome

> Yeah pretty cool... Be nice to have a bit of control over the process though...worth every penny.(a few days later)...i love this app and the amazing pic it creates. it does tend to crash off and on .. I thinkitsgreat [*sic*] though but i also se lots more potential... (a few days later) i must say am i getting a little tired of what i can only describe as the "digital Poonsification" that is the resultant image (google Larry Poons art of the 60's to see what i'm talking about) . It would be cool if you could get say a more planar/cubist kind of thing revisualization... As opposed to tinier and tinier "eyelike ellipses" that are always generated . M thinking of a slighlty [*sic*] more collage like result. Say like the work of Schwitters. Pretty cool app still.

> Instructions say 'pick a few pictures'. Four pictures: 'insufficient memory'. Three pictures: 'insufficient memory'. Two: 'insufficient memory'.... (I don't think you can 'synthesise' [*sic*] fewer than two things??). Guess I'll wait for the update (it's coming, right?)

> ... [*sic*] Reading the description for this app I thought WOW, but after the initial excitement ..... [*sic*] I suggest, the programmer(s) change the drugs they are being subscribed for serious delusions. They don't work, neither does the program, unless you call an indistinguishable mess being produced from your photos a synthesis, which I don't. I am going to throw some leftover food onto a canvas .... [*sic*] It is more exiting..... [*sic*] The star I had to give so someone may read this lines.

> Doesn't work at all if you deny location access even once !

> I've spent a surprising amount of time using this - its a great idea. It would be good to be able to delete images from the database as its really nice when you just use one image to make another, but this means quitting the app from the multitasking tray. Also, there seems to be some issues with the rendering on 3rd generation iPads.

> Nice idea but very low res output - deleted!

> Just did a couple of images and I like it very much. It has possibilities! I do backgrounds and videos for different musical performers and this will be very helpful. Will let you know more later?

> Picasso in an app. Smart download.

> The latest updates have really improved final results of your selected image. Only drawback is how long it takes to complete the image synthesis. Of course, you can always capture the image at less than 100%, amd [*sic*] you will still have a fine image for your collection. Biggest portion of the wait time seems to be with the finer details. I'll definitely be using this app more often now.

7.7 Discussion and Future Works

We have presented a framework for producing artistic stylizations of images or videos. A corpus of image material is automatically segmented, defining the possible strokes effecting the possible colors and textures in the stylization. Using a simple set of parameters, we have shown that many stylizations of a target image or video are possible, ranging from Impressionism, Expressionism, and Abstract Expressionism. By allowing

the interactive refinement of an image's stylization, we allow the user to experiment with a range of stylizations through simple parameters. This interactive refinement affords compositing, the ability to blend together stylizations from different parameters over time. We also demonstrate the extension of this framework to video-based stylization using simple motion tracking. As in image-based stylization, the user can influence the stylization through the same set of parameters in real-time to interactively refine the stylization.

The extension of video-based stylization is also particularly suited for real-time contexts as shown in "Memory Mosaicing", where a database is aggregated from learning representations in a target frame over time.

A number of issues could be addressed in future versions. For instance, synthesized regions with poor shape matches can be heavily distorted in a resulting synthesis. In these cases, it is likely that the database did not include any other matches with more similar shapes, or the shape descriptor had been weighted too low. As well, the speed of the synthesis in a real-time context can be greatly improved with other search methods such as tree or hash-table based indexes. As well, our approach to addressing the temporal coherence of the resulting stylization may be improved with investigating incorporating more recent models of optical flow, keyframe detection, and possibly spatiotemporal detection of representations rather than purely spatial ones.

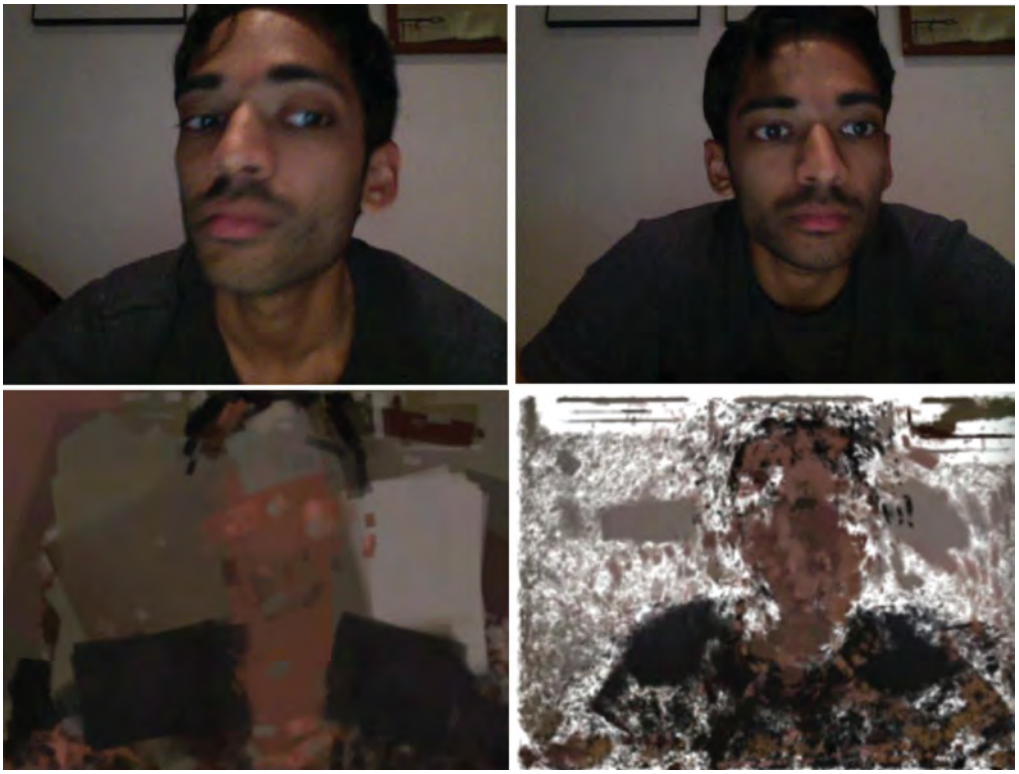


Figure 7.13: 2 examples of "Memory Mosaicing" showing the input (top) and resulting real-time stylization (bottom). Photos by the author.

CHAPTER 8

Computational Audiovisual Scene Synthesis

Contents

8.1 Introduction

Talk about attempts to combine models in perception and practice. DIEM model, how do you combine two modalities.

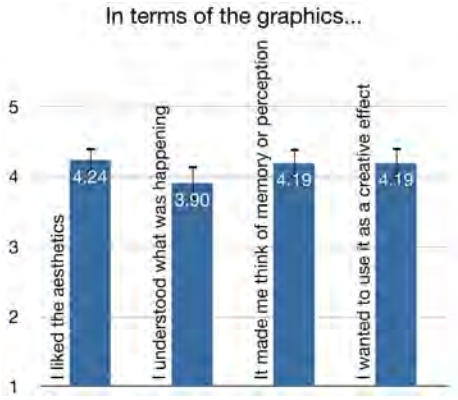
Classic effects in crossmodal/multisensory perception

Unresolved scientific questions... hence parallel...

8.2 Augmented Reality Hallucination

An interesting case of "Memory Mosaicing" is when a participant can actively explore a visual scene. By using augmented reality goggles, we allowed participants to explore their environment through the our proposed stylization process during an exhibition called "Augmented Reality Hallucinations" held at the Victoria and Albert Museum in London. Participants were invited to wear the goggles where two small CRT screens presented the same output of a "Memory Mosaicing" of a single camera mounted on the

(a) A few participants of the installation wearing the AR goggles. Photos by the author. Participants gave written consent to be photographed.



(b) Feedback where 21 participants were asked to rate different aspects of the visual synthesis. Error bars depict +/- 1 S.E.

Figure 8.1: "Augmented Reality Hallucinations", exhibited at the Victoria and Albert Museum in London, had participants wear Augmented Reality (AR) goggles with software running a real-time version of "Memory Mosaicing".

goggles right eye that faced the scene in front of them (see Figure 8.1a). As the only user interaction was in exploring a scene, a single preset was defined based on large region sizes and low number of timesteps.

Participants were also invited to give quantitative and qualitative feedback on their experience. The summary of the quantitative feedback is shown in Figure 8.1b. On the feedback form, 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 the qualitative feedback was 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." The lack of understanding of the process may also be revealed in the quantitative analysis in the second bar of the graph. However, on average, this number is still quite high across participants, though there is also no baseline to compare to.

8.2.1 Hardware

a

8.2.1.1 Vuzix Wrap20AR

a

8.2.1.2 Oculus Rift

a

8.3 YouTube Smash Up

a

8.3.1 YouTube Content ID

a In October 2007, YouTube introduced a content management tool, Video ID, to help content owners find infringing material. Block, track, monetize.

Margaret Gould Stewart, YouTube's head of content revealed in a TED talk (Stewart 2010)

> Here we can see the reference file being compared to the user generated content. The system compares every moment of one to the other to see if there's a match. This means we can identify a match even if the copy uses just a portion of the original file, plays it in slow motion, and has degraded audio or video. The scale and speed of this system is truly breathtaking – we're not just talking about a few videos, we're talking about over 100 years of video every day between new uploads and the legacy scans we regularly do across all of the content on the site. And when we compare those 100 years of video, we're comparing it against millions of reference files in our database. It'd be like 36,000 people staring at 36,000 monitors each and every day without as much as a coffee break.

Smitelli's reverse engineering of YouTube's Content ID system revealed a number of parameters which could and could not be detected by their system (Smitelli 2009).

Though one month after the 2012 fall democratic convention speech featuring first lady Michelle Obama was aired and subsequently taken down due to copyright infringement (Singel 2012), YouTube released a statement saying they would enforce manual reviewing of more of their copyright claims by the content owners in an effort to reduce erroneous claims (Alfishawi 2012).

Chapter 9

Conclusion

Contents

9.1 Summary

The world in front of us is measured by the various sensory mechanisms we have available to us. These mechanisms enable us to convert physical phenomena into electrical signals that we use to construct a perception of the world. However, we are often taught that the world we perceive is exactly as it is in front of us. Philosopher Alan Watts describes the situation:

> Most of us are brought up to feel that what we see out in front of us is something that lies beyond our eyes, out there. That the colors and the shapes that you see in this room are out there. In fact, that is not so. In fact, all that you see is a state of affairs inside your head. All these colors, all these lights, are conditions of the optical nervous system. There are, outside the eyes, quanta, electronic phenomena, vibrations, but these are not light, they are not colors until they are translated into states of the human nervous system. So if you want to know how the inside of your head feels, open your eyes and look. That is how the inside of your head feels
>
> Alan Watts

Though perhaps a bit lyrical, Watts describes the feeling inside of our heads as the one that we often mistakenly describe as what is "out there". However, this feeling is a construct. A representational theory of perception suggests that perception can be understood as process that encodes physical stimuli "out there" entering our eyes into a set of representations. These representations eventually get decoded through the cortical processes, resulting in our perceptual experience. What are the representations supporting these processes? How are they modeled, what do they look like, what can they explain, and what can they not explain? It is the aim of this thesis to develop a better understanding of questions such as these through a collage-based computational arts practice defined by a process called scene synthesis. Within this process, we aim to develop the encoding and decoding of a scene using some theoretical evidence of how we as humans may accomplish this incredible feat.

The work presented here details a computational arts practice exploring a specific type of collage called scene synthesis.

motivations, each chapter, what i did, and how i used them in practice,

9.2 Contribution

...Err...?

9.3 Limitations

Domain specific. No cross-modal/multi-sensory. Could have explored other data types, text, motion, etc...

9.4 Future Work

Shocking the mind, adding noise to the stored representations, exploring really massive data sets, better model selection, sparse coding, stochastic sampling.

Similar to the browser developed for the Daphne Oram archive, it would be interesting to create a similar tool for visual representations. This would allow one to organize

the objects within an archive based on their similarity in shape/color metric space, i.e. in terms of how similar they are represented in the brain. This tool could be used to further interrogate the representation. Further, the mapping of the proto-objects into a lower dimensional space for visualization purposes would likely reveal interesting patterns.

Ideally, the representations here of a simple shape and color vector could be extended. These could be incorporated to include co-occurrences, or how proto-objects are likely to occur with other proto-objects. This would likely be a histogram based vector, and would require some pre-training in order to build a vocabulary. Such a measure would likely bring the representation closer to a semantic object-ness and afford very interesting applications in collage-based applications using similar objects rather than proto-objects.

9.5 Final Discussion

Thank fuck it's over!

Appendix A

Appendix

Bibliography

[Ahlberg 1995] Christopher Ahlberg and Erik Wistrand. *IVEE: An Information Visualization & Exploration Environment Exploration Environment.* Proceedings of IEEE Viz'95, 1995. (Cited on page 47.)

[Alain 2001] C Alain, SR Arnott and TW Picton. *Bottomâup and topâdown influences on auditory scene analysis: Evidence from event-related brain potentials.* Journal of Experimental Psychology: Human Perception and Performance, vol. 27, no. 5, pages 1072–1089, 2001. (Cited on page 16.)

[Alain 2002] Claude Alain, Benjamin M. Schuler and Kelly L. McDonald. *Neural activity associated with distinguishing concurrent auditory objects.* The Journal of the Acoustical Society of America, vol. 111, no. 2, page 990, 2002. (Cited on page 16.)

[Alfishawi 2012] Thabet Alfishawi. *Improving Content ID*, 2012. (Cited on page 116.)

[Allamanche 2001] Eric Allamanche, J Herre and Oliver Hellmuth. *Content-based identification of audio material using MPEG-7 low level description.* Proceedings of the International Symposium on Music Information Retrieval, 2001. (Cited on page 27.)

[Aucouturier 2007] Jean-Julien Aucouturier, Boris Defreville and François Pachet. *The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music.* Journal of the Acoustical Society of America, vol. 122, no. 2, pages 881–91, 2007. (Cited on page 27.)

[Augoyard 2006] Jean-Francois Augoyard and Henry Torgue. Sonic Experience: A Guide To Everyday Sounds. Mcgill Queens Univ Pr, 2006. (Cited on page 8.)

[Azzopardi 2012] George Azzopardi and Nicolai Petkov. *A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model.* Biological cybernetics, vol. 106, no. 3, pages 177–89, March 2012. (Cited on page 69.)

[Bartsch 2001] Mark A. Bartsch and Gregory H Wakefield. *To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing.* In IEEE Workshop

on Applications of Signal Processing to Audio and Acoustics, 2001. (Cited on page 47.)

[Benetos 2011] Emmanouil Benetos and Simon Dixon. *Multiple-instrument polyphonic music transcription using a convolutive probabilistic model.* 8th Sound and Music Computing Conference, 2011. (Cited on page 28.)

[Bertin-Mahieux 2011] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman and Paul Lamere. *The million song dataset.* In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011. (Cited on page 42.)

[Biederman 1974] I Biederman, J C Rabinowitz, A L Glass and E W Stacy. *On the information extracted from a glance at a scene.* Journal of Experimental Psychology, vol. 103, no. 3, pages 597–600, 1974. (Cited on page 75.)

[Biederman 1987] I Biederman. *Recognition-by-components: a theory of human image understanding.* Psychological Review, vol. 94, no. 2, pages 115–147, 1987. (Cited on pages 75 and 79.)

[Bousseau 2007] Adrien Bousseau, Fabrice Neyret, Joëlle Thollot and David Salesin. *Video watercolorization using bidirectional texture advection.* ACM SIGGRAPH 2007 papers on - SIGGRAPH '07, page 104, 2007. (Cited on page 91.)

[Bregman 1990] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* May 1990. (Cited on pages 10 and 56.)

[Breton 1924] André Breton. *Manifeste du surréalisme.* 1924. (Cited on page 7.)

[Breton 1948] Andre Breton. *"Le Cadavre Exquis: Son Exaltation".* In Exhibition Catalog, pages 5–7, 9–11. La Dragonne, Galerie Nina Dausset, Paris, 1948. (Cited on page 7.)

[Buswell 1935] GT Buswell. *How people look at pictures.* 1935. (Cited on page 70.)

[Campbell 2007] Tom Campbell, István Winkler and Teija Kujala. *N1 and the mismatch negativity are spatiotemporally distinct ERP components: disruption of immediate memory by auditory distraction can be related to N1.* Psychophysiology, vol. 44, no. 4, pages 530–40, July 2007. (Cited on page 15.)

[Carmi 2006] Ran Carmi and Laurent Itti. *Visual causes versus correlates of attentional selection in dynamic scenes.* Vision research, vol. 46, no. 26, pages 4333–45, December 2006. (Cited on page 73.)

[Cartwright 2011] Mark Cartwright, Zafar Rafii, Jinyu Han and Bryan Pardo. *Making searchable melodies: Human vs. machine.* In Proceedings of the 2011 AAAI Workshop on Human Computation, San Francisco, USA., 2011. (Cited on page 46.)

[Casey 2001] Michael Casey. *General sound classification and similarity in MPEG-7.* Organised Sound, vol. 6, no. 02, pages 153–164, 2001. (Cited on pages 27 and 41.)

[Casey 2007] Michael Casey and Mick Grierson. *Soundspotter/Remix-TV: fast approximate matching for audio and video performance.* Proceedings of the International Computer Music Conference, 2007. (Cited on page 56.)

[Casey 2008a] MA Casey, Remco Veltkamp and Masataka Goto. *Content-based music information retrieval: current directions and future challenges.* Proceedings of the IEEE, vol. 96, no. 4, 2008. (Cited on page 42.)

[Casey 2008b] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes and Malcolm Slaney. *Content-Based Music Information Retrieval : Current Directions and Future Challenges.* Proceedings of the IEEE, vol. 96, no. 4, 2008. (Cited on page 46.)

[Celsis 1999] P Celsis, K Boulanouar, B Doyon, J P Ranjeva, I Berry, J L Nespoulous and F Chollet. *Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones.* NeuroImage, vol. 9, no. 1, pages 135–44, January 1999. (Cited on page 15.)

[Chang 2010] IC Chang, YM Peng, YS Chen and SC Wang. *Artistic Painting Style Transformation Using a Patch-based Sampling Method.* Journal of Information Science and Engineering, vol. 26, pages 1443–1458, 2010. (Cited on page 92.)

[Christel 1998] Michael Christel and David Martin. *Information visualization within a digital video library.* Journal of Intelligent Information Systems, no. June, 1998. (Cited on pages 47 and 55.)

[Chu 2009] Selina Chu, Shrikanth Narayanan and C.C.J. Kuo. *Environmental sound recognition with timeâfrequency audio features.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 6, pages 1142–1158, 2009. (Cited on page 27.)

[Collins 2004] Nick Collins. *On onsets on-the-fly: Real-time event segmentation and categorisation as a compositional effect.* In Perception. Citeseer, 2004. (Cited on page 56.)

[Copeland 2002] Roger Copeland. *Merce Cunningham and the aesthetic of collage.* TDR/The Drama Review, vol. 46, no. 1, pages 11–28, 2002. (Cited on page 9.)

[Cowan 1988] Nelson Cowan. *Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system.* Psychological bulletin, vol. 104, no. 2, pages 163–191, 1988. (Cited on pages 15 and 21.)

[Curcio 1990] CA Curcio and KA Allen. *Topography of ganglion cells in human retina.* Journal of Comparative Neurology, vol. 300, pages 5–25, 1990. (Cited on pages 66 and 68.)

[Davis 1939] P. A. Davis. *Effects of Acoustic Stimuli on the Waking Human Brain.* Journal of Neurophysiology, vol. 2, November 1939. (Cited on page 13.)

[Davis 1980] S Davis and Paul Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, no. 4, pages 357–366, 1980. (Cited on page 27.)

[DeCarlo 2002] Doug DeCarlo and Anthony Santella. *Stylization and abstraction of photographs.* ACM Transactions on Graphics, vol. 21, no. 3, pages 1–8, July 2002. (Cited on page 91.)

[Dominik 2009] L Dominik and Matthias Jarke. *Adaptive multimodal exploration of music collections.* Proceedings of the 10th International Society for Music Information Retrieval Conference, no. Ismir, pages 195–200, 2009. (Cited on pages 47 and 48.)

[Duan 2012] Zhiyao Duan, Gautham J Mysore and Paris Smaragdis. *Online PLCA for Real-time Semi-supervised.* Proceedings of the international conference on Latent Variable Analysis / Independent Component Analysis, pages 1–8, 2012. (Cited on pages 28 and 42.)

[Efros 2001] AA Efros and WT Freeman. *Image quilting for texture synthesis and transfer.* SIGGRAPH 2001: Proceedings of the 28th annual conference on Computer graphics and interactive techniques., 2001. (Cited on page 92.)

[Eronen 2006] AJ Eronen, VT Peltonen and JT Tuomi. *Audio-based context recognition.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 1, pages 321–329, 2006. (Cited on page 27.)

[Findlay 2003a] John M. Findlay and Iain D. Gilchrist. Active Vision: The Psychology of Looking and Seeing (Oxford Psychology). Oxford University Press, USA, 2003. (Cited on page 68.)

[Findlay 2003b] John M Findlay and Iain D Gilchrist. *Eye Guidance and Visual Search.* In Geoffrey Underwood, editeur, Cognitive Processes in Eye Guidance, pages 1–22. 2003. (Cited on page 72.)

[Fletcher 1940] Harvey Fletcher. *Auditory Patterns.* Reviews of Modern Physics, vol. 12, no. 1, pages 47–65, January 1940. (Cited on page 9.)

[Forssén 2007] Per-Erik Forssén. *Maximally stable colour regions for recognition and matching.* Computer Vision and Pattern Recognition 2007, (CVPR07)., 2007. (Cited on page 93.)

[Garrido 2009] Marta I Garrido, James M Kilner, Klaas E Stephan and Karl J Friston. *The mismatch negativity: a review of underlying mechanisms.* Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, vol. 120, no. 3, pages 453–63, March 2009. (Cited on pages 15 and 16.)

[Gooch 2002] Bruce Gooch, Greg Coombe and Peter Shirley. *Artistic vision: painterly rendering using computer vision techniques.* In NPAR '02 Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering, page 83, 2002. (Cited on page 91.)

[Gottlieb 1998] J P Gottlieb, M Kusunoki and M E Goldberg. *The representation of visual salience in monkey parietal cortex.* Nature, vol. 391, no. 6666, pages 481–4, January 1998. (Cited on page 69.)

[Griffiths 2004] Timothy D Griffiths and Jason D Warren. *What is an auditory object?* Nature reviews. Neuroscience, vol. 5, no. 11, pages 887–92, November 2004. (Cited on pages 9 and 10.)

[Guo 2003] G Guo and S Z Li. *Content-Based Audio Classification and Retrieval by Support Vector Machines.* IEEE Trans. Neural Networks, vol. 14, no. 1, pages 209–215, 2003. (Cited on pages 26 and 27.)

[Guo 2006] Yan-wen Guo, Jin-hui Yu, Xiao-dong Xu, Jin Wang and Qun-sheng Peng. *Example based painting generation.* Journal of Zhejiang University SCIENCE A, vol. 7, no. 7, pages 1152–1159, June 2006. (Cited on page 92.)

[Haeberli 1990] Paul Haeberli. *Paint by numbers: abstract image representations.* ACM SIGGRAPH Computer Graphics, vol. 24, no. 4, pages 207–214, September 1990. (Cited on page 90.)

[Hari 1984] R Hari, M Hämäläinen, R Ilmoniemi, E Kaukoranta, K Reinikainen, J Salminen, K Alho, R Näätänen and M Sams. *Responses of the primary auditory cortex to pitch changes in a sequence of tone pips: neuromagnetic recordings in man.* Neuroscience letters, vol. 50, no. 1-3, pages 127–32, September 1984. (Cited on page 15.)

[Harma 2005] A Harma and MF McKinney. *Automatic surveillance of the acoustic activity in our living environment.* Multimedia and Expo, 2005 IEEE International Conference on, vol. 1, no. 1, 2005. (Cited on page 27.)

[Healey 2004] Christopher G. Healey, Laura Tateosian, James T. Enns and Mark Remple. *Perceptually based brush strokes for nonphotorealistic visualization.* ACM Transactions on Graphics, vol. 23, no. 1, pages 64–96, January 2004. (Cited on page 92.)

[Heise 2012] S Heise, M Hlatky and J Loviscach. *Soundtorch: Quick browsing in large audio collections.* Audio Engineering Society Convention 125, 2012. (Cited on page 48.)

[Henderson 1999] J M Henderson and a Hollingworth. *High-level scene perception.* Annual review of psychology, vol. 50, pages 243–71, January 1999. (Cited on page 75.)

[Henderson 2003] J Henderson. *Human gaze control during real-world scene perception.* Trends in Cognitive Sciences, vol. 7, no. 11, pages 498–504, November 2003. (Cited on pages 68, 71, 72 and 73.)

[Hertzmann 1998] Aaron Hertzmann. *Painterly rendering with curved brush strokes of multiple sizes.* Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98, pages 453–460, 1998. (Cited on page 91.)

[Hertzmann 2000] Aaron Hertzmann and Ken Perlin. *Painterly rendering for video and interaction.* Proceedings of the first international symposium on Non-photorealistic animation and rendering - NPAR '00, pages 7–12, 2000. (Cited on pages 91 and 94.)

[Hertzmann 2001] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless and David H. Salesin. *Image analogies.* Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01, pages 327–340, 2001. (Cited on page 91.)

[Hertzmann 2002] Aaron Hertzmann, Nuria Oliver, Brian Curless and Steven M. Seitz. *Curve analogies.* In EGRW '02 Proceedings of the 13th Eurographics workshop on Rendering, pages 233–246, 2002. (Cited on page 92.)

[Hillyard 1973] SA Hillyard, RF Hink, VL Schwent and TW Picton. *Electrical signs of selective attention in the human brain.* Science, vol. 182, no. 4108, pages 177–180, 1973. (Cited on page 13.)

[Hillyard 1983] S a Hillyard and M Kutas. *Electrophysiology of cognitive processing.* Annual review of psychology, vol. 34, no. Haber 1974, pages 33–61, January 1983. (Cited on page 14.)

[Himmel 1998] Dave Himmel, Mark Greaves, Anne Kao and Steve Poteet. *Visualization for large collections of multimedia information.* Content Visualization and Intermedia Representations, 1998. (Cited on pages 47 and 55.)

[Hofmann 1999] Thomas Hofmann. *Probabilistic latent semantic analysis.* In Proc. of Uncertainty in Artificial Intelligence, UAIâ99, page 21. Citeseer, 1999. (Cited on pages 28, 29 and 30.)

[Hofmann 2001] Thomas Hofmann. *Unsupervised learning by probabilistic latent semantic analysis.* Machine Learning, pages 177–196, 2001. (Cited on page 29.)

[Hollingworth 2001] a Hollingworth, G Schrock and J M Henderson. *Change detection in the flicker paradigm: the role of fixation position within the scene.* Memory & cognition, vol. 29, no. 2, pages 296–304, March 2001. (Cited on page 76.)

[Horn 1981] BKP Horn and BG Schunck. *Determining optical flow.* Artificial intelligence, vol. 17, pages 185–203, 1981. (Cited on page 73.)

[Huang 2011] Hua Huang, Lei Zhang and Hong-Chao Zhang. *Arcimboldo-like collage using internet images.* Proceedings of the 2011 SIGGRAPH Asia Conference on - SA '11, vol. 30, no. 6, page 1, 2011. (Cited on page 92.)

[Hubel 1962] DH Hubel and T. N. Wiesel. *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.* The Journal of physiology, vol. 160, pages 106–154, 1962. (Cited on page 68.)

[Hubel 1968] DH Hubel and TN Wiesel. *Receptive fields and functional architecture of monkey striate cortex.* The Journal of physiology, pages 215–243, 1968. (Cited on page 68.)

[Hughes 1991] Robert Hughes. Shock of the New. 1991. (Cited on page 90.)

[Itti 1998] Laurent Itti, Christof Koch and Ernst Niebur. *A model of saliency-based visual attention for rapid scene analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. (Cited on pages 70, 71 and 72.)

[Itti 2001] L Itti and C Koch. *Computational modelling of visual attention.* Nature reviews. Neuroscience, vol. 2, no. 3, pages 194–203, March 2001. (Cited on pages 70, 71 and 72.)

[Itti 2005a] Laurent Itti. *Quantifying the contribution of lowâlevel saliency to human eye movements in dynamic scenes.* Visual Cognition, vol. 12, no. 6, pages 1093–1123, August 2005. (Cited on page 73.)

[Itti 2005b] Laurent Itti and Pierre Baldi. *A Principled Approach to Detecting Surprising Events in Video.* Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, no. June, 2005. (Cited on page 72.)

[Jaeger 2006] Timothy Jaeger. *VJ: Live Cinema Unraveled.* 2006. (Cited on page 13.)

[Joyce 1988] Paul Joyce. Hockney on Photography: Conversations with Paul Joyce. Harmony, 1988. (Cited on page 10.)

[Kahn 1994] Douglas Kahn. Wireless Imagination: Sound, Radio, and the Avant-Garde. The MIT Press, 1994. (Cited on page 8.)

[Kahneman 1984] D. Kahneman and A. Treisman. *Changing views of attention and automaticity.* In R. Parasuraman and R. Davies, editeurs, Varieties of Attention, pages 29–61. New York: Academic Press, 1984. (Cited on page 78.)

[Kalloniatis 2007] Michael Kalloniatis and Charles Luu. *Visual Acuity*, 2007. (Cited on page 81.)

[Kanoh 2001] S Kanoh, Toshinori Arai, R. Futami and N. Hoshimiya. *Properties of auditory temporal integration revealed by mismatch negativity.* Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, vol. 1, 2001. (Cited on page 15.)

[Kim 2002] Junhwan Kim and Fabio Pellacini. *Jigsaw image mosaics.* ACM Transactions on Graphics, vol. 21, no. 3, July 2002. (Cited on page 92.)

[Kim 2004] HG Kim and Nicolas Moreau. *Audio classification based on MPEG-7 spectral basis representations.* Circuits and Systems for Video, vol. 14, no. 5, pages 716–725, 2004. (Cited on page 27.)

[Kirn 2009] Peter Kirn. *How to Datamosh with Free Video Tools, âDatamoshâ is the Wrong Word, David OâReilly is Also Wrong*, 2009. (Cited on page 2.)

[Klein 2000] RM Klein. *Inhibition of return.* Trends in cognitive sciences, vol. 72, no. 1, pages 76–85, January 2000. (Cited on page 71.)

[Klier 2001] EM Klier, Hongying Wang and JD Crawford. *The superior colliculus encodes gaze commands in retinal coordinates.* Nature neuroscience, pages 627–632, 2001. (Cited on page 68.)

[Knees 2006] Peter Knees, Markus Schedl and Tim Pohle. *An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web.* Proceedings of the ACM, 2006. (Cited on pages 47 and 48.)

[Koch 1985] C Koch and S Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry.* Human Neurobiology, vol. 4, no. 4, pages 219–227, 1985. (Cited on page 70.)

[Koffka 1935] Kurt Koffka. Principles of gestalt psychology. Harcourt, Brace and Company, 1935. (Cited on page 67.)

[Köhler 1947] Wolfgang Köhler. Gestalt Psychology: An Introduction to New Concepts in Modern Psychology. Liveright, 1947. (Cited on page 67.)

[Kusunoki 2000] M Kusunoki, J Gottlieb and M E Goldberg. *The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance.* Vision research, vol. 40, no. 10-12, pages 1459–68, January 2000. (Cited on page 69.)

[Kyprianidis 2012] J Kyprianidis, John Collomosse, Tinghuai Wang and Tobias Isenberg. *State of the 'Art': A Taxonomy of Artistic Stylization Techniques for Images and Video.* IEEE transactions on Visualization and Computer Graphics, 2012. (Cited on page 91.)

[Lamy 2006] Dominique Lamy, Hannah Segal and Lital Ruderman. *Grouping does not require attention.* Perception & psychophysics, vol. 68, no. 1, pages 17–31, January 2006. (Cited on page 78.)

[Lee 2010] Hochang Lee, S Seo, S Ryoo and K Yoon. *Directional texture transfer.* NPAR '10 Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering, vol. 1, no. 212, pages 43–50, 2010. (Cited on page 92.)

[Leuba 1994] G. Leuba and R. Kraftsik. *Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age.* Anatomy and Embryology, vol. 190, no. 4, October 1994. (Cited on page 66.)

[Li 2002] Zhaoping Li. *A saliency map in primary visual cortex.* Trends in Cognitive Sciences, vol. 6, no. 1, pages 9–16, January 2002. (Cited on page 72.)

[Liang 2001] Lin Liang, Ce Liu, Ying-Qing Xu, Baining Guo and Heung-Yeung Shum. *Real-time texture synthesis by patch-based sampling.* ACM Transactions on Graphics, vol. 20, no. 3, pages 127–150, July 2001. (Cited on page 92.)

[Litwinowicz 1997] Peter Litwinowicz. *Processing images and video for an impressionist effect.* SIGGRAPH '97 Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pages 407–414, 1997. (Cited on page 91.)

[Ltd. 2013] N3krozoft Ltd. *// WALTER RUTTMANN*, 2013. (Cited on page 8.)

[Luo 2001] M. R. Luo, G. Cui and B. Rigg. *The development of the CIE 2000 colour-difference formula: CIEDE2000.* Color Research & Application, vol. 26, no. 5, pages 340–350, October 2001. (Cited on page 95.)

[Manjunath 2002] BS Manjunath, P Salembier and Thomas Sikora, editeurs. Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley and Sons, 2002. (Cited on page 27.)

[Marr 1978] D Marr and H K Nishihara. *Representation and recognition of the spatial organization of three-dimensional shapes.* Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain), vol. 200, no. 1140, pages 269–94, February 1978. (Cited on page 79.)

[Marr 1982] David Marr. Vision: A Computational investigation into the Human Representation and Processing of Visual Information. 1982. (Cited on pages 10, 79 and 90.)

[McDermott 2009] Josh H McDermott. *The cocktail party problem.* Current biology : CB, vol. 19, no. 22, pages R1024–7, December 2009. (Cited on page 9.)

[McGurk 1976] Harry McGurk and John Macdonald. *Hearing lips and seeing voices.* Nature, vol. 264, no. 5588, pages 746–748, December 1976. (Cited on page 66.)

[McKinney 2003] MF McKinney. *Features for audio and music classification.* Proc. ISMIR, vol. 4, 2003. (Cited on page 27.)

[McLeod 2011] Kembrew McLeod. Cutting Across Media: Appropriation Art, Interventionist Collage, and Copyright Law. Duke University Press Books, 2011. (Cited on pages 6, 10 and 0.)

[Meier 1996] Barbara J. Meier. *Painterly rendering for animation.* Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96, pages 477–484, 1996. (Cited on page 91.)

[Melucci 2000] Massimo Melucci and Nicola Orio. *SMILE: A system for content-based musical information retrieval environments.* RIAO'2000 Conference proceedings, 2000. (Cited on page 46.)

[Merrill 1968] RG Merrill and DR Metcalf. *COGNITIVE STYLES OF VISUAL PERCEPTION IN THE EVALUATION OF TELEVISION SYSTEMS.* Perceptual and Motor Skills, pages 1043–1046, 1968. (Cited on page 70.)

[Mesaros 2010] Annamaria Mesaros, Toni Heittola, Antti Eronen and Tuomas Virtanen. *Acoustic event detection in real-life recordings.* In 18th European Signal Processing Conference, 2010. (Cited on page 27.)

[Miller 2012] Jordan Miller and David Mould. *Accurate and Discernible Photocollages.* Computational Aesthetics in Graphics, Visualization, and Imaging, pages 115–124, 2012. (Cited on page 92.)

[Mital 2011] Parag K. Mital, Tim J. Smith, Robin L. Hill and John M. Henderson. *Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion.* Cognitive Computation, October 2011. (Cited on page 73.)

[Mital 2012] Parag Kumar Mital and Mick Grierson. *Audio Content-based Information Display: Mining Unknown Electronic Music Databases through Interactive Visualization of Latent Component Relationships.* In International Symposium on Music Information Retrieval 2012 (In Review), 2012. (Cited on page 29.)

[Moerel 2013] M M L Moerel. *Encoding of natural sounds in the human brain.* PhD thesis, 2013. (Cited on page 22.)

[Moore 2003] Tirin Moore and KM Armstrong. *Selective gating of visual signals by microstimulation of frontal cortex.* Nature, vol. 421, no. January, pages 370–373, 2003. (Cited on page 69.)

[Näätänen 1978] R Näätänen, AWK Gaillard and S Mäntysalo. *Early selective-attention effect on evoked potential reinterpreted.* Acta psychologica, vol. 42, pages 313–329, 1978. (Cited on pages 13, 14 and 15.)

[Näätänen 1987] R Näätänen and T Picton. *The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure.* Psychophysiology, vol. 24, no. 4, pages 375–425, 1987. (Cited on page 15.)

[Näätänen 2001] R Näätänen, M Tervaniemi, E Sussman, P Paavilainen and I Winkler. *"Primitive intelligence" in the auditory cortex.* Trends in neurosciences, vol. 24, no. 5, pages 283–8, May 2001. (Cited on page 16.)

[Näätänen 2007a] R Näätänen, P Paavilainen, T Rinne and K Alho. *The mismatch negativity (MMN) in basic research of central auditory processing: a review.* Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, vol. 118, no. 12, pages 2544–90, December 2007. (Cited on page 15.)

[Näätänen 2007b] R Näätänen, P Paavilainen, T Rinne and K Alho. *The mismatch negativity (MMN) in basic research of central auditory processing: a review.* Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, vol. 118, no. 12, pages 2544–90, December 2007. (Cited on page 16.)

[Näätänen 2011] Risto Näätänen, Teija Kujala and István Winkler. *Auditory processing that leads to conscious perception: a unique window to central auditory processing opened by the mismatch negativity and related responses.* Psychophysiology, vol. 48, no. 1, pages 4–22, January 2011. (Cited on pages 14 and 15.)

[Nam 2012] Juhan Nam, Gautham Mysore and Paris Smaragdis. *Sound Recognition in Mixtures.* Latent Variable Analysis and Signal, Lecture Notes in Computer Science, vol. 7191, pages 405–413, 2012. (Cited on page 28.)

[O'Donovan 2012] Peter O'Donovan and Aaron Hertzmann. *AniPaint: interactive painterly animation from video.* IEEE transactions on visualization and computer graphics, vol. 18, no. 3, pages 475–87, March 2012. (Cited on page 91.)

[Oliva 1997] Aude Oliva and Philippe G Schyns. *Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli.* Cognitive psychology, vol. 107, pages 72–107, 1997. (Cited on pages 75 and 81.)

[Oliva 2001] Aude Oliva and Antonio Torralba. *Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope â*. International Journal, vol. 42, no. 3, pages 145–175, 2001. (Cited on page 75.)

[Oliva 2005] Aude Oliva. *Gist of the scene*. In Neurobiology of attention, pages 251–257. 2005. (Cited on page 75.)

[Orabona 2007] Francesco Orabona and Giorgio Metta. *A proto-object based visual attention model*. Attention in cognitive systems. Theories, pages 198–215, 2007. (Cited on page 83.)

[Orchard 2008] Jeff Orchard and CS Kaplan. *Cut-out image mosaics*. ACM Transactions on Graphics, vol. 1, no. 212, 2008. (Cited on page 92.)

[Oswald 1985] John Oswald. *Plunderphonics, or Audio Piracy as a Compositional Prerogative*. In Wired Society Electro-Acoustic Conference, 1985. (Cited on page 15.)

[Oswald 2013a] John Oswald. *Plunderphonics - Essay*, 2013. (Cited on page 15.)

[Oswald 2013b] John Oswald. *Plunderphonics: Interviews*, 2013. (Cited on page 15.)

[Palmer 1999] Stephen E. Palmer. Vision science: Photons to phenomenology. 1999. (Cited on page 69.)

[Pampalk 2006] Elias Pampalk. *Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns*. International Symposium on Music Information Retrieval, 2006. (Cited on page 27.)

[Piazza 2013] Elise a Piazza, Timothy D Sweeny, David Wessel, Michael a Silver and David Whitney. *Humans Use Summary Statistics to Perceive Auditory Sequences*. Psychological science, June 2013. (Cited on page 22.)

[Posner 1984] MI Posner and Y Cohen. *Components of visual orienting*. In Attention and Performance X: Control of Language Processes., pages 551–556. 1984. (Cited on pages 71 and 72.)

[Posner 1985] Michael I. Posner, Robert D. Rafal, Lisa S. Choate and Jonathan Vaughan. *Inhibition of return: Neural basis and function*. Cognitive Neuropsychology, vol. 2, no. 3, pages 211–228, August 1985. (Cited on pages 71 and 72.)

[Potter 1969] Mary C Potter and Ellen I Levy. *Recognition memory for a rapid sequence of pictures.* Journal of Experimental Psychology, vol. 81, no. 1, pages 10–15, 1969. (Cited on page 75.)

[Potter 1976] M C Potter. *Short-term conceptual memory for pictures.* Journal of experimental psychology Human learning and memory, vol. 2, no. 5, pages 509–522, 1976. (Cited on page 75.)

[Pylyshyn 2001] Zenon W Pylyshyn. *Visual indexes, preconceptual objects, and situated vision.* Cognition, vol. 80, pages 127–158, 2001. (Cited on page 82.)

[Raj 2010] Bhiksha Raj, Tuomas Virtanen, Sourish Chaudhuri and Rita Singh. *Non-negative matrix factorization based compensation of music for automatic speech recognition.* Proceedings of the 11th Annual Conference of the International Speech Communication Association, pages 717–720, 2010. (Cited on page 27.)

[Rao 1999] R P Rao and D H Ballard. *Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.* Nature neuroscience, vol. 2, no. 1, pages 79–87, January 1999. (Cited on page 69.)

[Rauss 2011] Karsten Rauss, Sophie Schwartz and Gilles Pourtois. *Top-down effects on early visual processing in humans: a predictive coding framework.* Neuroscience and biobehavioral reviews, vol. 35, no. 5, pages 1237–53, April 2011. (Cited on page 71.)

[Rensink 2000] Ronald a. Rensink. *The Dynamic Representation of Scenes.* Visual Cognition, vol. 7, no. 1-3, pages 17–42, January 2000. (Cited on pages 76, 78 and 82.)

[Rensink 2001] RA Rensink. *Change blindness: Implications for the nature of visual attention.* In Vision & Attention, pages 169–188. 2001. (Cited on pages 76, 78 and 82.)

[Rensink 2002] RA Rensink. *Change detection.* Annual review of psychology, vol. 53, pages 245–77, January 2002. (Cited on pages 78 and 81.)

[Rhodes 2010] Christophe Rhodes, Tim Crawford and Michael Casey. *Investigating music collections at different scales with AudioDB.* Journal of New Music, pages 1–19, 2010. (Cited on pages 42 and 46.)

[Rinne 1999] T Rinne, G Gratton, M Fabiani, N Cowan, E Maclin, a Stinard, J Sinkkonen, K Alho and R Näätänen. *Scalp-recorded optical signals make sound processing in the auditory cortex visible?* NeuroImage, vol. 10, no. 5, pages 620–4, November 1999. (Cited on page 15.)

[Rosenholtz 1999] Ruth Rosenholtz. *Rapid communication A simple saliency model predicts a number of motion popout phenomena.* Vision Research, vol. 39, pages 3157–3163, 1999. (Cited on page 72.)

[Rubin 1915] Edgar 1886-1951 Rubin. *Synsoplevede figurer: studier i psykologisk analyse.* PhD thesis, København og Kristiania, Gyldendal, 1915. (Cited on page 66.)

[Sams 1983] M Sams, K Alho and R Näätänen. *Sequential effects on the ERP in discriminating two stimuli.* Biological psychology, vol. 17, no. 1, pages 41–58, August 1983. (Cited on page 15.)

[Samson 2010] Fabienne Samson, Thomas a Zeffiro, Alain Toussaint and Pascal Belin. *Stimulus complexity and categorical effects in human auditory cortex: an activation likelihood estimation meta-analysis.* Frontiers in psychology, vol. 1, no. January, page 241, January 2010. (Cited on page 22.)

[Schwarz 2006] D Schwarz. *Concatenative Sound Synthesis: The Early Years.* J. New Music Research, vol. 35, no. 1, 2006. (Cited on page 44.)

[Schwarz 2008] Diemo Schwarz, Roland Cahen and Sam Britton. *Principles and applications of interactive corpus-based concatenative synthesis.* Journées d'Informatique Musicale (JIM), GMEA, Albi, France, 2008. (Cited on pages 47 and 48.)

[Schyns 1994] Philippe G Schyns and Aude Oliva. *From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition.* Psychological Science, vol. 5, no. 4, pages 195–200, 1994. (Cited on pages 75 and 81.)

[Scotia 2010] Nova Scotia. *Inhibition of return in static but not necessarily in dynamic search.* vol. 72, no. 1, pages 76–85, 2010. (Cited on page 72.)

[Seo 2010] SangHyun Seo and KyungHyun Yoon. *Color juxtaposition for pointillism based on an artistic color model and a statistical analysis.* The Visual Computer: International Journal of Computer Graphics, vol. 26, no. 6-8, pages 421–431, April 2010. (Cited on page 91.)

[Shamma 2010] SA Shamma and Christophe Micheyl. *Behind the scenes of auditory perception.* Current opinion in neurobiology, vol. 20, no. 3, pages 361–366, 2010. (Cited on page 16.)

[Shamma 2011] Shihab a Shamma, Mounya Elhilali and Christophe Micheyl. *Temporal coherence and attention in auditory scene analysis.* Trends in neurosciences, vol. 34, no. 3, pages 114–23, March 2011. (Cited on pages 20 and 22.)

[Shams 2002] Ladan Shams, Yukiyasu Kamitani and Shinsuke Shimojo. *Visual illusion induced by sound.* Brain research. Cognitive brain research, vol. 14, no. 1, pages 147–52, June 2002. (Cited on page 66.)

[Sherrington 1906] Sir Charles Scott Sherrington. The Integrative action of the nervous system. Yale University Press, 1906. (Cited on page 68.)

[Simons 1997] Daniel J Simons and Daniel T Levin. *Change Blindness.* Trends in Cognitive Sciences, vol. 1, no. 7, pages 261–267, 1997. (Cited on page 81.)

[Simons 1998] Daniel J Simons and Daniel T Levin. *Failure to detect changes to people during a real-world interaction.* Psychonomic Bulletin & Review, vol. 5, no. 4, pages 644–649, 1998. (Cited on page 77.)

[Simons 1999] D J Simons and C F Chabris. *Gorillas in our midst: sustained inattentional blindness for dynamic events.* Perception, vol. 28, no. 9, pages 1059–74, January 1999. (Cited on page 76.)

[Singel 2012] Ryan Singel. *YouTube Flags Democratsâ Convention Video on Copyright Grounds*, 2012. (Cited on page 116.)

[Smaragdis 2006] Paris Smaragdis, Bhiksha Raj and Madhusudana Shashanka. *A Probabilistic Latent Variable Model for Acoustic Modeling.* In In Workshop on Advances in Models for Acoustic Processing at NIPS, numéro 1, 2006. (Cited on pages 28, 29 and 45.)

[Smaragdis 2007a] Paris Smaragdis and B. Raj. *Shift-invariant probabilistic latent component analysis.* Journal of Machine Learning Research, no. 5, 2007. (Cited on pages 28, 29 and 41.)

[Smaragdis 2007b] Paris Smaragdis, Bhiksha Raj and Madhusudana Shashanka. *Supervised and semi-supervised separation of sounds from single-channel mixtures.*

In Proceedings of the 7th international conference on Independent component analysis and signal separation, 2007. (Cited on page 28.)

[Smitelli 2009] Scott Smitelli. *Fun with YouTube's Audio Content ID System*, 2009. (Cited on page 116.)

[Smith 2009] Tim J. Smith and John M. Henderson. *Facilitation of return during scene viewing.* Visual Cognition, vol. 17, no. 6-7, pages 1083–1108, August 2009. (Cited on page 72.)

[Smith 2011] Tim Smith and Parag Kumar Mital. *Watching the world go by: Attentional prioritization of social motion during dynamic scene viewing.* In Vision Sciences Society (abstract), 2011. (Cited on pages 71 and 74.)

[Smith 2013] Tim J Smith and Parag K Mital. *Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes.* Journal of Vision, vol. 13, no. June, pages 1–25, 2013. (Cited on page 74.)

[Snyder 1976] Elaine Snyder and Steven A. Hillyard. *Long-latency evoked potentials to irrelevant, deviant stimuli.* Behavioral Biology, vol. 16, no. 3, pages 319–331, March 1976. (Cited on page 14.)

[Soltani 2010a] Alireza Soltani and Christof Koch. *Visual saliency computations: mechanisms, constraints, and the effect of feedback.* The Journal of neuroscience : the official journal of the Society for Neuroscience, vol. 30, no. 38, pages 12831–43, September 2010. (Cited on page 72.)

[Soltani 2010b] Alireza Soltani and Christof Koch. *Visual saliency computations: mechanisms, constraints, and the effect of feedback.* The Journal of neuroscience : the official journal of the Society for Neuroscience, vol. 30, no. 38, pages 12831–43, September 2010. (Cited on page 72.)

[Steenhuisen 2005] Paul Steenhuisen. Sonic Mosaics: Conversations with Composers. University of Alberta Press, 2005. (Cited on page 15.)

[Stewart 2008] R Stewart and M Levy. *3D interactive environment for music collection navigation.* Proc. DAFx-08, pages 1–5, 2008. (Cited on page 47.)

[Stewart 2010] Margaret Gould Stewart. *How YouTube thinks about copyright | Video on TED.com*, 2010. (Cited on page 115.)

[Su 2011] F Su, L Yang and Tong Lu. *Environmental sound classification for scene recognition using local discriminant bases and HMM.* Proceedings of the 19th ACM international, pages 1389–1392, 2011. (Cited on page 27.)

[Sussman 2005] Elyse S. Sussman. *Integration and segregation in auditory scene analysis.* The Journal of the Acoustical Society of America, vol. 117, no. 3, page 1285, 2005. (Cited on pages 17 and 22.)

[Sutton 1965] S Sutton, M Braren, Joseph Zubin and ER John. *Evoked-potential correlates of stimulus uncertainty.* Science, 1965. (Cited on page 13.)

[Sziklai 1956] George Sziklai. *Some studies in the speed of visual perception.* IEEE Transactions on Information Theory, vol. 2, no. 3, pages 125–128, 1956. (Cited on page 70.)

[Tatler 2009a] Benjamin Tatler. *Current understanding of eye guidance.* Visual Cognition, vol. 17, no. 6, pages 777–789, August 2009. (Cited on pages 72 and 73.)

[Tatler 2009b] Benjamin Tatler and Benjamin Vincent. *The prominence of behavioural biases in eye guidance.* Visual Cognition, vol. 17, no. 6, pages 1029–1054, August 2009. (Cited on page 74.)

[Tatler 2011] Benjamin W Tatler, Mary M Hayhoe, Michael F Land and Dana H Ballard. *Eye guidance in natural vision : Reinterpreting salience.* Journal of Vision, vol. 11, pages 1–23, 2011. (Cited on pages 71, 72 and 73.)

[Taylor 2006] Brandon Taylor. Collage: The Making of Modern Art. Thames & Hudson, 2006. (Cited on page 11.)

[Teki 2011] Sundeep Teki, Maria Chait, Sukhbinder Kumar, Katharina von Kriegstein and Timothy D Griffiths. *Brain bases for auditory stimulus-driven figure-ground segregation.* The Journal of neuroscience : the official journal of the Society for Neuroscience, vol. 31, no. 1, pages 164–71, January 2011. (Cited on page 20.)

[Teki 2013] Sundeep Teki, Maria Chait, Sukhbinder Kumar, Shihab Shamma and Timothy D Griffiths. *Segregation of complex acoustic scenes based on temporal coherence.* eLife, vol. 2, page e00699, January 2013. (Cited on page 20.)

[Temko 2007] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu and Maurizio Omologo. *CLEAR evaluation of acoustic event detection*

and classification systems. Multimodal Technologies for Perception of Humans, pages 311–322, 2007. (Cited on pages 26 and 27.)

[Tervaniemi 1994] M. Tervaniemi, J. Saarinen, P. Paavilainen, N. Danilova and R. Näätänen. *Temporal integration of auditory information in sensory memory as reflected by the mismatch negativity.* Biological Psychology, vol. 38, no. 2-3, pages 157–167, 1994. (Cited on page 15.)

[Tervaniemi 1997a] M Tervaniemi, E Schröger and R Näätänen. *Pre-attentive processing of spectrally complex sounds with asynchronous onsets: an event-related potential study with human subjects.* Neuroscience letters, vol. 227, no. 3, pages 197–200, May 1997. (Cited on page 15.)

[Tervaniemi 1997b] M Tervaniemi, I Winkler and R Näätänen. *Pre-attentive categorization of sounds by timbre as revealed by event-related potentials.* Neuroreport, vol. 8, no. 11, pages 2571–4, July 1997. (Cited on page 16.)

[Tervaniemi 2000] M Tervaniemi, S V Medvedev, K Alho, S V Pakhomov, M S Roudas, T L Van Zuijen and R Näätänen. *Lateralized automatic auditory processing of phonetic versus musical information: a PET study.* Human brain mapping, vol. 10, no. 2, pages 74–9, June 2000. (Cited on page 15.)

[Tholl 2010] Andrew Tholl. *Plunderphonics: A Literature Review.* 2010. (Cited on page 15.)

[Thompson 2005] Kirk G Thompson, Keri L Biscoe and Takashi R Sato. *Neuronal basis of covert spatial attention in the frontal eye field.* The Journal of neuroscience : the official journal of the Society for Neuroscience, vol. 25, no. 41, pages 9479–87, October 2005. (Cited on page 69.)

[Torralba 2006] Antonio Torralba, Aude Oliva, Monica S Castelhano and John M Henderson. *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.* Psychological review, vol. 113, no. 4, pages 766–86, October 2006. (Cited on page 71.)

[Treisman 1980] AM Treisman and Garry Gelade. *A feature-integration theory of attention.* Cognitive psychology, vol. 12, pages 97–136, 1980. (Cited on page 71.)

[Uhde 1994] Jan Uhde. *Jan Svankmajer: The Prodigious Animator From Prague.* Kinema: A Journal for Film and Audiovisual Media, vol. Spring, 1994. (Cited on page 9.)

[van Noorden 1975] LPAS van Noorden. *Temporal coherence in the perception of tone sequences.* PhD thesis, 1975. (Cited on page 10.)

[Walther 2006] Dirk Walther and Christof Koch. *Modeling attention to salient proto-objects.* Neural networks : the official journal of the International Neural Network Society, vol. 19, no. 9, pages 1395–407, November 2006. (Cited on page 83.)

[Wang 2004a] Bin Wang, Wenping Wang, Huaiping Yang and Jiaguang Sun. *Efficient example-based painting and synthesis of 2D directional texture.* IEEE transactions on visualization and computer graphics, vol. 10, no. 3, pages 266–77, 2004. (Cited on pages 92 and 96.)

[Wang 2004b] Jue Wang, Yingqing Xu, Heung-Yeung Shum and Michael F. Cohen. *Video tooning.* In SIGGRAPH '04 ACM SIGGRAPH 2004 Papers, pages 574–583, New York, New York, USA, 2004. ACM Press. (Cited on page 91.)

[Wang 2006] Avery Wang. *The Shazam Music Recognition Service.* Communications of the ACM, vol. 49, no. 8, 2006. (Cited on page 46.)

[Wang 2011] JC Wang, HS Lee and HM Wang. *Learning the Similarity of Audio Music in Bag-of-Frames Representation from Tagged Music Data.* International Symposium on Music Information Retrieval, 2011. (Cited on page 27.)

[Weiss 2011] Ron J. Weiss and Juan Pablo Bello. *Unsupervised Discovery of Temporal Structure in Music.* IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 6, pages 1240–1251, October 2011. (Cited on pages 28 and 41.)

[Wiener 1948] Norbert Wiener. Cybernetics; or control and communication in the animal and the machine. 1948. (Cited on page 3.)

[Winkler 2009] István Winkler, Susan L Denham and Israel Nelken. *Modeling the auditory scene: predictive regularity representations and perceptual objects.* Trends in cognitive sciences, vol. 13, no. 12, pages 532–40, December 2009. (Cited on page 11.)

[Winkler 2010] István Winkler. *In search for auditory object representations.* In Unconscious Memory Representations in Perception: Processes and mechanisms in the brain (Advances in Consciousness Research), pages 71–106. John Benjamins Publishing Company, 2010. (Cited on page 9.)

[Wolfe 1989] J M Wolfe, K R Cave and S L Franzel. *Guided search: an alternative to the feature integration model for visual search.* Journal of Experimental Psychology: Human Perception and Performance, vol. 15, no. 3, pages 419–433, 1989. (Cited on page 70.)

[Xiong 2003] Ziyou Xiong and Regunathan Radhakrishnan. *Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification.* , Speech, and Signal, 2003. (Cited on page 27.)

[Yang 2006] HL Yang and CK Yang. *A Non-Photorealistic Rendering of Seurat's Pointillism.* Advances in Visual Computing, pages 760–769, 2006. (Cited on page 91.)

[Yarbus 1967] Alfred Yarbus. Eye movements and vision. 1967. (Cited on pages 70, 71 and 73.)

[Young 2008] MW Young, JL Drever, Mick Grierson and Ian Stonehouse. *Goldsmiths Electronic Music Studios: 40 Years.* In Proceedings of the 2008 International Computer Music Conference, pages 8–11, 2008. (Cited on page 45.)

[Zeng 2009] K Zeng, M Zhao, C Xiong and SC Zhu. *From image parsing to painterly rendering.* ACM Transactions on Graphics (TOG), vol. 29, no. 1, 2009. (Cited on page 92.)

[Zhang 2012] Xilin Zhang, Li Zhaoping, Tiangang Zhou and Fang Fang. *Neural activities in v1 create a bottom-up saliency map.* Neuron, vol. 73, no. 1, pages 183–92, January 2012. (Cited on page 72.)

[Zimmer 2003] Robert Zimmer. *Abstraction in art with implications for perception.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 358, no. 1435, pages 1285–91, July 2003. (Cited on page 90.)