A STUDY ON IMAGE-BASED MUSIC GENERATION

by

Xiaoying Wu
B. ASc, Nanyang Technological University, 1997
M. Comp, National University of Singapore, 2001

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
School of Computing Science

SIMON FRASER UNIVERSITY

Spring 2008

APPROVAL

Name: **Xiaoying Wu**

Degree: **Master of Science**

Title of Thesis: **A Study on Image-based Music Generation**

Examining Committee:

Chair: **Greg Mori**
Assistant Professor, School of Computing Science

Ze-Nian Li
Senior Supervisor
Professor, School of Computing Science

Tamara Smyth
Supervisor
Assistant Professor, School of Computing Science

Mark Drew
Examiner
Professor, School of Computing Science

Date Defended: April 7, 2008

ABSTRACT

Visual and auditory forms have noticeable associations that can inspire similar cognitive and aesthetical experiences. This study investigates the possibility of applying various visual-auditory associations in music generation. Several algorithms are proposed in order to create music segments from image features such as contour, colour and texture. Test results show that the algorithms can produce interesting and pleasant music segments in many cases. Integrating the generated segments together can create longer and more intriguing results.

The generation process totally depends on the image content and avoids the problem of introducing pseudo randomness in many algorithmic music composers. Different images with similar features are able to generate similar music patterns. For very different images, some of the algorithms can generate music of different styles. This study shows that using image as a source of music generation has a high potential in creating interesting and fresh music.

Keywords: Image sonification, Image-based music generation, Auditory display, Algorithmic composition.

ACKNOWLEDGEMENTS

I wish to express my deepest appreciation to Dr. Ze-Nian Li for his thorough and patient guidance through the whole period of my study in Simon Fraser University. This thesis cannot be successfully accomplished without his great help and invaluable support.

I also wish to thank Dr. Tamara Smyth for her insightful suggestions on the direction of this study and helpful advices on my work.

Dr. Mark Drew receives my special thanks for inspiring me the initial idea of this work and his precious advices.

Finally, I wish to thank my husband, Zhengang Wang, for his unlimited support and help in this way or that during the period of this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1. Motivation and objective

Associations between musical and visual forms have been noticed since very early time of human history. There are many examples of poetries and artists describing scenes in musical terms, or musicians describing music in visual terms. This kind of phenomena can be found in many different cultures in the world, hence is considered a common aspect of human cognitive and perceptive experience. It suggests that, though musical and visual forms are perceived differently, they can arouse similar cognitive and aesthetic experiences. Inspired by this point, this study attempts to generate music from images, especially images of natural scenes whose aesthetic values are generally appreciated by human.

Algorithmic music composition has achieved great success in the past decades, especially in the direction of music style learning and imitation [Cope05]. To model the creative process of music composition computationally, many algorithms make use of stochastic processes to allow changes and variations. On the contrary, this study focuses on transforming existing features (colour, contour, and texture) of images into musical materials. More precisely, this study investigates possible visual-auditory associations and tries to apply them to music generation.

1.2. Scope and application

Music generation can be a very large topic that includes all aspects of music, like melody, chord structure, rhythm, timbre, sound synthesis, sound effect, performance indicator, etc. In the scope of this study, music generation is restricted to pitch, chord, and duration only (in this context chord is in its general meaning: notes that sound simultaneously). In other words, the task is to convert an image into one or more series of notes of certain pitches (or chords) and durations.

Experimental results suggest that this approach has a high potential in creating interesting music segments. These segments can be used to create more complete music in real compositional process. Another application is accompanying music generation in various contexts, especially those requiring short pieces only. For example, electronic greeting card, game environment, interactive multimedia system, etc.

1.3. Thesis structure

This chapter introduces the motivation and objective of this study. The next chapter presents some related works in the past. Chapter 3 describes the details of generating music segments from image features, like contour, colour and texture. Chapter 4 discusses the methods and issues in combining music segments together. Finally, Chapter 5 gives conclusions and some suggestions for further work.

CHAPTER 2: BACKGROUND REVIEW

2.1. Visual-auditory associations

Many evidences exist in linking colour and music by people with or without musical backgrounds. Traditionally, musicians often use colours to depict musical notes. In fact, the use of colour is so natural for musicians that they name the 12-semitone music scale as *chromatic* scale. Tonal colour is often used as a synonym of timbre. In some psychological study on synaesthesia, people claim that they can "see" colours while listening to music.

The advancement in physics in 17th century has given people more scientific understanding of colour and sound. Visible light is an electromagnetic wave at various wavelengths and its colour is an overall sensation of its light spectrum. Audible sound is a pressure wave in air and its pitch is determined by the fundamental frequency in its sound spectrum. These findings suggest an association between colour and sound in a more scientific way.

Perhaps the first scientific attempt of associating colour and sound is Newton's colour music wheel in his famous work Opticks [Newton52]. This wheel maps the seven prism colours (red to violet) to the seven tones of a diatonic scale (D, E, F, G, A, B and C), as shown in Figure 1. The mapping is based on the wavelength distributions of colours and tones in their respective spectra.

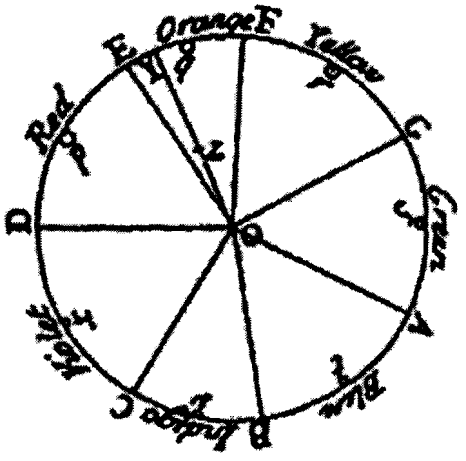


Figure 1. Newton's Colour Music Wheel, from source [Colour Music] by permission.

In this wheel, each colour corresponds to an interval between two consecutive pitches. For example, Violet correlates to C-D, Red correlates to D-E, etc. Most intervals between two pitches are full tones (2 semitones), except intervals E-F and B-C, which cross only one semitone. They are mapped to Orange and Indigo respectively, which are considered occupying less space than others in the colour spectrum. The reason for a wheel representation is from two perceptual observations. In the colour spectrum, the last colour Violet meets the first colour Red seamlessly in human eyes, though their wavelengths are far apart. A similar phenomenon exists in the sound spectrum, where two pitches one octave apart are strongly consonant in human ears and are considered of the same tone.

More mappings of colour and sound appeared in the days after Newton. Most of them are similar to the colour music wheel except the ways of separating colours and pitches [Caivano94]. One common practice is to separate the colour circle into 12 colours such that they can be mapped to the 12 pitches of the chromatic scale. One interesting observation is: at two ends of the visible colour spectrum, the wavelength of red is twice

the wavelength of violet [Caivano94] [Garner78]. Though the wavelengths are chosen only approximately (one source uses 760 and 380 while another one uses 800 and 400 nanometres), this phenomenon coincides well with the fact that an octave interval has one end twice the frequency of the other.

Wells [Wells80] compares the colour circle formed from traditional painters' primary colours (Red, Yellow and Blue) and secondary colours (Orange, Green and Violet), with the chromatic scale, which equally divides one octave into 12 semitones. He notices that opposite colours (e.g. Red and Green) and chords formed by opposite tones (e.g. C and F#) suggest a similarity of clashing, struggling and dramatic effect in art and music respectively. Based on that correlation, he also suggested some chord formation methods.

Besides these similarities between hue and pitch, there are also many arguments about their differences [Garner78] [Pridmore92]. One problem with the early mappings is that there are many (9-10) audible octaves but only one colour circle. Pridmore addresses this issue by using a spiral to represent multiple octaves such that pitches in the same tone map to the same hue but in different cycles [Pridmore92]. In a recent study, Giannakis adopts a similar approach such that pitches in the same tone but higher octaves map to colours in the same hue but higher luminosity levels [Giannakis06].

Many other arguments exist in associating hue with pitch. Perhaps the most important difference between them is the ways human perceive them. For example, when played together, pitches in a chord are still separately distinguishable to ear. However, hues mixed together produce a different colour whose ingredients are not separable to eye [Garner78]. Human perception on pitches mainly depends on distances between them,

not their exact frequencies. Hence, a song can start at any key and sounds almost the same. A chord of fixed intervals can have different roots and remains recognizable. On the contrary, hues are determined by their absolute frequencies, though sometimes the perception of a hue is affected by its neighbouring colours.

In addition, hues do not have perceptual ordering though one can physically order them by their frequencies. There is no reason that red should come before blue or vice versa. Hues are qualitative instead of quantitative entities. However, pitches have natural ordering to human ear. For this reason, Viora finds that it makes more sense to correlate hue with timbre (also known as sound quality or sound colour), as timbre is considered perceptually qualitative [Viora82]. On the same basis, Barrass goes further by associate pitch with luminosity as both are quantitative data [Barrass96].

Other aspects of sound and colour are also associated in various studies. Besides supporting the hue-pitch association, Caivano [Caivano94] correlates sound loudness to colour luminosity, timbre to colour saturation, and duration of sound to size of colour, as shown in Table 1. These correlations are all based on the physical characteristics of light and sound. For example, both sound volume and colour luminosity are proportional to their intensity or energy consumption. The correlation of timbre and colour saturation is because they are both determined by their compositional complexities. A saturated colour corresponds to a pure sound since both occupy a narrow portion of their spectra (single wavelength); a dull colour corresponds to a complex sound since both occupy a wide portion. Caivano also supports his associations by some psychological experiments where 80% of the people relate saturated colour with pure sound, while only 20% of the people choose the opposite association (i.e. saturated colour with complex sound).

Table 1. Caivano's visual-auditory mappings

Visual Elements	Auditory Elements
Hue	Pitch
Saturation	Timbre
Luminosity	Loudness
Size	Duration

To empirically identify the possible associations between colour and sound in human perception, Giannakis [Giannakis01] conducted an experiment based on a group of 24 users. In the experiment, users need to associate a sequence of pure sounds of various pitch and loudness, to colours from a colour palette of various hue, saturation and luminosity. His result shows a strong correlation between colour saturation and sound loudness, i.e. a highly saturated colour is more associated with a loud sound. Giannakis also concludes the existence of an association between pitch and colour luminosity, though this assumption is not strongly supported by the experiment statistics. Another interesting observation is that users constantly ignore hue while doing the experiment. As timbre is excluded from this experiment, any possible association between timbre and colour is not under consideration. This could be the possible reason why people totally ignore hue, if in fact hue should be associated with timbre.

To explore the perceptual associations between other auditory and visual elements, Giannakis continues his investigation and suggests a mapping between timbre and visual texture [Giannakis06]. An empirical study is carried out to compare his mapping with *sonogram*, a well-known visual representation of sound in two-

dimensional space. In sonogram, the horizontal dimension represents time; the vertical dimension represents frequency; and the pixel intensity represents the amplitude of sound at the particular time and frequency. Giannakis associates three dimensions of timbre, i.e. sharpness, compactness and sensory dissonance, with three dimensions of texture: coarseness, granularity and periodicity. Though constrained by a small sample size (8), the experiment shows that visual representations of sounds based on these associations are more comprehensible by users than sonogram. Table 2 summaries the visual-auditory associations identified by Giannakis:

Table 2. Giannakis's visual-auditory mappings

Visual Elements	**Auditory Elements**
Hue, luminosity	Pitch
Saturation	Loudness
Texture	Timbre
Coarseness	Sharpness
Granularity	Compactness
Periodicity	Sensory dissonance

Discussion

The research in the associations of colour and sound remains inconclusive. Though there are some physical correlations between them, these correlations do not necessarily translate to corresponding correlations in human perception. On the other hand, it is a subjective and complex phenomenon when people associate colour and

sound, and is largely affected by the number of variables to be considered at the moment. Different patterns of association could result when given a different combination of the variables. In many cases, it is natural for people to associate a series of proportionally increasing values to another series of proportionally increasing values, though they may be in very different domains. If only one pair of variables is considered, e.g. loudness and luminosity, people tend to associate louder volume to brighter luminosity. In case of qualitative variables (e.g. hue, timbre) from two different domains, they do not inspire this kind of obvious association and are likely to show either no association between them or different people will have their own associations. All the proposed associations will make sense under some particular conditions, but no universal relationship predominates at this moment.

2.2. Algorithmic composition

Large amount of research effort has been dedicated to algorithmic music composition in the past decades. The main approaches of algorithmic composition can be categorized into mathematical models (e.g. stochastic processes and Markov chains), knowledge based systems, grammars, evolutionary methods (e.g. genetic algorithms), systems that learn (e.g. neural networks and machine learning), and hybrid systems [Papadopoulos99][Cope05].

A fundamental problem common to most, if not all algorithmic composers, is how to measure the quality of the generated music. The simplest method is to use a human to evaluate each music piece subjectively [Robertson98]. This method lacks of efficiency, especially in evolutionary methods where a number of music segments need to be evaluated in every evolution iteration. Instead, objective fitness functions can replace a

human evaluator to test the "goodness" of the generated music. A fitness function can be based on concepts in musical theory like consonance degrees of musical intervals, on statistical models empirically created by induction from a corpus of music genre, on priori training of a neural network, etc [Papadopoulos98][Moroni00][Conklin03]. It is also common to use both a human evaluator and some fitness functions together to compensate the shortcomings of each other, especially in real-time interactive systems [Moroni00].

2.3. Visual interface in music composition

With the fast development of graphical user interface, visual elements have been widely accepted as a direct and efficient means of communication. Music composition systems have also made extensive use of visual elements in their interfaces to enable non-traditional composition experience.

Some music systems like UPIC, Phonogram and Metasynth, directly make use of a graphical representation of sounds similar to *sonogram* (see *Section 2.1.*) in their compositional process [Xenakis92] [Lesbros96] [Metasynth].

Lesbros' representation, named *phonogram*, uses logarithm of frequency (instead of simple frequency in sonogram) in the vertical axis to represent uniformly scaled pitches [Lesbros96]. The horizontal axis represents time. A composer creates music by drawing patterns on paper, which is scanned and processed by 16 synthesizers to produce sounds. This system requires some understanding about the associations between the drawing patterns and the corresponding sounds before successful composing. Though phonogram is a physical representation of sound, it does give composers some perceptual

ideas of interpreting visual elements as music, especially after training. For example, an upper position in the image corresponds to a higher pitch, a wave of line corresponds to a wave of melody, a darker pixel corresponds to a louder volume, a pattern of sound can be pasted to anywhere and remains its internal structure, etc.

Metasynth, an award-winning software, uses a similar approach to sound synthesis and composition [Metasynth]. An interesting point about Metasynth is that it makes use of colour (RGB) to provide spatial information of sound in the three-dimensional space (green for left, yellow for centre and red for right).

Many motion-based interactive musical systems use image-processing techniques to produce music. Iamascope is an interactive system that produces large-screen imagery with accompanying music in response to user's movements in front of a camera [Fels99]. It uses a pie slice of the captured image to produce kaleidoscope-like imagery and to generate the corresponding music. The pie slice is divided into N bins; each bin corresponds to a note. If the slice in the next frame of the captured video has a big intensity difference from the current one (i.e. some movement happened), the corresponding bin will be turned on by signalling a MIDI NoteOn event. The pleasant music produced by Iamascope is partly due to its limited choices of notes, as only notes C, E, F are used in an octave (in fact these notes constitutes the most popular C Major chord). To get more varieties in the produced music, Iamascope allows user to choose a sequence of keys as the root note and cyclically goes through them at given intervals.

2.4. Image-based music generation

2.4.1. Image sonification

Data sonification is traditionally viewed as an alternative to visualization that provides different perception and new insight into the data. Image sonification for musical purpose, however, has gained some research attention in recent years [Kabisch05][Yeo05].

As image contains 2D planer data, the first problem of converting image to music is to map 2D planar data to the time axis. Yeo proposes two concepts of time mapping (scanning and probing) to address this problem [Yeo05]. Scanning refers to reading the image data in a fixed order. For example, reading each column of pixels from left to right of an image is a typical scanning process. Probing, on the contrary, does not have a fixed order in reading image data. For example, a user can arbitrarily go to visually salient areas of interest in the image.

In an image sonification system, natural and manmade landscapes are used as source material to create accompanying music [Kabisch05]. An edge detection algorithm extracts landscape contours (like city skylines). As the user's position moves horizontally along the landscape image, the vertical location of the corresponding contour pixel is mapped to pitch. This approach is quite intuitive as the resulting sound mapping allows the user to trace the landscape contour as pitch increasing or decreasing.

2.4.2. Converting image to music using colour

A recent study shows the possibility of converting image to music using image colour information [Margounakis06]. This work is based on a previous research that

computes "chromatic indices" for a piece of music [Politis04]. The chromatic index of a segment of music is a real number (in the range of 1.0 to 2.1) indicating how "chromatic" the segment of music is (coarsely how strong the feelings are evoked by the music). The underlying scale of the music determines a base chroma value, while the melodic flow plots a curve of chroma values based on how big the intervals are between consecutive notes. The chromatic index of the music is then the average of the chroma values on the curve. With a mapping of chromatic indices to colours, a piece of music can be visualized as a sequence of coloured boxes indicating their chromatic indices. The length of each box indicates the duration of the corresponding music segment.

The chromatic index of a music scale is defined from its notes intervals. Higher weights are given to less commonly used intervals such as semitones and quarter-tones and notes foreigner to the scale (they are considered adding more *chroma* to the music). In this sense, a Minor scale has a higher chromatic index than a Major scale, while an Arabic scale has an even higher value. The author produces a table of chromatic indices of about 70 music scales from both western and eastern cultures.

The mapping of colours to chromatic indices is based on the association between colours and emotional states. A table of 12 colours is listed to represent 12 levels of emotional states, from the weakest to the strongest feelings. In this order, the 12 colours are: white, sky blue/turquoise, green, yellow/gold, orange, red, pink, blue/royal blue, purple, brown, gray, and black. The chromatic values are linearly mapped to these colours, with white mapping to the smallest chromatic value (weakest feelings), and black to the biggest value (strongest feelings).

Using a reverse method of computing chromatic index, a sequence of coloured boxes can be transferred to a sequence of music segments and form a whole piece of music [Margounakis06]. Margounakis uses the terms "chromatic brick" and "chromatic wall" to represent the segments and the whole piece respectively. So composing becomes a task of building a chromatic wall. By choosing the colour and length of a brick, the composer decides the music feeling and duration of a segment. However, as many music segments have the same chromatic index, it is not sufficient to determine the music segment uniquely from a colour. The author proposes two solutions. One is to use a predefined brick database such that each brick can have a set of candidates to choose. Another approach is to get random notes that produce the desired chromatic index value, guided by some heuristics. The author also shows an encouraging result of a piece of music generated from a small image of smiling face.

2.5. Visual music

The long perceived visual-auditory associations have led to numerous attempts to create visual arts musically, or visual music in general term. In a recent definition, visual music is "time-based visual imagery that establishes a temporal architecture in a way similar to absolute music" [Evans05]. With its "non-narrative and non-representational" nature, visual music is well accepted as a form of abstract art. It is popular for nowadays' music players to show changing shapes and colours synchronized with the music.

The approaches used in visual music composition are generally related to the links between visual and musical elements, possible in various levels of abstraction. For example, Pridmore's sound-to-light transducer directly associates low-level elements like hue and brightness to pitch and loudness respectively [Pridmore92]. In another attempt,

dominant patterns in melody are abstracted and matched to predefined visual pattern candidates [Pocock92]. Evans tries to address higher-level phenomena in music, like resolving from dissonance to consonance, to similar occurrences in visual domain [Evans05].

As a form of abstract art, there is no absolute theory of visual music creation, though some general principles exist. As Karinthi points out, it is essentially a matter of choice by visual music artists [Karinthi91]. However, the success in this area has suggested both possibility and feasibility of translation in the reverse direction: converting visual arts to music.

CHAPTER 3: GENERATE MUSIC SEGMENTS FROM IMAGE FEATURES

This chapter describes a few basic methods of generating music segments from image features. The generated segments are of various lengths, generally several bars (or measures) of notes. The explored image features include *contour*, *colour* and *texture*.

Despite the image feature in use, the music generation process consists of three basic steps: *partitioning*, *sequencing* and *mapping*. *Partitioning* is the first step that separates the image into individual units. The *sequencing* step decides the sequence of these units along the time axis. Finally, each unit is converted into one or more musical note in the *mapping* step. Figure 2 shows the flow diagram of this process. An optional pre-processing step is included if any pre-processing is necessary for the input image.

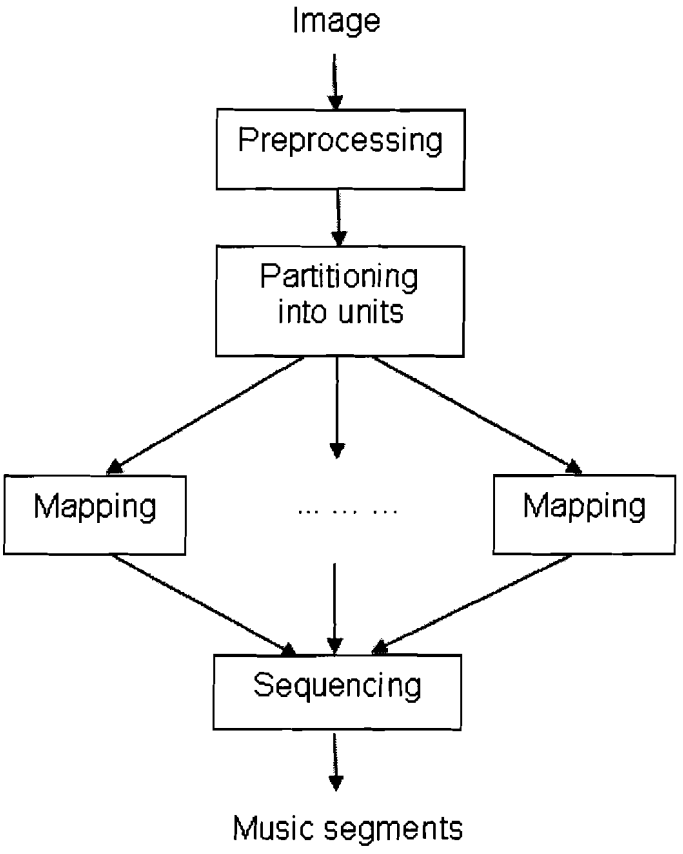


Figure 2. Flow diagram for the image to music conversion process.

3.1. Contour

Many images of natural scenes contain interesting contours. For example, leaves, mountain ranges, waves, etc. Chapter 2 describes some examples of using contour in music composition and image sonification in the literature [Lesbros96][Kabisch05]. These examples have shown that interpreting image contour as pitch contour is quite intuitive to user.

The usual approach is to treat the horizontal axis of the image as time and the vertical axis as pitch level. This approach, however, pays more attention to the global location of the contour in the image instead of its shape information. This study attempts to address this issue by tracing along the contour. More pitch mapping methods are also proposed and discussed.

3.1.1. Pre-processing, partitioning and sequencing

OpenCV is a popular and powerful open source library of computer vision algorithms [OpenCV]. This study makes use of this library in many image-related tasks. Contours are extracted by using an available function in this library. The extracted contours are approximated by straight-line segments. Finer approximation introduces more line segments. Figure 3 shows the extracted contour of a leaf object. The dots are the vertices connecting the line segments.

There are two possible directions of tracing a contour. For an open contour, one can read it from one end to the other or the other way round. For a closed contour, one can read it in either clockwise or counter-clockwise direction.

In the actual implementation, an interactive window is available to allow the user to select an interesting contour (or portion of it) and to specify the direction of reading it. This is done by simply selecting three vertices over a connected contour. The selected contour will be read starting from the first vertex, following the second vertex and ending at the third vertex. The second vertex is necessary as the contour is possibly closed. Without the second vertex, it is not possible to tell whether the reading direction is clockwise or counter-clockwise if the contour is closed.

For example, in the contour shown in Figure 3, the user selects three vertices by circling them in order, as indicated by the black circles labelling 1, 2 and 3. With this ordering, the contour starts from the red vertex, follows through all the green vertices, and ends at the blue vertex.

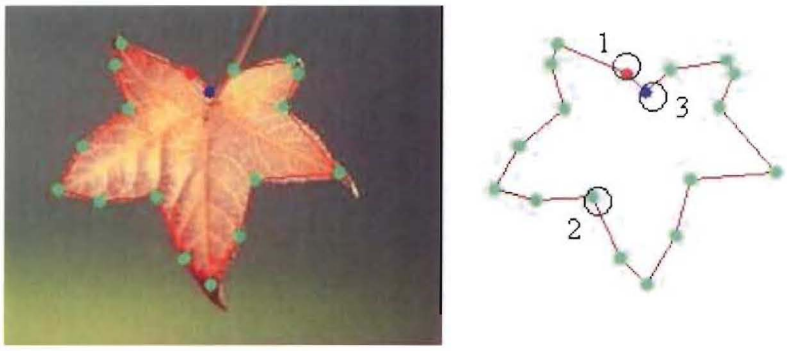


Figure 3. Extracted contour of a leaf.

3.1.2. Mapping

Once decided the contour and the sequence of reading it, a sequence of musical notes can be generated by mapping each line segment into a musical note. Three mapping methods are considered.

The first method maps the Y coordinate of each line segment's starting vertex to pitch. By setting a scale of pitches, the Y coordinates are proportionally mapped to the scale. Two pitch scales are used. One is the uniform Chromatic scale. Another is the non-uniform Major scale.

Figure 4 shows the mapping of a leaf contour to one octave of pitches. The right part of the figure shows the pitches of the Chromatic scale and the Major scale. The Major scale contains 7 pitches: C, D, E, F, G, A and B. The Chromatic scale contains 12 pitches: C, C#, D, D#, E, F, F#, G, G#, A, A# and B. The contour is placed in a local coordinate system with the X-axis at the bottom-most vertex and the Y-axis at the left-most point of the contour. In this system, the Y coordinate of each line segment's starting vertex is mapped to an octave of pitches proportionally. For example, when using the Chromatic scale, the vertex crossed by the red line is mapped to the note C#. If using the Major scale, it is mapped to the note D.

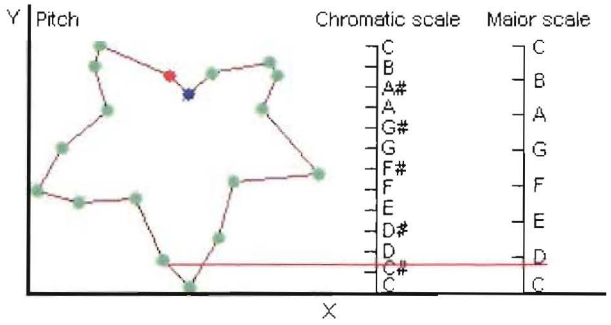


Figure 4. Mapping contour height to pitch scales.

The duration of each note is proportional to the line segment's length. A longer line segment produces a longer note. Figure 5 shows the results of mapping the contour in Figure 4 to both the Chromatic scale and the Major scale. The vertical axis represents

pitch. The horizontal axis represents time in the unit of whole notes. One whole note is subdivided into eight portions, indicating eighth notes. For example, the first note takes two eighth notes and is a quarter note, the second, eighth.

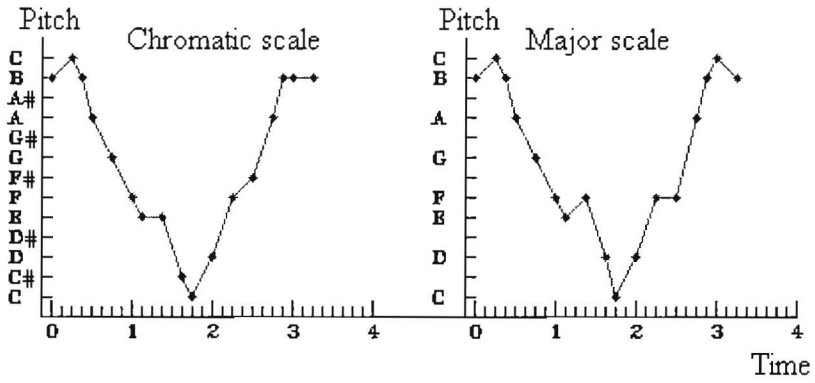


Figure 5. Mapping contour height to pitch results.

The other two mapping methods are similar to the first one. The only difference is that the second method maps the line segment's slope to pitch. The third one maps the derivative (slope change) to pitch.

Figure 6 shows the mapping of contour slopes to pitches. Each line segment is considered as a vector pointing from its starting vertex to its ending vertex. The slope is defined as the angle of the vector in the coordinate system introduced in Figure 4. The angle range is from -180° to 180°, as illustrated in the left part of Figure 6. The middle part of the figure shows the pitches of the contour mapping to an octave of the Major scale. The first segment starts from the red vertex and ends at the neighbouring green vertex. Its slope is around 150 ° (the biggest slope along the contour) and is mapped to the higher C note. The smallest slope (about -170 °) is the second last line segment. It is mapped to the lower C note.

The right part of Figure 6 shows the mapping of contour derivatives to an octave of pitches in the Major scale. Contour derivative is defined as the angle between two neighbouring line segments. The range is from 0° to 360°. The sharp corners in the contour are converted to pitch peaks. For example, in the pitch flow there are five peaks. Figure 6 corresponding to the five acute corners in the contour.

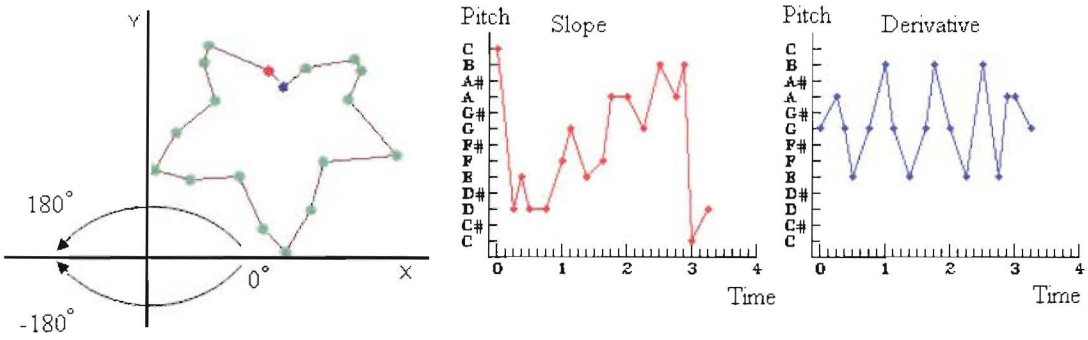


Figure 6. Mapping contour slopes and derivatives to pitches.

3.1.3. Discussion on results

All three mappings generate interesting results in many cases. They can be used as the raw materials in music composition. The number of line segments increases with a better approximation of the contour. In this case, the pitch flow takes a similar outline but the number of big pitch intervals is reduced in all three mapping methods.

Comparing the Chromatic scale and the Major scale, the generated pitch flows are similar (like the example shown in Figure 5). Only a small amount of differences exists, especially when the pitch range is small. Those differences are generally at pitches in the Chromatic scale but outside the Major scale. There are five of them: C#, D#, F#, G# and A#. They may introduce an unpleasing feeling if the listener is more familiar and comfortable with the Major scale that is popular over the world for centuries.

Similar contours are able to generate similar pitch flows from different images. For example, the two leaves in the first column of Figure 7 have a similar shape. The pitch flows of the upper leaf are plotted in red; those of the other leaf are in blue. The second column of Figure 7 shows the pitch flows using the height mapping method. The third and the last columns show the slope and the derivative mapping methods respectively. Comparing the pitch flows of the two leaves, they are highly similar to each other under each mapping method.

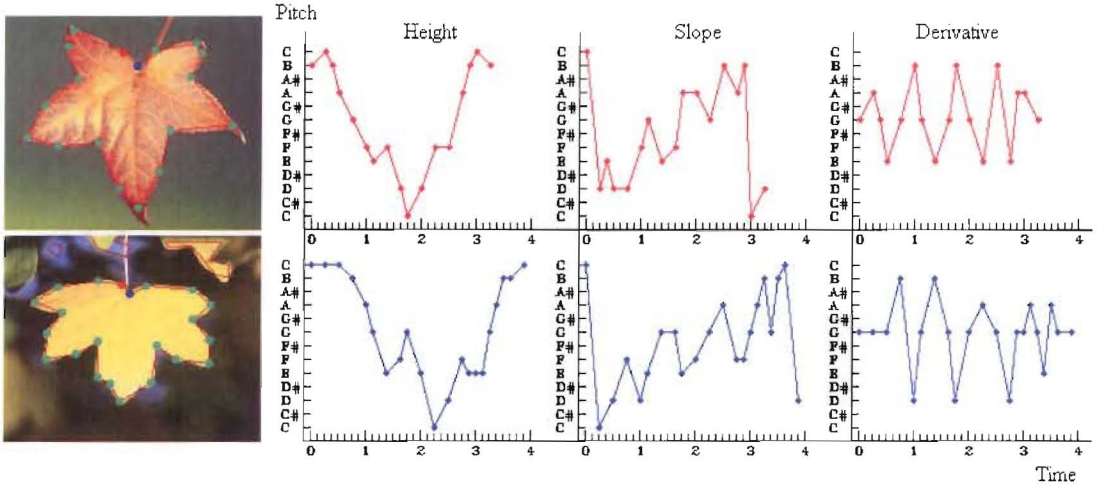


Figure 7. Comparing two similar contours.

This approach of using contour is different from some previous studies [Kabisch05]. Besides considering contour slope and derivative in addition to height, the path of reading the contour is also different. It traces along the contour starting from an arbitrary vertex, instead of scanning the contour from one side of the image to the other side horizontally as in some previous studies. Scanning the image horizontally is straightforward and contour-independent. However, tracing along the contour is orientation-free in the case of the derivative mapping method. If two shapes are identical

but in different orientations, they will still generate the same or very similar pitch flow. An image with multiple occurrences of a specific shape can have the similar pitch flow occurring multiple times. This is conforming to a common musical phenomenon that the same theme is repeated several times in a piece of music with possible small variations.

Comparing the three mapping methods, the first one is quite intuitive to the user as the direction of the contour can be synchronized with the rising and falling of the pitch flow, especially in an environment where the image and the generated music are presented together. The result of the third method can also be easily associated with the contour as it produces high pitches at acute corners and low pitches at obtuse ones. However, the derivative method produces less fluctuation in pitch flow for smooth contours. In an extreme case, a circle-like contour will give almost the same pitch. The second method is not very intuitive as it is not easy to differentiate the slope of each line segment.

3.2. Colour

Besides contour, colour is another important feature of image. Many images contain rich colour information. The variations and transitions of colour are generally smooth and soft and sometimes even suggest certain rhythmic pattern. This is especially true for images of natural scenes.

A previous study has shown a method of converting colours to music segments based on their *chroma* values (see *Section 2.4.2.* for more details). However, the generated music is not totally depending on the image colour information. As the mapping of colour to music segment is not unique, the music segment is either chosen from a predefined database or generated randomly with some heuristics. Moreover, as one pixel may generate several musical notes, a small image (15×15) can generate long result.

This section introduces the study of using the colour information of the whole image to generate music. Instead of reading the image pixel by pixel, the perceptually apparent colours are extracted and used. Several methods of mapping colours to pitches are proposed and discussed.

3.2.1. Colour model and colour distance

As the purpose of this study focuses on the perceptual connection between image and music, the perceptual categorization and interpretation of colour is considered more important than its physical understanding. For this reason, the colour model chosen is the Hue-Saturation-Value (HSV) model as it better describes human perception of colour than the common RGB model while remains relatively simple and easy to implement.

The HSV model builds a colour space with three dimensions: Hue, Saturation and Value. The whole colour space can be represented by a cylinder, as shown in Figure 8(a). The height of the cylinder represents the Value dimension; the cylinder radius represents the Saturation dimension and the angle represents the Hue dimension. A point in the cylinder represents a colour with its angle as hue, its distance to the centre axis as saturation, and its height as value (or luminosity). Along the Value axis, colour changes from black (the smallest luminosity) to white (the biggest luminosity). As one can see, when luminosity decreases, the number of perceptually distinguishable colours also reduces. In the extreme case, when luminosity is at the smallest end, there is only one colour, black, regardless of what hue and saturation may take. To rectify this uneven colour distribution problem, the base of the cylinder can shrink to form a cone, as shown in Figure 8(b).

The cone representation provides a relatively simple way of computing perceptual distance between any two colours – the Euclidean distance between the two points corresponding to the two colours. Though the cone representation is not proven uniform in perceptual colour measurement, it is definitely better than the cylinder representation or the RGB model. Without a single well-established model for perceptually uniform colour measurement, the HSV cone model is used as a practical substitution in this study.

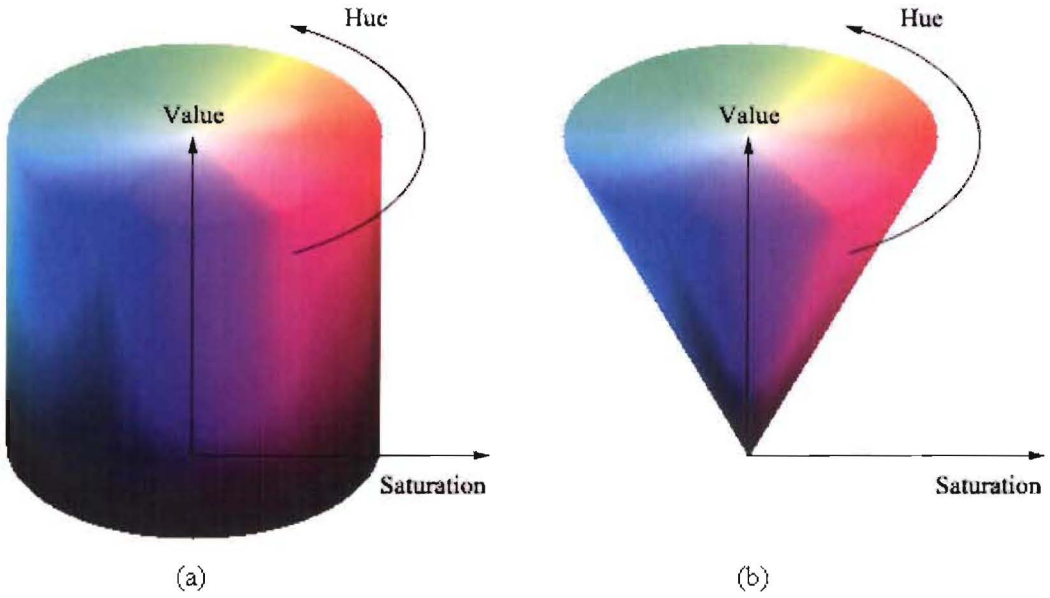


Figure 8. HSV colour model representations: (a) cylinder; (b) cone.

The remaining work is to decide the proportion of the cone. This is done by making some assumptions about the distances between several typical colours. Consider the prominent points of the cone. At the centre of the base circle is the brightest white. At the apex of the cone is the darkest black. Edge points on the base circle represent the brightest and purist colours, like pure red, pure orange, etc. Among them, points opposite to each other are complementary colours, like red and cyan, blue and yellow, green and magenta, etc. If assuming the perceptual distance between the complementary colours is the biggest and the range of distance is from 0 to 1, then the cone base has diameter 1. Assuming the distance between black and any pure colour is the same as the distance between complementary colours, then the cone side edge is also 1. Based on these assumptions the base circle diameter and the two side edges connecting the apex form an equilateral triangle.

Using this cone model one can compute the distance between any two colours. For example, the distance between black and white is about 0.86; the distance between red and white is 0.5, etc.

3.2.2. Partitioning and sequencing

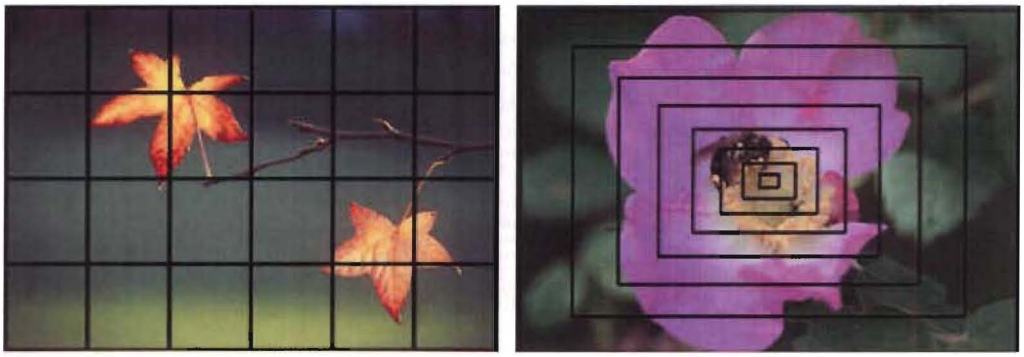
The basic constituting unit of image is pixel. There are many pixels in an im ge. For example, a small 100×100 image has 10,000 pixels. Hence generating pitches at the pixel level is not desired, as too much information is available. More importantly, people generally do not perceive image at the pixel level.

Two approaches are considered to partition an image into blocks and sequence the blocks in order. In the first approach image is divided into equal size blocks, as shown in Figure 9(a). The assumption here is that, since neighbouring blocks are continuous in the image, it is possible that the generated notes also have some kinds of continuity in music.

The sequence of reading the blocks is from left to right, row by row from top to bottom, similar to the conventional image scanning process in computer literature. Though this is not usually how people perceiving image, it is consistent with most people's reading habit. In addition, it is more conforming to the top-down process of eye movement than other possible sequences [Yeo05].

As many photos are taken by active observers who often place the object of interest at the centre, another approach of partitioning and sequencing is to read image from centre to edge, as illustrated in Figure 9(b). This would emulate one of the human attention models when one's sight is attracted by the object at the centre of an image first

and then the remaining part of it. This is particularly useful when we deal with videos when camera movements are used to capture and pursue a moving subject.



(a) (b)

Figure 9. Partition image into (a) equal-size blocks (b) circular blocks. Image (a) is adapted from [Kristie Shureen Photography] by permission.

3.2.3. Mapping colour average

Once the blocks are obtained and ordered, they are mapped to musical notes. The first mapping method simply maps each block's average colour to a single note. The average colour of a block is the average of all its pixels. For example, the block average colours of Figure 9(a) are shown in Figure 11(a).

To test the possible associations of colour and pitch, a block's average hue, saturation and value are mapped to pitch individually. Similar to contour mapping methods, both the Chromatic scale and the Major scale are used. The note duration is proportional to the block *complexity*, which is defined as the number of horizontal and vertical edges in the block. This definition of complexity measures whether the block looks simple or complex and gives a short or long duration respectively.

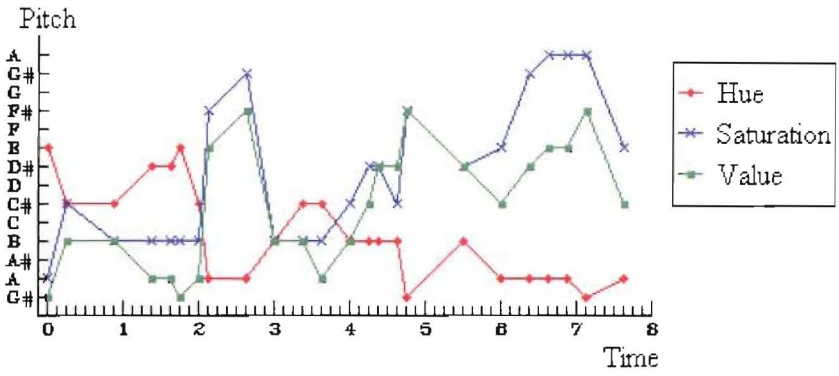


Figure 10. Average hue, saturation and value mappings under the Major scale.

Figure 10 shows the mappings of the 24 blocks in Figure 9(a) to about 8 whole notes. The red plot is the mapping of average hue to pitch; the blue one is average saturation and the green one is average value. In the real implementation, the ranges of hue, saturation and value are (0, 180°), (0, 255), and (0, 255) respectively. The average hue, saturation and value of each block are linearly mapped to two octaves of pitches under the Major scale. As the average colour attributes remove extreme values, the results are usually in a range smaller than two octaves. Blocks with more complex content have longer duration. For example, the first block (the left-most one at the top-most row in Figure 9(a)) has rather uniform colour and occupies two eighth notes (or a quarter note). The second block (at the right of the first one) contains part of a leaf and is more complex. It occupies five eighth notes.

Similar to contour mapping, average colour attributes can give interesting and pleasant results in many cases. When using the Major scale, the generated music is simpler than that when using the Chromatic scale in general. This is because only seven tones are used in case of the Major scale (instead of 12 tones in the Chromatic scale).

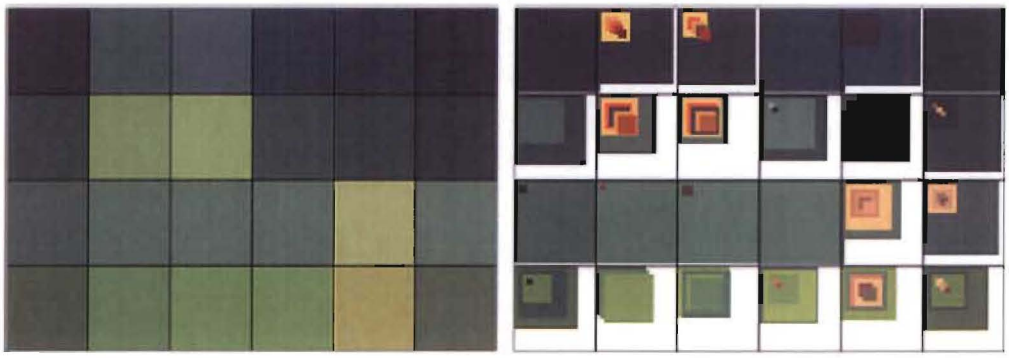
Comparing the three colour attributes, the value mapping generates more intuitive results than the other two. Listening to the generated music with the image in front, one feels easier to observe the luminosity changing and associate it with pitch rising and falling than the other two. Perhaps this is because human has better perception of luminosity than hue and saturation.

3.2.4. Most representative colours

Block average colour only represents a small amount of information of the image. More detailed colour information can be extracted from blocks and be considered in pitch mapping.

Many images, especially those of natural scenes, consist of rich colours. Among the colours, some occur more frequently than others do. Most people can easily figure out the representative colours in an image by observation. For example, one can roughly describe that the image in Figure 9(a) contains red, yellow and a few shades of green colours.

Most representative colours give more information about the colour structure of an image block than the average colour. For example, Figure 11(b) shows the most representative colours of blocks in the image of Figure 9(a). The colours are shows as squares of various sizes to indicate how frequent they occur in the block. Comparing with the average colours in Figure 11(a), the most representative colours give more number of distinctive colours and shades and can better describe the original image.



(a) (b)

Figure 11. Block (a) average colours; (b) most representative colours.

To find out the most representative colours computationally, the HSV colour histogram of an image (or a block) is calculated and searched for possible clusters of colours. The histogram equally divides the HSV colour space into 12 hues, 8 saturation levels and 8 value levels. Overall, there are 12×8×8 bins in the histogram. Every pixel of the image falls into a bin of the matched colour. Each bin counts the number of pixels falling into it and the total number is called the *size* of the bin. In addition, each bin also calculates the average colour of those pixels falling into it to allow more precise colour representation within the bin.

The most representative colours are obtained at the biggest clusters of bins of similar colours. The algorithm of finding bin clusters simply merge bins together if their colour distance (as discussed in *Section 3.2.1.*) is small enough. The pseudo code gives the details of the algorithm, as shown in Figure 12.

```
Arrange all bins in a list
For every bin b in the list
      Find the bin a before b in the list that is closest to b
            If the distance between a and b <= threshold
                  Add b's size to a
                  Update a's colour with b's pixels
                  Remove b from the list
Sort remaining bins in the list in the order of size
Output k biggest bins as the most representative colours
```

Figure 12. Pseudo code for finding clusters in colour histogram.

This algorithm is simple and effective. The use of colour distance as the merging criteria ensures that colours with small perceptual difference can be grouped together. In the actual implementation, the threshold is 0.13. This value comes from some observation. Figure 13 shows the base circle of the HSV cone. If dividing the circle into 18 arcs, as shown in Figure 13(a), the two colours at the two ends of an arc still have noticeable difference. If dividing the circle into 24 arcs, as shown in Figure 13(b), the difference between the two ends is less obvious and considered negligible. This observation suggests that the colour distance threshold is appropriate when it is the length between the two ends of an arc in the latter case. As the radius of the circle is 1, the threshold is about 0.13 in this case.

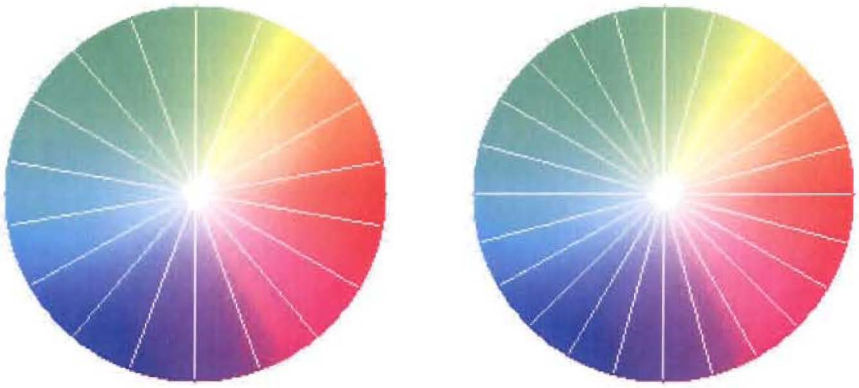


Figure 13. HSV cone base circle divided into (a) 18 divisions; (b) 24 divisions.

3.2.5. Mapping colour confliction

As noticed by painters and many others, opposite colours often suggest a clashing and contradicting feeling in arts [Wells80]. The feeling evoked by opposite colours is generally stronger than that evoked by similar colours or colours of small distance. From this observation, the maximum distance between the most representative colours of a block can be used to measure the degree of feeling evoked by these colours. That distance is defined as the *colour confliction* of the block in the context of this research.

Colour confliction describes how strong the feeling is associated with a block from visual observation. A block with contradicting colours is considered more excited while a block with similar colours is considered calmer. This suggests a mapping of colour confliction to pitch, as higher pitches are considered more excited than lower pitches in general.

Figure 14 shows the mapping of block colours of Figure 11(b) to two octaves of pitches under the Major scale. The results of this mapping are quite intuitive to the user

as it is obvious that blocks with more interesting colours have higher pitches. For example, the second and the third blocks in the first row of Figure 11(b) are mapped to high pitches while the rest blocks in the same row are mapped to low pitches.

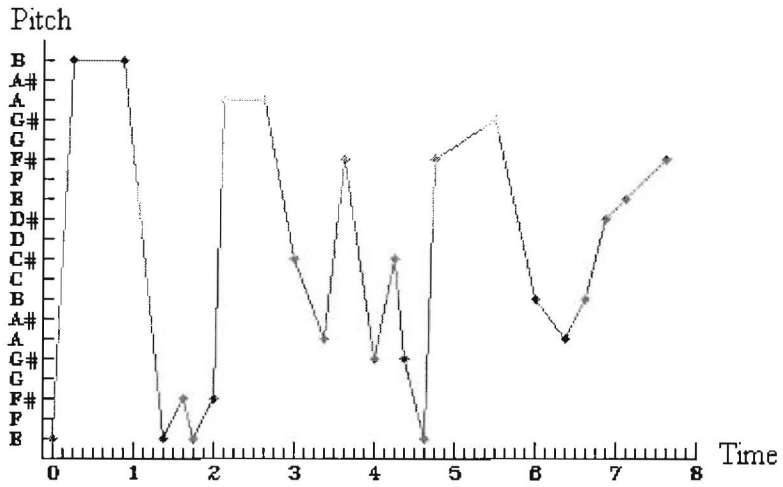


Figure 14. Mapping colour confliction to pitch under the Major scale.

3.2.6. Tonal hierarchy and Global Colour Hierarchy (GCH)

The mapping methods introduced in the previous sections all follow some sort of linear scale. That means all pitches in the target range have similar chance of being chosen. If using the Chromatic scale, all the 12 tones are equally likely to appear in the generated pitches.

However, this does not necessarily conform to the general music convention. Studies in music cognition and perception have shown that tones are not perceived uniformly and are not of equal importance. Instead, they are differentiated into a hierarchical structure, called *tonal hierarchy* [Krumhansl00]. Figure 15 illustrates the tonal hierarchy of the 12 tones in the C Major context. The tonic C sits at the highest level, followed by G and E at the second level, then the rest four tones of the scale (D, F,

A and B) at the third level. The bottom level contains all non-scale tones (C#, D#, G# and A#). Loosely speaking, tonal hierarchy shows the stability and popularity of tones in the specific tonal context. Empirical studies show that tonal hierarchy exists in not only western tonal music, but also in music of other cultures [Krumhansl00]. Though further research is necessary to identify the tonal hierarchies of different kinds of music, it is evident that tonal hierarchy forms the basis of most types of music.

C

E G

D F A B

C# D# F# G# A#

Figure 15. Tonal hierarchy in the C Major context.

The existence of tonal hierarchy suggests that a linear mapping scheme may not suit the nature of music. Instead, a hierarchical mapping may be more desired. For this purpose, a hierarchical structure is created from image and is used as a tonal hierarchy to guide pitch mapping. This hierarchical structure is called Global Colour Hierarchy (GCH) in this study.

Figure 16. Global Colour Hierarchy of the image in Figure 9(a).

GCH is the list of 12 most representative colours of the image, sorted in the descending order of their size. Here size means the number of pixels one colour occupies in the image. The algorithm of computing the most representative colours is described in *Section 3.2.4.* Figure 16 shows the GCH of the image in Figure 9(a). The colours are shown as overlapped squares of their sizes. For example, the biggest square is a dark green colour, the second biggest is a lighter green colour, and the smallest square is a red colour. This GCH shows the perceptually apparent colours in the original image. A few shades of green are from the background area. The yellow, brown and red colours are from the leaves.

3.2.7. Mapping GCH

The 12 colours in GCH are mapped to a tonal hierarchy, for example, the one shown in Figure 15. In this case, the biggest colour maps to C, the second biggest maps to E, and so on. It is possible that some image has less than 12 most representative colours if the image is relatively simple. For example, the GCH in Figure 16 has only 10 colours. If that is the case, the last several tones in the tonal hierarchy are not used.

GCH provides a hierarchical mapping of pitches. A block's most representative colours are compared with the colours in GCH. The closest matches in GCH are used to obtain the corresponding pitches. For example, if a block has three most representative colours, their matches in GCH will give three corresponding tones. In such a way, the generated notes will follow the guidance of the tonal hierarchy such that low-level tones will appear less frequent than high-level ones. This is because colours mapped to the high-level tones appear more frequent in the image and the image blocks.

Typically, many blocks have several most representative colours and the corresponding pitches. If a block can generate more than one pitch, they can be played concurrently or sequentially. If played concurrently, chords are formed. In this case, the results are more pleasant than those containing single notes only. This is because chords provide richer sounds. More importantly, when multiple notes are played together, one's brain tends to choose the most suitable ones automatically to form the main melody. The number of concurrent pitches is restricted to at most three as too many sounds can cause adverse effect.

If playing the notes from the same block sequentially, their order needs to be determined. There are many possible ways of ordering them. For example, the notes can be sorted by pitch, colour brightness, colour size, etc. As many blocks contain several most representative colours, the number of notes created from the whole image may be too large. If this is the case, the generated pitch flow may be too long to show adequate cohesiveness between the parts. Hence, if playing the notes sequentially it is more desirable to divide the image into fewer blocks and/or restrict the number of most representative colours of blocks.

Most of these methods can generate repeated patterns at similar blocks. Two blocks will have high chance of getting the same pitch sequence if they have similar colour composition, which is very likely in simple images. For example, Figure 17(a) shows an enlarged view of the first row of the image in Figure 9(a). The second and the third blocks have similar most representative colours, as shown in Figure 17(b) (the number of most representative colours is restricted to four in this case). Figure 17(c) shows the partial result of sorting block notes in the ascending order of pitch. As

highlighted at the circled positions, the same sequence of pitches appears twice for the two similar blocks. The duration of the notes is proportional to their colour size.

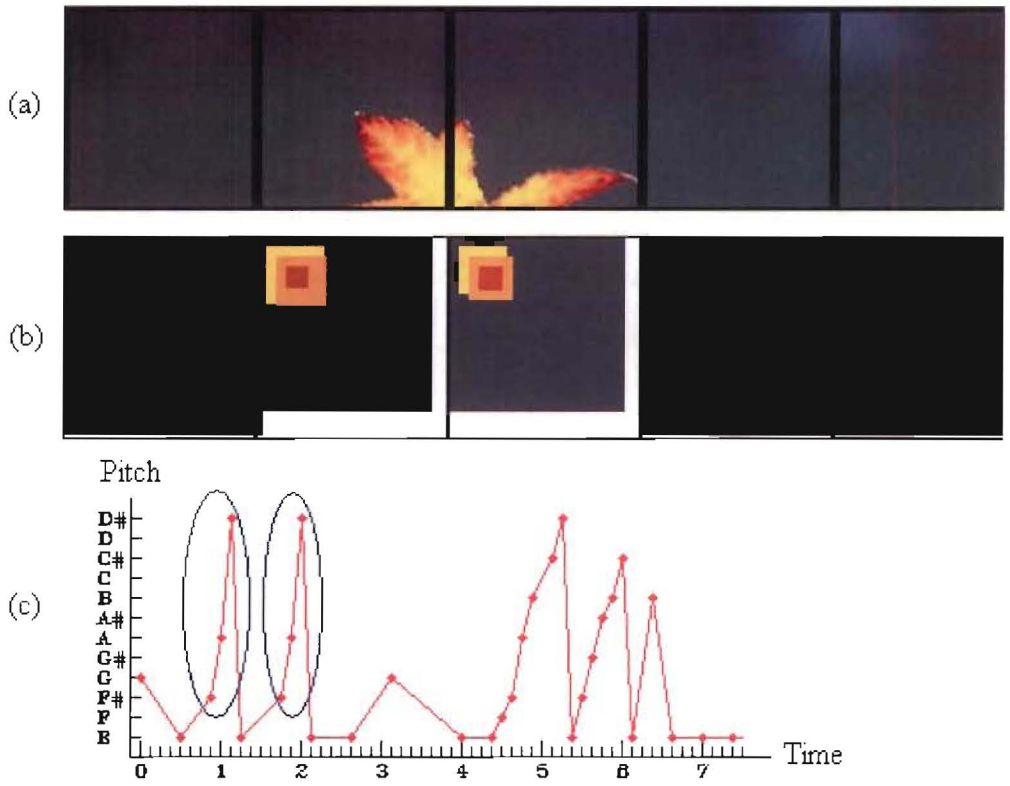


Figure 17. Similar pattern occurs for similar blocks. (a) First row of an image; (b) the most representative colours of the blocks in (a); (c) the generated pitch flow of the blocks. Note the second and the third blocks in (b) are similar and they generate similar patterns in (c).

To comply with the tonal context further, an enhancing method called *melodic anchoring* can be applied to the created notes. Melodic anchoring refers to a phenomenon in tonal music that notes at a lower level of the tonal hierarchy will have an inclination to resolve to a note at a higher level of the hierarchy. For example in the context of the C Major scale, if note B (level 3 in the hierarchy) is heard at the end of a phrase, a strong feeling of incompletion will arise in the listener's mind. However if B is followed by a C note (level 1), the sense of incompletion is effectively resolved. Bharucha suggested that

melodic anchoring is effective if the distance between the two notes is close enough [Bharucha96].

Based on the melodic anchoring principle, an algorithm is used to adjust the order of the block notes locally such that melodic anchoring happens for notes at level three or four of the tonal hierarchy if possible. The results show a better smoothness in the melody in most cases. For example, Figure 18 shows the ending part of two similar pitch sequences created from the same source. The blue sequence is the same as the red one except that it has applied the melodic anchoring principal. The last note of the red sequence is A, which is at level three of the tonal hierarchy. In the blue sequence, the sense of incompletion caused by note A is effectively reduced by the following note G that is at the second level of the tonal hierarchy.

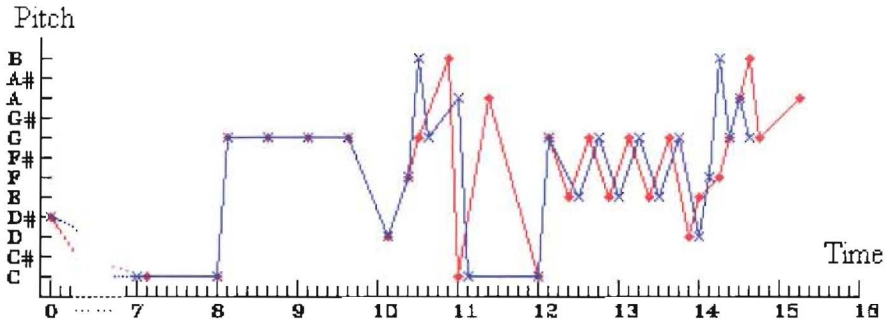


Figure 18. Sorting notes with the melodic anchoring principle.

3.2.8. Forming new tonal hierarchy

As tonal hierarchy exists in most types of music, it is possible to use different tonal hierarchy to create music in different tonal context. The tonal hierarchy used in the previous examples is in the context of the C Major scale. Across different images, the tonal hierarchy is fixed, while the colour to pitch mapping is not. For example, a dark

green colour is mapped to the note C in one image. The same colour could be mapped to the note F# in another image, depending on the position of the colour in the list of the most representative colours in the image. Under the same tonal hierarchy, the colour to pitch mapping totally depends on the sizes of the most representative colours, regardless of the actual colour attributes.

If a fixed colour-to-pitch mapping is used, then the same colour is always mapped to the same pitch. In this case, the tonal hierarchy for each image may be a different one. The colour composition of an image defines its tonal hierarchy and the generated music.

Figure 19 shows a possible fixed colour-to-pitch mapping by choosing 12 reference points in the HSV cone space. The 12 reference points are chosen such that they can be mapped to the 12 tones of the Chromatic scale and are representative colours in natural scene images. The first 6 points represent the six primary colours: red, yellow, green, cyan, blue and magenta. The next three points are black, white, and gray. The last three points are dark red, dark green and dark blue. Only three dark colours are chosen as the perceptual difference is small when the luminosity is low. The 12 reference points are mapped to the 12 tones of the Chromatic scale, as listed at the right of the figure.

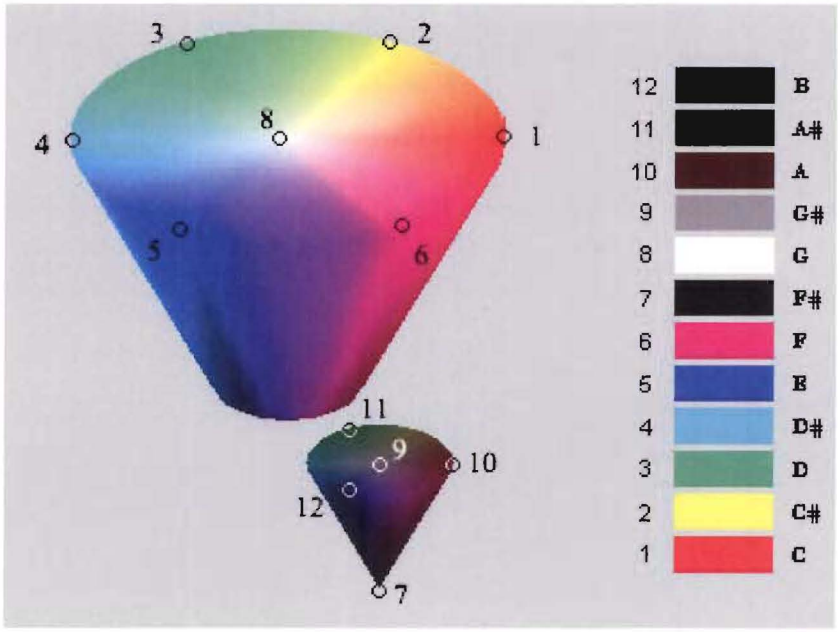


Figure 19. Fixed colour mapping. Twelve colours in the HSV cone are mapped to 12 tones of the Chromatic scale.

Each block's most representative colours are mapped to pitches according to this fixed colour-to-pitch mapping. Using this mapping, tones corresponding to more popular colours in the image will appear more frequently. Consequently, a different tonal hierarchy emerges in the generated music for every different image.

If two images have similar colour compositions, it is likely that they will form similar tonal hierarchies. For example, the two images shown in Figure 20(a) and (b) have similar colours. Figure 20(c) and (d) show the tone frequency distributions of the pitch flows generated from images (a) and (b) respectively. As image (b) is more complex than (a), it has more block representative colours and generates more notes in consequence. However, the generated notes of both images fall into several common categories: C, C#, F#, G#, A and A#. Only one note from image (b) is in a different

category: the G note. Another apparent difference is that the frequency of note A is much bigger in (b) than in (a).

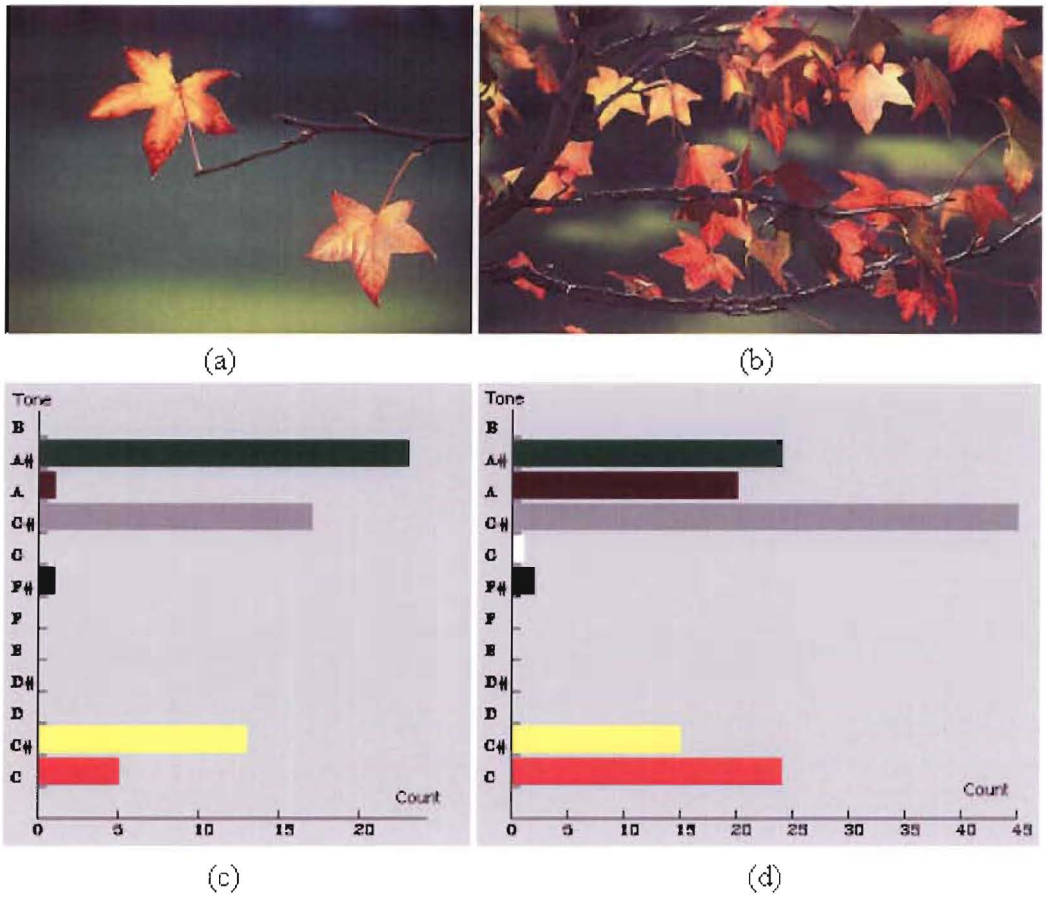


Figure 20. Images with similar colour composition show similar tonal hierarchy. (a) one image; (b) another image with similar colours as (a); (c) tonal distribution of image (a); (d) tonal distribution of image (b). Images (a) and (b) are adapted from [Kristie Shureen Photography] by permission.

Another feature of the results from the two images reinforces the closeness of their tonal hierarchy. When the block colours are forming chords, the most common intervals are 1, 2, 7, and 9 in both images. These intervals are formed from the same notes the two images share. For example, as C, C#, G# and A# are frequent in both images, so are the intervals formed by them.

3.3. Texture

Many images contain different types of textures. Textures can be characterised by different features. Tamura et al. proposed six basic textural features, namely, coarseness, contrast, directionality, line-likeness, regularity, and roughness [Tamura78]. In fact, some of the textural features are already involved in the previous section. Colour confliction can be considered as a measurement of textural contrast. Block complexity represents textural roughness to a certain extent. More textural features, however, are believed to be suitable for determining sound colour (or timbre), as supported by some empirical work [Giannikis06]. Since timbre is not included in the scope of this study, the use of more textural features is postponed to further research.

In an image, the same texture element may occur at many places. If this is the case, they are referred to as structure-based texture. The individual texture elements are called *texels*. For example, an image of autumn scene may contain many similar maple leaves. A pond may have many lotus flowers at various places. A painting may present many stars in exaggerated size and shape over the deep blue sky. The distribution of these texture elements may look random in many cases, but sometimes they are also highly regular. For example, leaves on a trig are sometimes ordered in specific pattern, depending on the type of the tree.

If an image contains structure-based texture, it could be an interesting source for generating musical pattern. To test this idea, an interactive method is implemented in this study. First, the user chooses a texture element (*texel*) of interest by using mouse to highlight the main part of the texel in the image. Second, an algorithm is run to analyze the colour composition of the selected area and to find similar texles in the image. For

example, in an image of many autumn leaves, the user can select one leaf and the algorithm will find other similar leaves. Finally, the obtained texels are mapped to pitches. Taking the image's bottom-left corner as the origin, the horizontal axis represents time and the vertical axis represents pitch. The y coordinate of the centre of each texel is mapped to pitch, while the distance to the next texel (or the right end of the image if the last texel) is mapped to duration. If two texels have very similar x coordinates, their sounds are played concurrently.

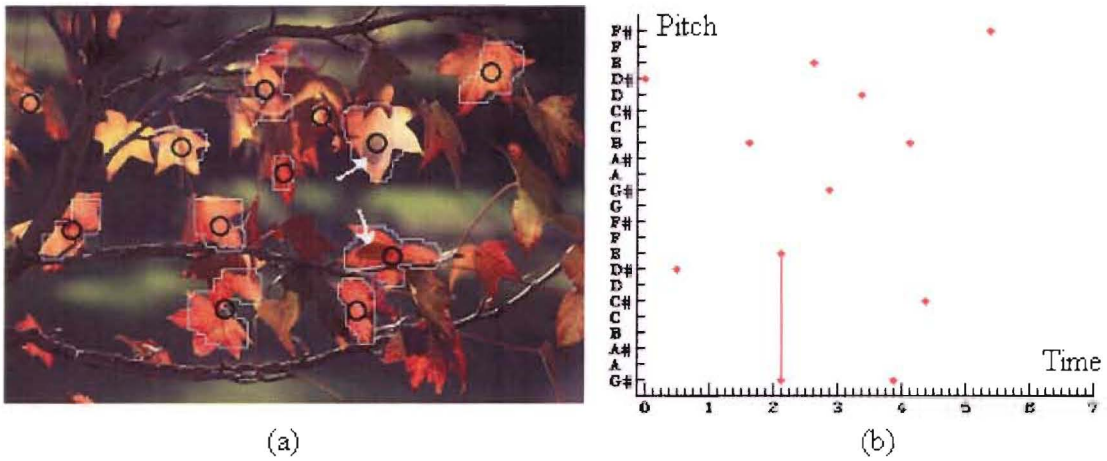


Figure 21. Mapping structure texture to pitch; (a) the original image with texels marked by black circles; (a) the generated pitches of the texels. Image (a) is adapted from [Kristie Shureen Photography] by permission.

Figure 21(a) shows the obtained texels of some red/yellow leaves. The texels are outlined by white polylines and their centres are marked by black circles. The two white arrows point to the initial selection of texel as shadowed area. The initial selection can include parts from different texels to form a wider selection. In this case, parts from two different texels are included to have both red and yellow colours. Figure 21(b) shows the mapping result of the texels. The pitches have similar geometrical distribution as the texels in the original image. As the image height is mapped to three octaves, the pitch

flow has rather big steps. The vertical red line shows the two texels that are sound at the same time.

This method of converting texels to pitches is simple and similar to the approach in several previous studies [Lesbros96][Kabisch05]. If the number of texels is too large, the generated result may contain too many sounds and become unpleasant. If the number of texels is too small (for example only two), it is hard to form any interesting pattern.

The algorithm of finding similar texels makes use of colour histogram intersection as described in [Swain91]. A small square window glides over the whole image. At any time, if the pixels in the window have a similar colour composition with the initial selection, then the pixels are marked to indicate that they belong to some texel. Finally, texels are extracted by performing a connected component analysis over all marked pixels. Comparison of the window pixels and the initial selection is by colour histogram intersection.

The current use of colour histogram as the only criteria to find texels will be less effective in some cases. It cannot work successfully if the colours of the texel are not distinguishable from the background, for example, green leaves over a green background. Moreover, the internal structure of the texels may not be properly reflected in the colour histogram.

CHAPTER 4: INTEGRATION OF MUSIC SEGMENTS

The music segments generated by the methods described in Chapter 3 can be grouped together to create longer and more interesting music. This chapter discusses some methods and issues about integrating them.

When integrating music segments into a more complete piece, some global attributes need to be decided, such as tempo, instrument, key, etc. A simple method is proposed to use the global attributes of image to determine the global attributes of music based on their emotional associations.

4.1. Integrating music segments

There are two possible directions in integrating music segments. One is to arrange the segments horizontally in order to create longer and more structured results. The other is to place the segments vertically in different tracks so that richer and more interesting results can be formed.

4.1.1. Arranging segments horizontally

Repetition is a significant characteristic in music. The same theme may appear many times in a whole piece of music. After hearing the same pitch sequence several times, the listener gets more familiar with it and may expect to hear it again in the remaining part of the music. It becomes more interesting if some variations are added to the same sequence. For example, the same sequence is played at a higher key, in a faster

or slower tempo, by a different instrument, with accompanying chord, adding sharps or flats at certain pitches, etc. As a complete piece of music may contain several themes, intermediate phrases may need to be inserted to connect them.

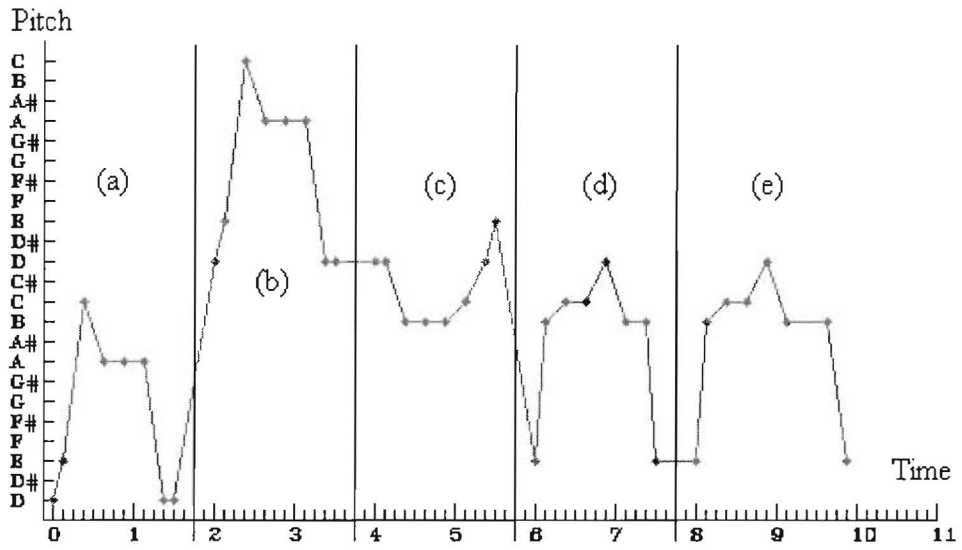


Figure 22. Arrange music segments horizontally. (a) Segment 1; (b) Segment 1 raised by one octave; (c) Segment 2; (d) Segment 3; (e) Segment 3 with slow ending.

Figure 22 shows a melody created from three music segments. All the three segments are generated from the image shown in Figure 9(b). The image is divided into eight columns to create short and simple segments. Segment 1 is created from the average hue method. Segment 2 is created from the average saturation method. Segment 3 is created from the average confliction method.

The whole melody contains five parts. They are created from the three music segments by repeating Segment 1 and Segment 3 once. Segment 1 is first played in part (a) and is repeated in part (b) by raising one octave of pitch. Segment 3 is played in part (d) and is repeated in the last part (e) by slowing down its ending notes to create a closing effect. Segment 2 is played in part (c) to connect them.

This example illustrates that using several simple segments one can created music that is longer and more complex. Repeating segments make the melody more structured and coherent. Variations added to the repeated segments make the result more interesting.

4.1.2. Arranging segments vertically

In the vertical dimension, several tracks can be played at the same time to get richer and more interesting music. Among the methods introduced in Chapter 3, some of them are particularly interesting when vertically combined.

One combination contains the average hue, saturation and value methods, as introduced in *Section 3.2.3.* This combination creates a music segment containing three pitch flows. If played by different instruments, the pitch flows interfere with each other and create an interesting and pleasing effect. One example is shown in Figure 10.

Another combination includes two methods presented in *Section 3.2.7.* One method is the mapping of block most representative colours to pitch via GCH. The other is the same method but with embellishment by melodic anchoring.

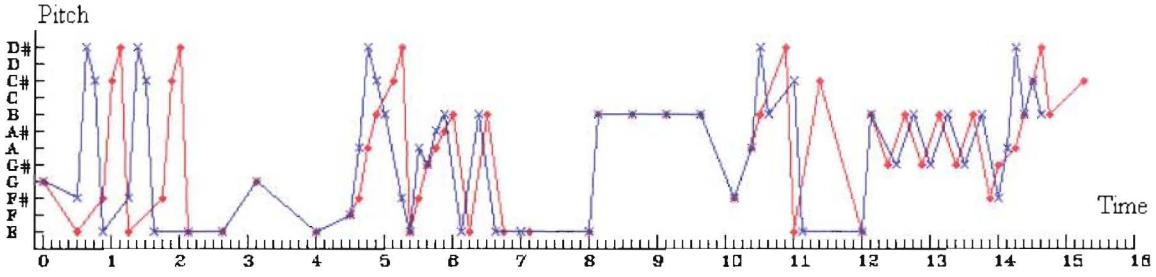


Figure 23. Combining two methods with and without melodic anchoring.

Figure 23 shows an example of this combination. The red track is the original method. The blue track is the method with melodic anchoring. The two tracks are close to

each other in the main direction but different at many local positions. When played by different instruments, the two tracks create many interesting effects at different places, like contradicting to each other, joining with each other, one ahead of the other, etc.

4.2. Associating image and music via emotion

As both image and music arouse emotional response from observer, it is desirable to investigate the possible application of this association to music generation.

4.2.1. Previous studies

Many studies exist in understanding the linkage between music and emotion in various aspects [Collier01] [Makris03] [Sloboda91]. Conventionally, music pieces played in Major modes are considered happier and brighter than when played in Minor modes. Experimental studies support this linkage. Faster tempo is considered more exciting and happier. Ascending pitch directions and higher keys are rated happier and brighter than descending pitches and lower keys in an experimental study [Collier01]. In judging the pleasantness of musical combinations, a simple additive rule is applied to different aspects of music, including timbre, pitch contour, tempo, etc [Makris03].

The linkage between image and emotion is at least as complex as that between music and emotion. Painters usually refer to colours as warm or cold to describe the feeling about them. An empirical study has investigated the effects of colours on emotions, and a model of mapping colour to emotion is presented [Valdez94]. In that study, emotion is modelled by a space of three dimensions: Pleasure, Arousal and Dominance (PAD). Pleasure measures happiness/unhappiness of an emotion. Dominance measures the emotion's passiveness/activeness towards the outside world. Arousal shows

how strong the emotion is. For example, impressed and admired are both considered positive in pleasure and arousal, but impressed is negative while admire is positive in dominance. The experimental results on emotional reactions to hue are relatively weak. However, strong evidence has been shown that saturation and brightness are highly linked to emotion, as summarized by the linear equations below:

$$Pleasure = .69\ Brightness + .22\ Saturation \qquad (1)$$

$$Arousal = -.31\ Brightness + .60\ Saturation \qquad (2)$$

$$Dominance = -.76\ Brightness + .32\ Saturation \qquad (3)$$

4.2.2. Mapping global attributes of image to music

The previous studies have shown the possibility of associating image and music via emotion. Since emotion is generally an overall feeling about the image or the music, the mapping of image to music should also occur at the global level.

A simple method is implemented to map the global attributes of image to music as suggested by their associations with emotion. For simplicity, the average brightness and saturation of an image are mapped to emotion (PAD) using the equations (1), (2) and (3). As arousal represents how strong the emotion is, it is mapped to tempo that indicates how excited the music sounds. Pleasure is mapped to key so higher pleasure corresponds to a higher key in music. The interpretation of dominance is not clear in music, so it is not used in mapping.

Figure 24 shows four images with different average colours. Image (a) has both high brightness and high saturation. Image (b) has high brightness but low saturation.

Image (c) has low brightness but high saturation. Image (d) is low in both brightness and saturation.

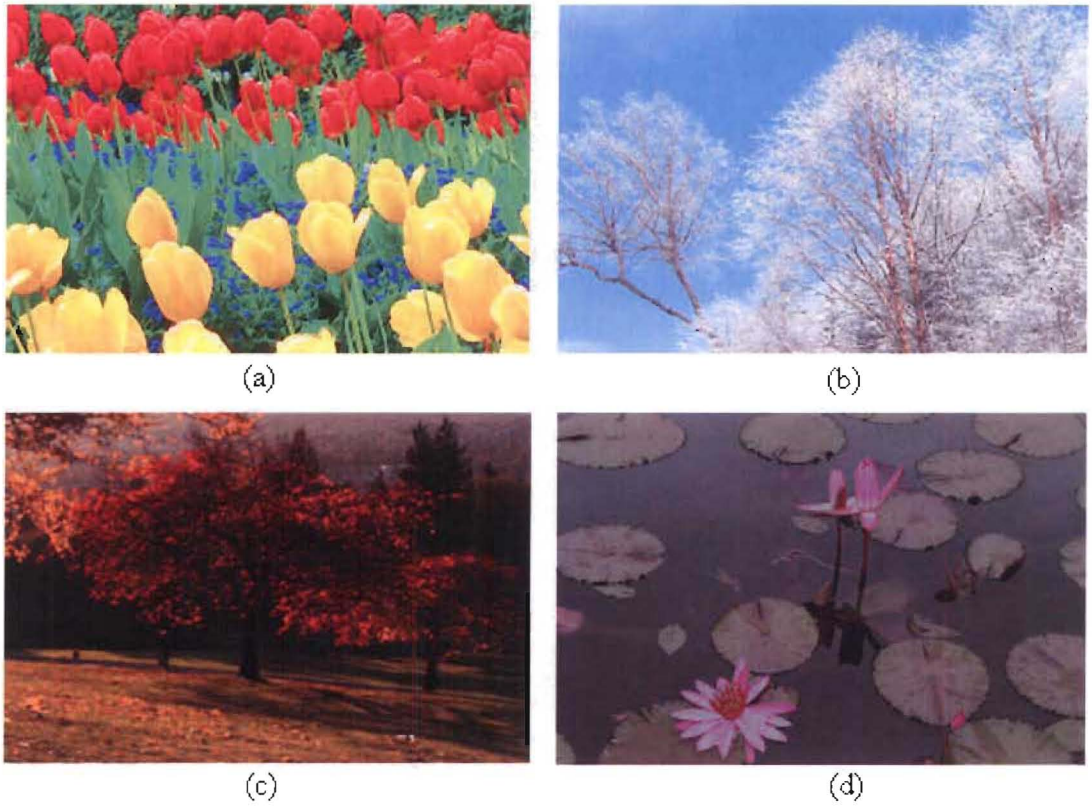


Figure 24. Four images with different average colours: (a) high brightness, high saturation; (b) high brightness, low saturation; (c) low brightness, high saturation; (d) low brightness, low saturation.

Figure 25 shows the results of mapping average colour to key and tempo. For comparison purpose, all four images shown in Figure 24 are applied the same colour confliction method to generate music segments. Image (a) has the highest key since its brightness and saturation are both high. Image (d) has the lowest key since its brightness and saturation are both low. The key of image (b) is slightly higher than that of (c), as brightness has higher contribution than saturation in measuring pleasure. Image (c) gets

the highest tempo as it has low brightness and high saturation. Image (b) gets the lowest tempo, as it is the opposite of (c) (high brightness and low saturation).

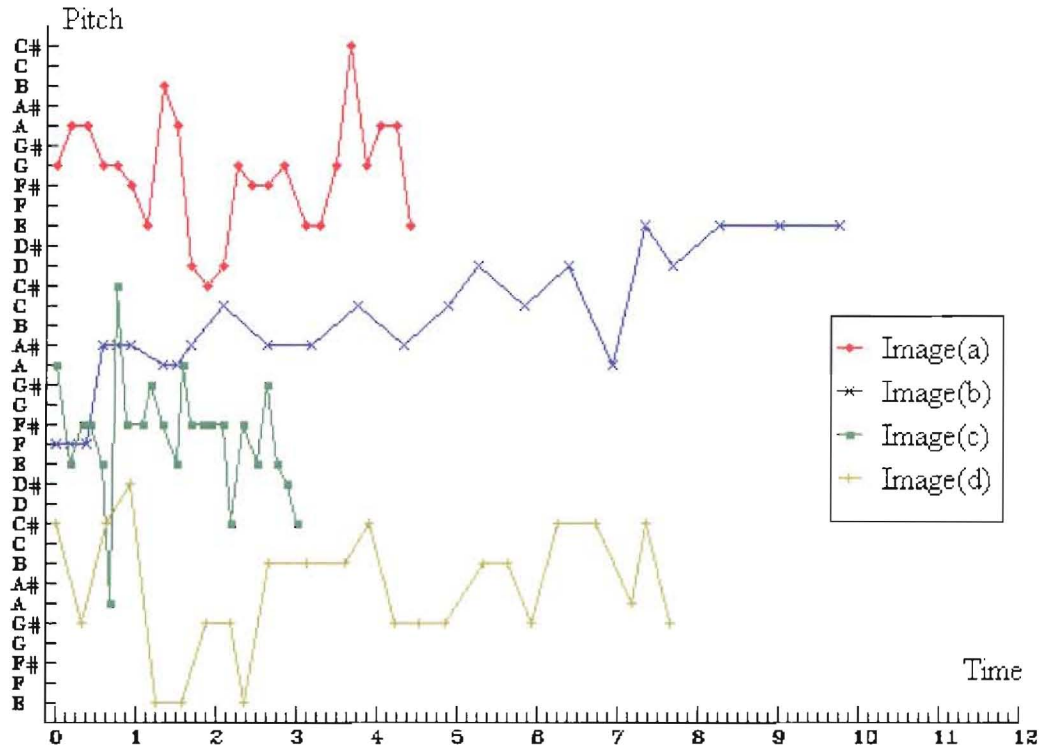


Figure 25. Mapping image average colour to tempo and key.

Using this method, brighter and more colourful images are played at higher keys, which give a brighter feeling in music (e.g. (a)). Darker and more greyish images are played at lower keys and have a more depressing feeling (e.g. (d)). Fast tempo is associated with images of strong colours and dark brightness and sounds exciting and intense (e.g. (c)). Slow tempo is played when the image has bright and soft colours, such that a calm and dreamy feeling is aroused (e.g. (b)).

CHAPTER 5: CONCLUSION AND DISCUSSION

5.1. Conclusion

This study investigates the possibility of applying visual-auditory associations in music generation. It presents and discusses several algorithms of generating music segments from image features, including contour, colour and texture. Object contours are extracted and converted to pitch contours. Image colours are mapped to pitches and chords. Global attributes of image are used to control global attributes of music. Research results in music perception have been applied to enhance the algorithms, e.g. tonal hierarchy and melodic anchoring. Test results show that the algorithms can produce interesting and pleasant music segments in many cases. Combining the generated segments together can create results that are more appealing.

The approaches presented in this study are different from those of previous studies. Image contour is traced along the contour itself instead of horizontally scanning the image. In such a way, contours of the same shape but different orientations can produce similar results. When considering image colour information, the perceptually apparent colours (most representative colours) are used instead of the individual pixel colours. This is because the most representative colours can better describe the user perception of the image than the individual pixel colour.

As the music generation process totally depends on the image content, it avoids the problem of introducing pseudo randomness in many algorithmic music composers.

The same image will always generate the same result under the same algorithm. Moreover, different images with similar features are able to generate similar music patterns. Similar features in the same image are also able to generate similar patterns. Hence, the generated music inherits some of the regularities of the images.

Using image as a source for music generation can create music of different styles for different images. The fixed colour mapping method creates similar tonal hierarchy for images of similar colour compositions. For images of very different colour compositions, it may create very different tonal hierarchies as well. As tonal hierarchy is generally considered the basis of many music styles, the corresponding musical results are of different styles.

Generated music segments can be integrated to form longer and more intriguing results. Currently the way of integrating them and adding variations is essentially the choice of the composer. However, music segments created from the same image but under some different methods can be vertically integrated to get richer and more pleasing results. Using a method of mapping image average colour to tempo and key, the resulted music exhibits some emotional linkage to the original image.

5.2. Further work discussion

This approach for music generation has a high potential and is worth of further investigation. An interesting extension is to apply the algorithms to video frames as video has a natural time line. A possible approach is to synchronize musical notes with content change of frames. Rapid change of content will generate fast flow of notes while slow change of content generates slow flow of notes. Many of the methods developed in this

study can also be applied. For example, a brighter frame generates a higher pitch. The result can be played with the video sequence as background music.

Due to the limited time and resource, the generated music is only subjectively evaluated by the author at this point. A more systematic method could be conducted to better estimate the quality of the results. As it is also of great interest to examine the existence of the perceptual association between the generated music and its original image, some user study may be appropriate to fulfil this purpose. Further work is needed to design an effective user study so that the essential linkage between image and its generated music can be quantitatively evaluated.

The scope of this study does not include timbre. As timbre is an important part of music, another interesting extension may be converting image attributes to timbres. A previous study has investigated the topic on mapping texture to timbre [Giannakis06].

REFERENCES

Barrass, S., "Sculpting a Sound Space with Information Properties", *Organised Sound*, 1(2), pp. 125-136, 1996.

Ben-Tal, O. and Berger, J., "Creative Aspects of Sonification", *Leonardo*, 37(3), pp. 229-233, 2004.

Bharucha, J. J., "Melodic Anchoring", *Music Perception*, 13, pp. 383-400, 1996.

Caivano, J. L., "Colour and Sound: Physical and Psychophysical Relations", *Colour Research and Application*, 19(2), pp. 126-132, 1994.

Collier, W. G., and Hubbard, T. L., "Musical Scales and Evaluations of Happiness and Awkwardness: Effects of Pitch, Direction, and Scale Mode", *American Journal of Psychology*, 114, pp. 355-375, 2001.

Colour Music, Website, http://home.vicnet.net.au/~colmusic/zac.gif

Conklin, D., "Music Generation from Statistical Models", *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, Aberystwyth, Wales, 2003.

Cope, D., *Computer Models of Musical Creativity*, MIT Press, Cambridge, Mass., 2005.

Evans, B., "Foundations of a Visual Music", *Computer Music Journal*, 29(4), pp. 11-24, 2005.

Fels, S. S. and Mase, K., "Iamascope: A graphical musical instrument", *Computers and Graphics*, 2, pp. 277-286, 1999.

Garner, W., "The Relationship between Colour and Music", *Leonardo*, 2, pp. 225-226, 1978.

Giannakis, K. and Smith, M., "Imaging Soundscapes: Identifying Cognitive Associations between Auditory and Visual Dimensions", Godoy, R. I., Jorgensen, H. (eds.): *Musical Imagery*, Swets & Zeitlinger, pp. 161-179, 2001.

Giannakis, K., "A Comparative Evaluation of Auditoryvisual Mappings for Sound Visualisation", *Organised Sound*, 11(3), pp. 297-307, 2006.

Johanson, B. and Poli, R., "GP-music: An Interactive Genetic Programming System for Music Generation with Automated Fitness Raters", *Proceedings of the third International conference on Genetic Programming*, Cambridge, MA, 1998.

Kabisch, E., Kuester, F., and Penny, S., "Sonic Panoramas: Experiments with Interactive Landscape Image Sonification", *Proceedings of the 2005 international Conference on Augmented Tele-Existence*, 157, pp. 156-163, 2005.

Karinthi, P. Y., "A Contribution to Musicalism: An Attempt to Interpret Music in Painting", *Leonardo*, 24(4), pp. 401-405, 1991.

Kristie Shureen Photography, Website, http://www.ksphotography.com.au/index.htm

Krumhansl, C. L., "Rhythm and Pitch in Music Cognition", *Psychological Bulletin*, 126, pp. 159-179, 2000.

Lesbros, V., "From Images to Sounds: A Dual Representation", *Computer Music Journal*, 20(3), pp. 59-69, 1996.

Makris, I., and Mullet, E., "Judging the Pleasantness of Contour-Rhythm-Pitch-Timbre Combinations", *American Journal of Psychology*, 116, pp. 581-611, 2003.

Margounakis, D. and Politis., D., "Converting Images to Music Using Their Colour Properties", *Proc. 12th International Conference on Auditory Display*, London, UK, 2006.

Metasynth, Software, http://www.uisoftware.com/PAGES/index.html.

Moroni, A., Manzolli, J., Zuben, F. V., and Gudwin, R., "Vox Populi: An Interactive Evolutionary System for Algorithmic Music Composition", *Leonardo Music Journal*, 10, pp. 49-54, 2000.

Newton, Isaac, Sir, *Opticks: or a treatise of the reflections, refractions & colours of light*, Dover Publications, New York, 1952.

OpenCV, Open computer vision library, http://sourceforge.net/projects/opencvlibrary/

Papadopoulos, G. and Wiggins, G., "AI Methods for Algorithmic Composition: A Survey, a Critical View and Future Prospects", *Proceedings of the AISB'99 Symposium on Musical Creativity*, Edinburgh, UK, 1999.

Papadopoulos, G. and Wiggins., G., "A Genetic Algorithm for the Generation of Jazz Melodies", *STeP'98*, Finland, 1998.

Pocock-Williams, L., "Toward the Automatic Generation of Visual Music", *Leonardo*, 25(1), pp. 445-452, 1992.

Politis, D., Margounakis., D., and Mokos, K., "Vizualizing the Chromatic Index of Music", *Proc. 4th International Conference on Web Delivering of Music (WEDELMUSIC2004)*, Barcelona, Spain, pp. 102-109, 2004.

Pridmore, R. W., "Music and Color: Relations in the Psychophysical Perspective", *Colour Research and Application*, 17(1), pp. 57-61, 1992.

Robertson, J., Quincey, A., Stapleford, T., and Wiggins, G., "Real-Time Music Generation for a Virtual Environment", *ECAI98 workshop on AI/Alife and Entertainment*, Brighton, England, 1998.

Sloboda, J. A., "Music Structure and Emotional Response: Some Empirical Findings", *Psychology of Music*, 19(2), pp. 110-120, 1991.

Swain, M., and Ballard, D., "Color Indexing", *Int. J. Comput. Vision*, 7(1), pp. 11-32, 1991.

Tamura, H., Mori, S., and Yamawaki, T., "Textural features corresponding to visual perception", *IEEE Transactions on Systems, Man and Cybernetics*, 8, pp. 460-473, 1978.

Valdez, P., and Mehrabian, A., "Effects of Color on Emotions",*Journal of Experimental Psychology: General*, 123, pp. 394–409, 1994.

Viora, J., "Music and Visual Colour: A Proposed Correlation", *Leonardo*, 15(4), pp, 335-336, 1982.

Wells, A., "Music and Visual Colour: A proposed Correlation", *Leonardo*, 13(1), pp. 101-107, 1980.

Xenakis, I., *Formalized Music*, Revised edition, Pendragon Press, 1992.

Yeo, W. S., and Berger, J., "Application of Image Sonification Methods to Music", *Proc. International Computer Music Conference*, Barcelona, Spain, pp. 219-222, 2005.