Goldsmiths
UNIVERSITY OF LONDON

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

THANKS...

Contents

Abstract

Abstract...

Keywords: synthesis, scene analysis, encoding, decoding

Chapter 1

Introduction

Contents

These fragments I have shored
against my ruins.
– T.S. Eliot, *The Waste Land*

1.1 Motivation

Collage is a practice that uses existing fragments of media to create new meanings. The juxtaposition of appropriated cultural fragments such as words, photos, or sound clips, the situation of the content's reuse, and the collage's overall composition produces new meanings that the original fragments alone could not have provided. Indeed, more than an artistic technique, it has also been described as a "philosophical attitude" that can be applied to virtually any medium.

As collage practice has progressed over the last 100 years, media too has evolved, meaning the tools and technologies available for manipulating and accessing media have altered the practice. This thesis looks at how developments in machine listening and machine vision can create new representations for the units of collage and automate the generation of collage in order to build interactive collage experiences that a participant can experience in real-time. In one incarnation, "Memory Mosaic", a participant experiences a real-time collage of aggregated memories, associating the ongoing audiovisual world to fragments of previous experiences. In another, the collage is experienced through an augmented reality headset, where the vision and sound of the environment is presented through a headset built for

immersive gaming environment. The resulting experience, "Augmented Reality Hallucinations", captures the experiences of the viewer through their attention to the world, translating them into a collage of fragments.

1.2 Background

1.2.1 Early Collage

The first-half of the 20th century saw an explosion of collage practices spanning visual, textual, and sonic mediums. In the early 20th-century, Pablo Picasso and Georges Braque were experimenting with gluing visual paper fragments of culture represented by stamps, newspaper clippings, and photos onto the canvas of their paintings. The practice was later extended to wood by Kurt Schwitters in the 1920's, and to purely cut-up photographic material in the 1950's, a technique known as photomontage.

During the same time, at a Dadaist rally in the 1920's, Tristan Tzara created a poem by taking cut-up fragments of text-based media such as newspapers or brochures out of a hat and verbalizing them eventually leading to a riot destroying the theatre on location and the expulsion of Tzara from the movement by Andre Breton. The technique, also known as "cut-up technique", would later form the basis of literary works such as T. S. Eliot's *The Waste Land*, James Joyce's *Ulysses*, and William Burrough's *Naked Lunch*.

Collage did not stop with static media, however. At about the same time 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 crowds 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. While not strictly collage as it had been, the notion of producing new meanings from the collection of individual fragments is certainly shared.

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 Schaffer was also experimenting with splicing and recombining magnetic tape recordings of sound in a practice later called musique concrète. Later theorized in "The Guide to the Sonic Object" ...

1.2.2 Modern Collage

Since then, media has become increasingly digitized. As a result the capabilities of editing software have afforded practitioners with faster methods of composition. For instance, Adobe Photoshop and Adobe After Effects support the automatic parsing of an image or video into object regions that can be individually manipulated enabling artists to finely segment and compose visual media.

For sound-based collage, early digital samplers such as the Fairlight CMI or 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.

Lee "Scratch" Perry. King Tubby. Public Enemy. The artist Abstract, also known as Q-Tip, of the early 1990's hip-hop group "A Tribe Called Quest" remark on the influence of the previous generation's culture within their music in his lyrics:

> Back in the days when I was a teenager
> Before I had status and before I had a pager
> You could find the Abstract listening to hip hop
> My pops used to say, it reminded him of Bebop
> I said, well daddy don't you know that things go in cycles
> Way that Bobby Brown is just amping like Michael

1.2.3 Information Visualization/Auralization

As collage also works with visualizing large amounts of existing data, the work also shares motivations with information visualization and information auralization, where the data are presented in order to tell a story. In another vein, the work also shares motivations with encoding and decoding, or compression. The purpose in compression is to take an existing dataset and reduce it down to "perceptually" similar information, while removing extraneous information.

1.2.4 Compression/Cryptography

In a related note, this process can also be used for encryption/decryption, where the same process is used to embed meaningful information that can only be deciphered through arcane processes, such as a table look-up or transformation of the data.

1.3 Goals

This thesis investigates a computational model for automating collage generation. These parameters effect the two major modular components of the algorithm: attention modeling and representation. Interaction is focused on the selection of material used in the collage, how the source content is parsed, stored, and for dynamic media, how the collage should be composited over time. This system affords entirely new experiences around collage, such as the ability to aggregate source content in real-time or experience the collage through an augmented reality headset that mosaics the worlds as it is experienced.

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.

1.4 Overview

Attention literature

Representation literature

Chapter 2

Basics

Contents

2.1 Attention

2.2 Representation

Chapter 3

Auditory Scene Analysis

Contents

The growth of digital audio archives has built the need for intelligent content-based analysis systems. Within audio archives, an acoustic class may not occur as an isolated stream but rather within a mixture of other acoustic classes. As such, content-based information retrieval algorithms should also be capable of classifying the separate acoustic classes that give rise to an acoustic scene's mixture. This paper investigates the performance of three classifiers; (1) a full-frequency and (2) reduced frequency classifier 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 features, MFCCs. We evaluate these models in a variety of cases: (1) classifying acoustic textures, (2), classifying acoustic textures in the presence of noise, and (3), classifying acoustic mixtures. 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.

3.1 Introduction

The growth of digital audio media archives has built the need for intelligent and automatic preprocessing of stored data in order for composers, sound designers, or analysts to search and retrieve items of interest. In a typical scenario, a user would

like to search an archive based on their interests in the *contents* of a file rather than the file-systems own characteristics, e.g. their name, size, or last modified date. Solutions to the former scenario are known generally as content-based information retrieval (CBIR), an active topic in all forms of multimedia archives such as text, picture, video, and sound.

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 much 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.

An efficient model should be able to classify an acoustic scene into its constituent classes, allowing for a parts-based analysis of the stored data such as which acoustic events appear within an audio clip. For example, a street scene may be better described by the parts that describe it such as "car" and "horn". An analogous model in vision is one that detects objects in a visual scene rather than the entire scene itself. However, the difficulty of classifying the parts comprising a mixture of acoustic events is well noted in the literature of acoustic event detection 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).

Our work in classifying mixtures is inspired by research in auditory perception highlighting the importance of the separation/segregation of sound information for structuring sensory input for high-level perceptual processes (Winkler 2009; Teki 2011). For instance, the predictive regularities and temporal coherence of frequency information may lend listeners a cue for discovering sources of sound information (Winkler 2009; Shamma 2011). Such evidence is reminiscent of approaches in Auditory Scene Analysis (Bregman 1990) that claim that the perceptual organization of an auditory scene is represented by a decomposition into streams. According to Bregman, each stream is encoded by one of two formations: (1) primitive, low-level characteristics such as frequency, intensity, and location; and (2), schema-driven integration of sensory evidence where schema are defined by Gestalt-like regularities such as similarities, differences, common-fate, or continuity in frequency information from a continuous signal. Attention is then thought to act upon one of these streams of information. Thus, in investigating computational models of acoustic information, we are motivated by algorithms able to capture the predictive regularities describing *subspaces* of frequency distributions.

3.2 Previous Work

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 acting on the MFCC feature vector are generally unsuited for describing the *parts* that make up an acoustic scene as such measures 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 subspaces that 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 work 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, 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 textures, (2), classifying acoustic textures in the presence of noise, and (3), classifying acoustic mixtures.

3.3 Probabilistic Latent Component Analysis

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:

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

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 the Algorithm below:

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 Methods

3.4.1 Material

Sounds were sourced from both the Sound Ideas archive and the BBC Sound Library and selected based on whether the sound file consistently represented a single sound class. 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.

3.4.2 Models

We tested 3 kinds of models, (1), a Gaussian Mixture Model of Mel-Frequency Cepstral Coefficients (MFCC Model), (2), a probabilistic Latent Component Analysis of a frequency transformation (PLCA Model), and (3), a probabilistic Latent Component Analysis of a Mel-frequency transformation (Mel-PLCA Model). In the following section, the models are explained in more detail. The models are also summarized graphically in Figure 3.1.

3.4.2.1 MFCC model

The first model we built describes each acoustic class by first decomposing each training example into a vector of Mel-frequency cepstral coefficients (MFCCs) and then building a classifier using 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 signal 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.

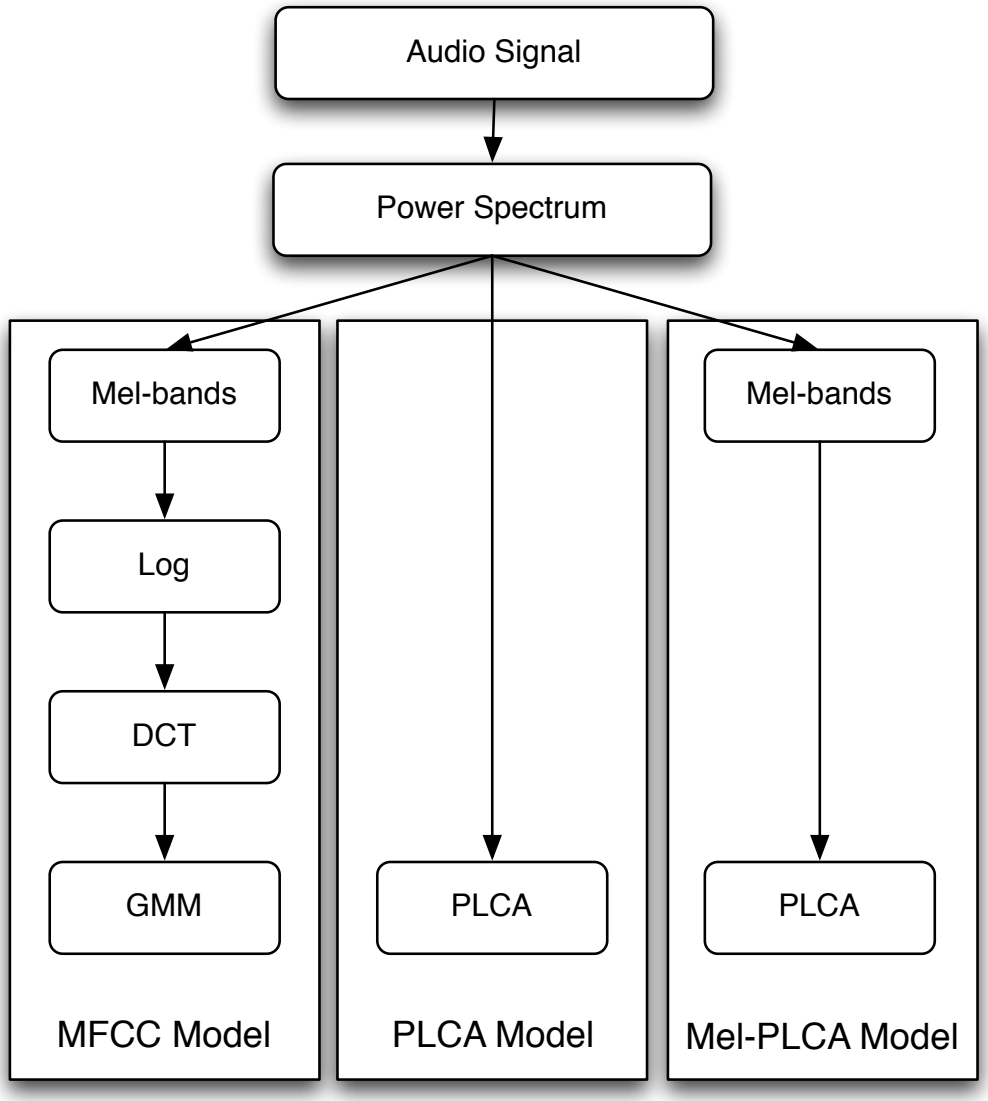

Figure 3.1: Overview of the different models

GMM In order to describe our set of acoustic classes, 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 was then calculated using the posterior probabilities $p(k|\mathbf{x})$ of each of the k components in the Gaussian mixture distribution:

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

3.4.2.2 PLCA model

As pLCA operates on a matrix in order to describe its latent decomposition, we first transposed each audio signal 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), 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 ran pLCA on each of the training example's STFT using 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 Section 3.3. 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 $p(f|k)$.

3.4.2.3 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.

3.4.3 Experiments

3.4.3.1 Experiment 1

Most previous studies in acoustic classification use multiple examples of a single class in isolation. However, as our investigation focused on classification performance during mixtures of classes, we only trained a single example of each class, building a set of 37 classifiers for experiment 1. As a sanity check, we tested whether the MFCC and PLCA models were able to correctly classify the test example built using the 37 classifiers.

3.4.3.2 Experiment 2

For experiment 2, we 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).

3.4.3.3 Experiment 3

For experiment 3, 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 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.

3.4.4 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:

$$TPR_i = \frac{TP}{TP + FN} \tag{3.14}$$

and the false positive rate as:

$$FPR_i = \frac{FP}{TN + FP} \tag{3.15}$$

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 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.

3.5 Results

Experiment 1: Classifying isolated acoustic textures

We tested the performance of a single class in isolation as a sanity check, and as we expected, the performance of the MFCC and PLCA models as determined by the ROC analysis are excellent, with an AUC of within 0.001 of perfect discrimination.

Experiment 2: Classifying acoustic textures 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 1332 mixtures are depicted in Figure 3.2 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 far greater 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.

Experiment 3: Classifying acoustic mixtures

The last test we performed measures the performance 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

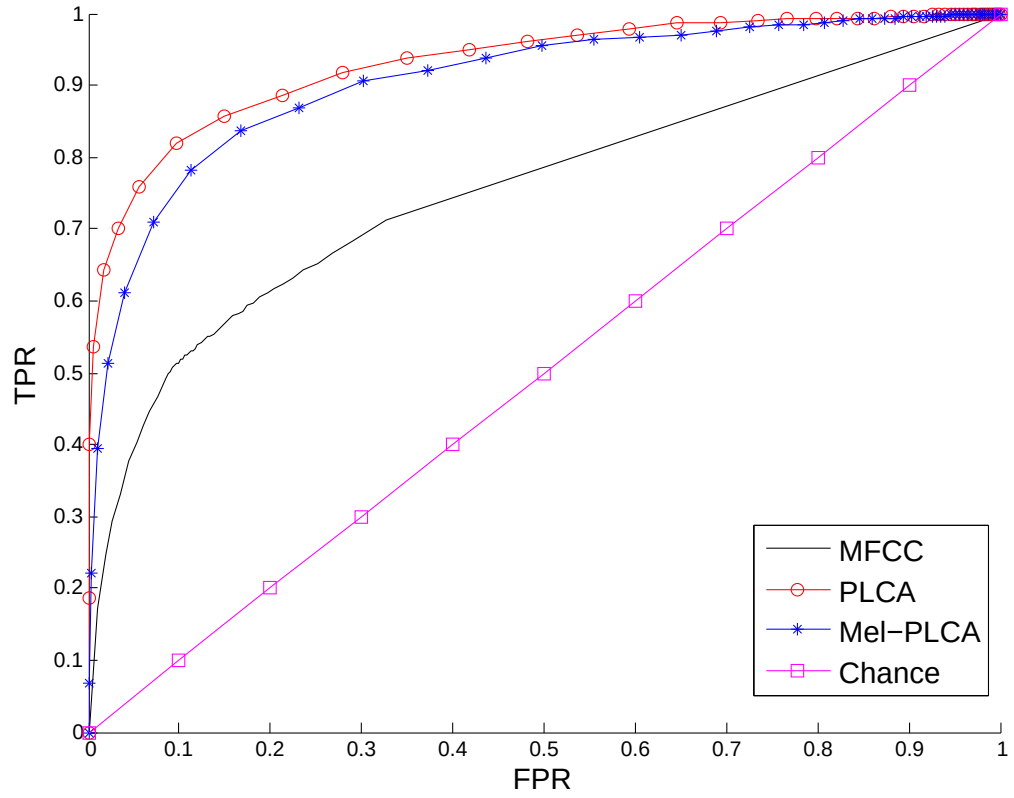

Figure 3.2: Experiment 2: Classification masked by noise.

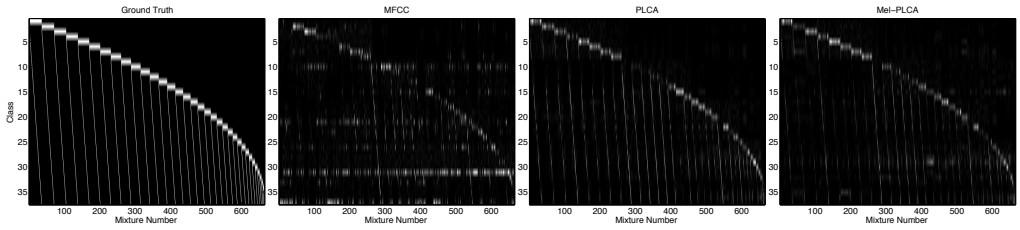

Figure 3.3: Experiment 3: Classification performance of acoustic mixtures depicting the ground truth classes for each of the 666 mixtures and the MFCC model, the PLCA model, and the Mel-PLCA model's classification likelihoods for each of the 666 mixtures. Images represent likelihood of a class in a given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

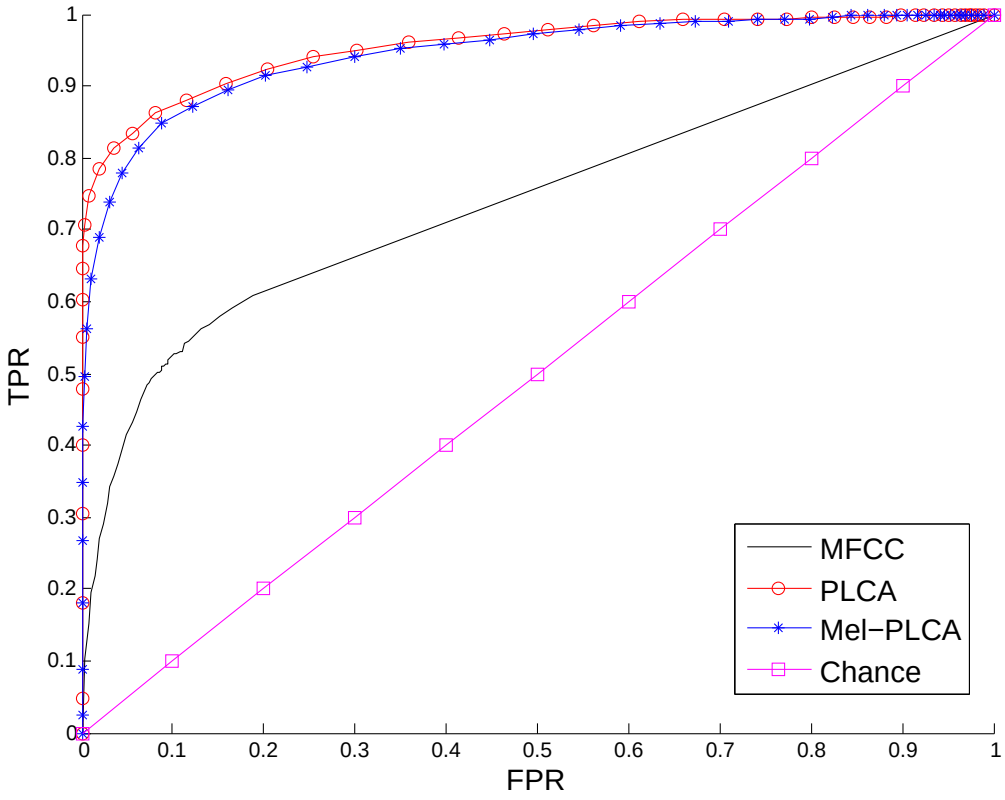

Figure 3.4: Experiment 3: Classification of acoustic mixtures.

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

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 3.4.4 produces 666 ROC curves. The average of these curves are depicted in Figure 3.4, 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.

3.6 Discussion

We tested the performance of 3 types of acoustic classification algorithms, 1 based on MFCCs and 2 based on pLCA. All 3 models performed with excellent results when a single acoustic class appeared in isolation. However, our interests were in how such models performed when classifying the *parts* that make up an acoustic scene. We therefore devised two additional experiments: Experiment 2 masked a known acoustic class by an unknown acoustic class, effectively adding noise; and Experiment 3 tested the performance of each model to classify multiple parts of an acoustic scene by mixing 2 classes together. The MFCC model performed well above chance in both cases with an AUC of 0.74, but the models built on pLCA performed with much stronger results, exhibiting > 0.9 AUC in both experiments.

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 then ambiguity is created, and it becomes 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 make 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.

3.7 Future Work

This work presents an early prototype of a broader framework capable of acoustic source separation and classification for content-based information retrieval. 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 denoising (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-based CBIR algorithms exists. Though the problem is well noted in music information retrieval (Casey 2008; 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.

Chapter 4

Auditory Synthesis

Contents

4.1 Introduction

Chapter 5

Visual Scene Analysis

Contents

Building convincing augmented realities requires creating perceptual mappings between an agent and the augmented content of the environment they perceive. These mappings should be both continuous and effective, meaning the intentions of an agent should be taken into consideration in any affective augmentations. How can an embedded intelligence controlling the augmentation infer the expectations of an agent in order to create realistic and perceivable augmented realities? The current sub-chapter begins to answer this question by reviewing the literature in two essential mechanisms of visual perception: attention and object representation. Beginning with an overview of eye-movements, the review continues to discuss two well-studied phenomena indicative of the architecture of early visual representation: Gist and Change Blindness. Finally, the review concludes in a discussion on building a computational model of visual perception based on the presented literature.

5.1 Introduction

In building an augmented reality, perceptual mappings between an agent and the augmented content of the environment they perceive should be both continuous and effective, meaning the intentions of an agent should be taken into consideration in any affective augmentations. How can an embedded intelligence controlling the augmentation infer the expectations of an agent in creating realistic and perceivable augmented realities? The current sub-chapter begins to answer this question by presenting an overview of literature on visual perception and its possible representations with the final aim of motivating the basis for a computational model.

Beginning with an overview of attention and eye-movements, the sub-chapter continues to discuss a number of presiding architectures for visual perception built

as a result of studies in Gist and Change Blindness. Finally, the review concludes in a discussion on developing a computational model of early visual perception using the presented evidence.

5.2 Attention

Our experience of the world is a rich, continuous, and fully-detailed illusion. Yet, our eyes rapidly move 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*). Going away from the fovea (the *parafovea*), resolution for spatial detail drops logarithmically, while resolution for motion detail increases, a relationship due to the distribution of photo-receptive cells in the eye combined with the lens of the eye itself. We cannot encode with high spatial detail an entire visual scene, as a camera with a small aperture may be able to do, and require saccades (and head-movements) to move our eyes to stabilize (a *fixation*) of our eyes to the region of interest, a process lasting on average 330 ms. During this time, it is thought that encoding into memory occurs as well as planning of the next eye-movement.

The earliest studies in eye-movement behavior (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 relation of a scene and one's endogenous influences on the scene are much less consistent across viewers.

5.2.1 Exogenous Influences on Attention

In seminal work investigating the speed of visual perception using Gestalt 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 known as a *bottom-up* model (Koch 1985; Itti 1998; Wolfe 1989; Itti 2001). Such models are built around theories of feature-integration (Treisman 1980) and are further supported by physiological evidence of the receptive fields and visual architecture of the visual cortex of cats (Hubel 1962). To discover the attentional biases for portions of a scene

(*saliency*), bottom-up models recompose a full resolution image using filter banks tuned to multiple frequency orientations and scales corresponding to pre-attentive visual features also found in early visual cortex such as luminance, oriented edges, and color contrasts. Saliency is then computed as a weighted linear summation (*integration*) of the resulting "feature maps" formed of different scales.

It is thought that basic feature levels of models of integration are modulated by "top-down" influences (Itti 2001) such as the current ongoing task (Yarbus 1967; Smith 2011) and the context of a scene in order to reduce processing load (Henderson 2003; Torralba 2006). Though, the level at which top-down influences may affect processing is still open to debate. 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.

5.2.2 Endogenous Influences on Attention

In a seminal study on how task affects eye-movements during static scene viewing, (Yarbus 1967) 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. It is thought that task, therefore, is an endogenous influence.

In a similar study on dynamic scene viewing, Smith studied task-based effects on viewers' eye-movements looking at unedited videos of natural scenes from a camera mounted on a tripod (Smith 2011). Participants were natives to the city of Edinburgh and viewed a variety of indoor and outdoor scenes from the city. The 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. The study re-asserts the findings of Yarbus, though for a dynamic time-course. Further, it also provides 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.

5.3 Gist

The ability to classify scenes with rapid pre-attentive processing lasting only 45-135 ms (*Gist*) (Potter 1969; Biederman 1974; Potter 1976; Schyns 1994; Henderson 1999) suggests that the general shape and structure of a scene leading one to infer its context are 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). A scene's 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 subsequent processing of gist seem 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 (*priming*). 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. Thus, 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 (textitpriming), 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.

5.4 Change and Inattentional Blindness

Research over the last century 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, unnoticed by 75% of participants. The phenomena of "Change Blindness" was demonstrated in a real-world psychology experiment (Simons 1998) where participants arrived at a kiosk to fill in a consent form and hand the completed form to a man behind the counter. The man ducks behind the counter as to pretend to file the paper, while a different man comes up from behind the counter, again unnoticed by a majority of the participants.

Failing 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. What form, and to what detail this representation encodes is still an open question. Rensink takes this evidence in developing a theory of coherence, proposing that object representation depends on focal attention. 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, representationalless 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).

5.5 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. For example, I may focus on a cup, but not build the representation of the fingerprints on the cups as I was not intending to look at this particular scale. In this case, the endogenous influence of perceiving the object representation of fingerprints on the cup was necessary for building such a representation, even though focal attention will have brought my eyes to the cup. It may be that my task of drinking from the cup saw the cup as what it afforded:

a drink. In a free-viewing task, if such a thing exists, it may be more likely that an exogenous influence such as the mis-representation of the cup will provoke more detailed representations and cause additional focal attention to the cup. Thus, it may be the case that focal attention is necessary for explaining an object, however, it seems it is not sufficient and the cause of focal attention should still be considered.

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 scenes 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.

Considering both the implicit, unmediated representation and the attentional and contextual mechanisms, at least two 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, geons, or proto-objects, where this representation is affected by a logarithmic filter around the point of fixation

based on the evidence of response properties of photo-receptors; (2), 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

Visual Synthesis

Contents

We investigate an approach to the artistic stylization of photographic images and videos that uses an understanding of the role of abstract representations in art and perception. We first learn a database of representations from a corpus of images or image sequences. Using this database, our approach synthesizes a target image or video by matching geometric representations in the target to the closest matches in the database based on their shape and color similarity. We show how changing a few parameters of the synthesis process can result 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, our 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

Figure 6.1: 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.

Mosaicing". Finally, we report the user feedback of 21 participants using an augmented reality version of "Memory Mosaicing" in an installation called ÒAugmented Reality HallucinationsÓ, where the target scene and database came from a camera mounted on augmented reality goggles. This information was collected during an exhibition of 15,000 participants at the Digital Design Weekend at the Victoria and Albert Museum (co-located during the London Design Festival).

6.1 Introduction

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". Finally, we report the feedback of 21 participants using an augmented reality version of "Memory Mosaicing" in an installation called ÒAugmented Reality HallucinationsÓ, where the target scene and database came from a camera mounted on augmented reality goggles.

6.2 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).

6.3 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 6.5.

6.3.1 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 6.4.1, 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.

6.3.2 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:

$$f(i, j) = I_t(i, j) - I_{t-1}(i, j) \tag{6.1}$$

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:

$$R_T = \{R_i \mid f(C_{R_i}) > threshold, \forall i = 1...N_T\} \tag{6.2}$$

6.3.3 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:

$$d_{R_i} = \left(\mu_{00}, \eta_{11}, \eta_{20}, \eta_{02}, L, a^*, b^*\right) \tag{6.3}$$

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:

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{\left(1+\frac{i+j}{2}\right)}} \tag{6.4}$$

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.

6.3.4 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:

$$d(R_t, R_c) = d_s(R_t, R_c) + d_c(R_t, R_c) \tag{6.5}$$

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.

6.3.5 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:

$$\Theta = \frac{1}{2} * \arctan \frac{2 * \frac{\mu_{11}}{\mu_{00}}}{\frac{\mu_{20}}{\mu_{00}} - \frac{\mu_{02}}{\mu_{00}}} \tag{6.6}$$

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.

6.4 Parameters

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

6.4.1 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 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 6.3.4) 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.

6.4.2 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-6.2). 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 (see Figure-6.3). As well, this will also produce more regions that coincide

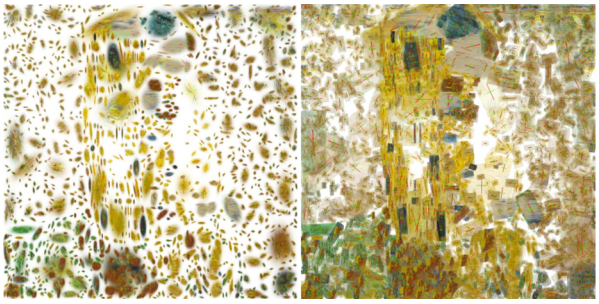

Figure 6.2: Using the target image and database shown in Figure-6.1, we show an example stylization with (first image) and without (second image) spatial blending. We also draw the region's orientation depicted by red/green axes in order to better show the regions (best viewed in the color manuscript at 200%).

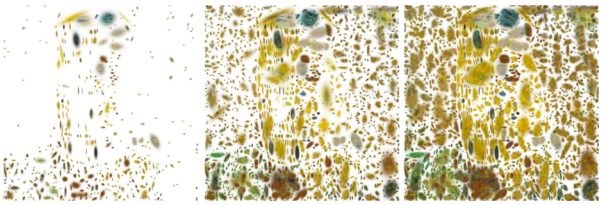

Figure 6.3: Using the target image and database shown in Figure-6.1, the timesteps are increased over time. This allows the user to detect more regions and develop a denser and higher contrast stylization.

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-6.4).

- *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-6.6). Together with an increased number of timesteps, this parameter can be used to change the contrast of the overall image (as shown in Figure-6.7).

Figure 6.4: Using the target image and database shown in Figure-6.1, the minimum region size is decreased over time, allowing the user to detect smaller regions and produce finer detailed stylizations.

Figure 6.5: Using the target image and database shown in Figure-6.1, 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.

- *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-6.5). 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.

6.5 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

Figure 6.6: Using the target image and database shown in Figure-6.1, we increase the temporal blending factor. This influences the opacity of every region drawn.

Figure 6.7: Using the target image and database shown in Figure-6.1, we use temporal blending as well as decreasing minimum region size and increased timesteps to begin to produce the final synthesis.

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". Finally, we present an augmented reality version of "Memory Mosaicing" including feedback collected from 21 participants of an installation at the Victoria and Albert Museum in London.

6.5.1 Image: Landscape

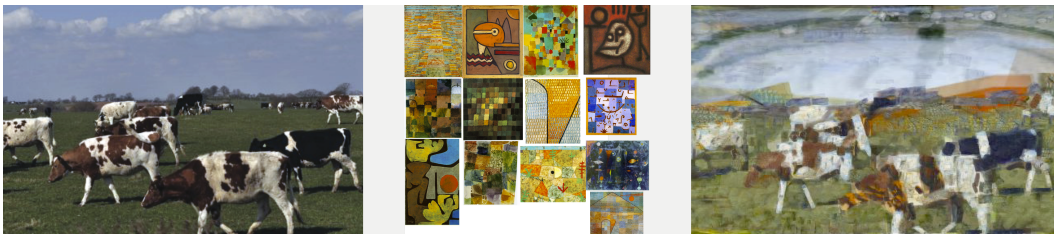

Figure 6.8: A landscape picture of cows grazing is synthesized using 13 images of Expressionism painter Paul Klee to produce the image on the right. Images representing faithful reproductions of Paul Klee sourced from Mark Harden's Artchive are public domain. Photo of cows taken by the author.

In Figure-6.8, 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).

6.5.2 Image: Abstract

Figure 6.9: 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.

In Figure-6.9, 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.

6.5.3 Image: Painterly

Figure 6.10: Van Gogh's "The Bedroom" is synthesized using 3 images of Monet paintings to produce the image on the right. Images representing faithful reproductions of Van Gogh and Claude Monet sourced from Wikimedia Commons are public domain.

We demonstrate how CBVS can stylize existing painterly images into other styles. In the teaser graphic in Figure-6.1, 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 smaller brush strokes, and allow the timesteps to be high as we want to bring out as much contrast as possible.

In Figure-6.10, 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.

6.5.4 Video: Portrait

Figure 6.11: Left: 4 frames from a target video; Right: Stylization using Paul Klee's corpus in Figure-6.8. 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.

Two examples in video-based stylization are presented: one of a subject rowing a boat and another of abstract imagery. In Figure-6.11, we can see 4 frames taken from a video stylization. We use the same corpus as in Figure-6.8 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.

6.5.5 Video: Abstract

Figure 6.12: Left: 4 frames from a target video; Right: Stylization using Paul Klee's corpus in Figure-6.8. Here we aim to stylize with greater abstraction than in Figure-6.11, 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.

In Figure-6.12, we stylize a video using the same corpus as in Figure-6.8 and set the minimum region size to be very large. Thus, instead of producing an Expressionist style as in Figure-6.11, 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.

6.5.6 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 6.13, 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.

6.5.7 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 goggles right eye that faced the scene in front of them (see Figure 6.14). 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, as shown in the left column in Figure 6.13.

Participants were also invited to give quantitative and qualitative feedback on their experience. The summary of the quantitative feedback is shown in Figure 6.15. 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.

6.6 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. Extending this case to an augmented reality setting, a participant of this system can actively view their world, creating a hallucinogenic experience, as validated by a number of participants during an exhibition at the Victoria and Albert Museum. However, the feedback from this installation also revealed a lack of understanding in how the process works.

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 6.13: 2 examples of "Memory Mosaicing" showing the input (top) and resulting real-time stylization (bottom). Photos by the author.

Figure 6.14: An exhibition at the Victoria and Albert Museum in London had participants wear Augmented Reality goggles with software running a real-time version of "Memory Mosaicing". Photos by the author.

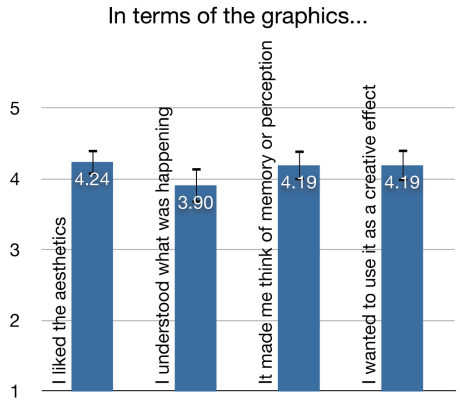

Figure 6.15: Results of the "Augmented Reality Hallucination" installation feedback where 21 participants were asked to rate different aspects of the visual synthesis. Error bars depict +/- 1 S.E.

Appendix A

Appendix

Bibliography

[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 11.)

[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 11.)

[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 11.)

[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 23.)

[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 30.)

[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 page 30.)

[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 37.)

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

[Buswell 1935] GT Buswell. *How people look at pictures.* 1935. (Cited on page 28.)

[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 11 and 22.)

[Casey 2008] 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 23.)

[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 38.)

[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 11.)

[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 11.)

[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 37.)

[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 11 and 23.)

[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 37.)

[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 11.)

[Forssén 2007] Per-Erik Forssén. *Maximally stable colour regions for recognition and matching.* Computer Vision and Pattern Recognition 2007, (CVPR07)., 2007. (Cited on page 39.)

[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 37.)

[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 page 11.)

[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 38.)

[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 36.)

[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 11.)

[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 38.)

[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 30.)

[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 page 29.)

[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 37.)

[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 37 and 39.)

[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 37.)

[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 37.)

[Hofmann 1999] Thomas Hofmann. *Probabilistic latent semantic analysis.* In Proc. of Uncertainty in Artificial Intelligence, UAIâ99, page 21. Citeseer, 1999. (Cited on pages 11, 12 and 13.)

[Hofmann 2001] Thomas Hofmann. *Unsupervised learning by probabilistic latent semantic analysis.* Machine Learning, pages 177–196, 2001. (Cited on page 12.)

[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 30.)

[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 38.)

[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 28.)

[Hughes 1991] Robert Hughes. Shock of the New. 1991. (Cited on page 36.)

[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 page 28.)

[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 28 and 29.)

[Kim 2002] Junhwan Kim and Fabio Pellacini. *Jigsaw image mosaics.* ACM Transactions on Graphics, vol. 21, no. 3, July 2002. (Cited on page 38.)

[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 11.)

[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 28.)

[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 37.)

[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 31.)

[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 38.)

[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 37.)

[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 37.)

[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 40.)

[Manjunath 2002] BS Manjunath and P Salembier. *Introduction to MPEG-7: multimedia content description interface.* WWW-address: http://ipsi. fhg. de/delite/Projects/, 2002. (Cited on page 11.)

[Marr 1982] David Marr. Vision: A Computational investigation into the Human Representation and Processing of Visual Information. 1982. (Cited on page 36.)

[McKinney 2003] MF McKinney. *Features for audio and music classification.* Proc. ISMIR, vol. 4, 2003. (Cited on page 11.)

[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 37.)

[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 28.)

[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 11.)

[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 38.)

[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 12.)

[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 11.)

[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 37.)

[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 30 and 32.)

[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 30.)

[Oliva 2005] Aude Oliva. *Gist of the scene.* In Neurobiology of attention, pages 251–257. 2005. (Cited on page 30.)

[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 33.)

[Orchard 2008] Jeff Orchard and CS Kaplan. *Cut-out image mosaics.* ACM Transactions on Graphics, vol. 1, no. 212, 2008. (Cited on page 38.)

[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 11.)

[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 30.)

[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 30.)

[Pylyshyn 2001] Zenon W Pylyshyn. *Visual indexes, preconceptual objects, and situated vision.* Cognition, vol. 80, pages 127–158, 2001. (Cited on page 32.)

[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 11.)

[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 30, 31 and 32.)

[Rensink 2001] RA Rensink. *Change blindness: Implications for the nature of visual attention.* In Vision & Attention, pages 169–188. 2001. (Cited on pages 30, 31 and 32.)

[Rensink 2002] RA Rensink. *Change detection.* Annual review of psychology, vol. 53, pages 245–77, January 2002. (Cited on pages 31 and 32.)

[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 page 23.)

[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 30 and 32.)

[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 37.)

[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 page 10.)

[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 31.)

[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 31.)

[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 pages 30 and 31.)

[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 11 and 12.)

[Smaragdis 2007a] Paris Smaragdis and B. Raj. *Shift-invariant probabilistic latent component analysis.* Journal of Machine Learning Research, no. 5, 2007. (Cited on pages 11, 12 and 23.)

[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 11.)

[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 page 29.)

[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 11.)

[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 28.)

[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 page 29.)

[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 10.)

[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 10 and 11.)

[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 29.)

[Treisman 1980] AM Treisman and Garry Gelade. *A feature-integration theory of attention.* Cognitive psychology, vol. 12, pages 97–136, 1980. (Cited on page 28.)

[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 33.)

[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 38 and 40.)

[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 37.)

[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 11.)

[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 11 and 23.)

[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 10.)

[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 28.)

[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 11.)

[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 37.)

[Yarbus 1967] Alfred Yarbus. Eye movements and vision. 1967. (Cited on pages 28 and 29.)

[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 38.)

[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 36.)