A STUDY OF IMAGE-BASED MUSIC COMPOSITION

Xiaoying Wu and Ze-Nian Li

School of Computing Science
Simon Fraser University, Burnaby, BC, Canada V5A 1S6
{xwa6, li}@cs.sfu.ca

ABSTRACT

Visual and auditory forms have some noticeable associations that can inspire similar cognitive and aesthetical experiences. This paper presents a study on the possibilities of applying low-level visual-auditory associations to music generation. A few methods are implemented to directly convert visual features extracted from images into musical elements (pitch, duration, and chord). Some initial results suggest a high potential of composing interesting music along this direction[†]. Possible applications of this study include getting new ideas for music composers and generating accompanying music in various contexts.

Index Terms— Algorithmic composition, Image sonification, Image-based music composition, Multimedia application

1. INTRODUCTION

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 cognition and perception. It suggests that, though musical and visual forms are perceived differently, they can sometimes arouse similar cognitive and aesthetic experiences. Inspired on this point, this study tries 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 [2]. To computationally model the creative process of music composition, many algorithms make use of stochastic processes to allow changes and variations. On the contrary, this study focuses on transforming existing features (color, contour, and texture) of images into musical elements. More precisely, this study investigates possible visual-auditory associations and tries to apply them to music composition.

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

Some initial results suggest that this approach has a high potential in creating fresh and interesting music, though the results may not be in complete piece. This is one important advantage of this approach. One immediate application is to help music composers to get new ideas and create new pieces. Another main application is accompanying music generation in various contexts, like electronic greeting card, game environment, interactive multimedia system, etc.

In the next section, some related background studies are reviewed. Section 3 introduces our approach to image-based music generation and gives details of 3 implemented methods. Finally, conclusions and further work discussions are given in Section 4.

2. BACKGROUND REVIEW

2.1. Visual-auditory associations

Perhaps the first scientific attempt of associating color and music is Newton's Color Music Wheel, which maps the 7 prism colors (red to violet) to the 7 tones of a diatonic scale (D to C) based on their wavelength distribution in their respective spectrums. The reason of forming a wheel is based on two perceptual observations: violet meets red seamlessly in human eyes, while two pitches one octave apart are strongly consonant in human ears.

Other mappings of color and sound are mostly similar to Newton's except the ways of separating colors and pitches [1]. One common practice is to separate the color circle into 12 colors such that they can be mapped to the 12 tones of the chromatic scale. One problem that has not been addressed in the early mappings is the fact that there are 9-10 audible octaves but only 1 color circle. Pridmore represents color by a spiral so that colors of the same hue but at different cycles are mapped to pitches of the same tone but in different octaves [7]. Similarly, Giannakis maps color luminosity levels to octaves in a recent study [4].

Besides acknowledging the hue-pitch association, Caivano [1] also correlates luminosity to loudness, saturation to timbre, and size of color to duration of sound. To identify the perceptual associations between color and sound, Giannakis [3] conducted an experimental study with a group of 24 users. His result shows strong evidence in associating high saturation with high loudness.

2.2. Algorithmic composition

Large amount of research effort has been dedicated to algorithmic music composition in the past decades. The main approaches of

[†] *Experimental results are available at http://www.sfu.ca/~xwa6/*

algorithmic composition can be categorized into mathematical models (e.g. stochastic processes and Markov chains), knowledge based systems, grammars, evolutionary methods (e.g. genetic algorithm), systems which learn (e.g. neural network and machine learning), and hybrid systems [6][2].

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 subjectively evaluate each music piece [16]. However, this method lacks of efficiency, especially in evolutionary methods. Instead, objective fitness functions can be used to replace a human evaluator to test the "goodness" of generated music. A fitness function can be based on concepts in musical theory like consonance levels of musical intervals, on statistical models empirically created by induction from a corpus of music genre, on priori training of a neural network, etc [14][15]. It is also common to use both human evaluator and fitness function to compensate the shortcomings of each other, especially in real-time interactive systems [15].

2.3. Image-based music composition

In some computer music systems, graphical representation of sounds is used as a composing tool [5][8]. Lesbros' representation, named *phonogram*, uses the vertical axis to represent the uniformly scaled pitches and the horizontal axis to represent time [5]. A composer creates music by drawing patterns on paper, which is scanned and processed to produce sounds. Though phonogram is a physical representation of sound, it does give composers some perceptual ideas of interpreting image 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.

Image sonification for musical purpose has gained some research attention in recent years [9][11][13]. As image contains 2D planer data, the first problem of converting image to music is to map 2D planar data to the time axis. Two concepts of time mapping (scanning and probing) are proposed to address this problem [13]. Landscape contours (e.g. city skylines) are extracted as pitch contours to create music in a landscape image sonification system [11]. In an attempt to synthesize appropriate music to video based on the video content, frame hue is mapped to pitch and brightness to loudness [9]. The generated pitch profile is matched with a selected example to create new music.

3. COMPOSITION FROM IMAGE FEATURES

In our approach, the music generation process consists of 3 basic steps. *Partitioning* is the first step that separates image into individual pieces. The *sequencing* step decides the sequence of these pieces along the time axis. Finally, each piece is converted into one or more musical notes in the *mapping* step. Following this process, various image features are extracted and converted into music, including contour, color and texture. They are described in this section.

3.1. Contour

Section 2.3 has shown some examples of using contours in music composition and image sonification [5][11]. As many images of natural scenes contain interesting contours, a similar method is implemented to use them to guide pitch contours.

Object contours are extracted from the image and approximated by straight line segments. Reading the line segments along the contour in one direction, each line segment can be mapped to a musical note. For testing purpose, the portion of the contour and the direction of reading it can be interactively selected in this method.

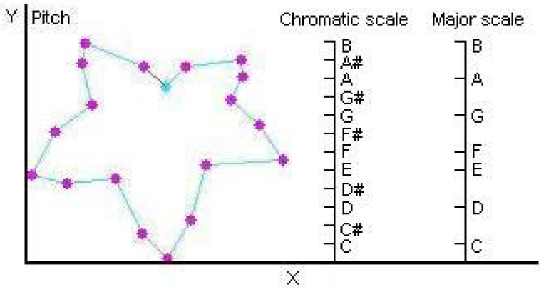

Figure 1: Contour of a leave. If using the first mapping, the y coordinates of the points are mapped to the chromatic scale or the Major scale.

Fig. 1 shows the contour of a leave in a coordinate system with X axis at the bottom-most point and Y axis at the left-most point of the contour. Three mapping are tested. The first one maps the y coordinate of each line segment's ending point to pitch. The second one maps the line segment's slope (-180°~180°) to pitch. The third uses derivative (slope change). In all three mappings, duration is proportional to the line segment's length. So a longer line segment produces a longer note. Two pitch scales are used in mapping. The first is the uniform chromatic scale of 12 tones. The second is the non-uniform Major scale of 7 tones.

All three mappings generate interesting results in some cases. Using the Major scale, however, gives smoother results than the chromatic scale in most cases. This is true for people who are more familiar with the Major scale than the chromatic scale

Comparing the three mappings, the first one is the most intuitive, as people can directly see the rising and falling of pitch from object contour. The third mapping is also easy to interpret as it produces high pitches at sharp corners and low pitches at obtuse ones. It gives less fluctuation for smooth contours. In an extreme case, a circle-like contour will give almost the same pitch. The second mapping is less intuitive as the slope of each line segment is not easy to differentiate.

3.2. Color

Besides contour, color is another important attribute to consider in image-based music generation. In this study, the HSV color model is used as it better describes human perception of color than the RGB model.

3.2.1. Partitioning and sequencing

Two ways of partitioning and sequencing are used. First, image is divided into equal size blocks, as shown in Fig. 2(a). The assumption here is that, since neighboring blocks are continuous in the image, the generated notes are expected to have similar continuity in music as well.

The sequence of reading the blocks is from left to right, top to bottom. This is similar to the conventional image scanning process found in computer literature. Though this is not usually the way of perceiving images, 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 [13].

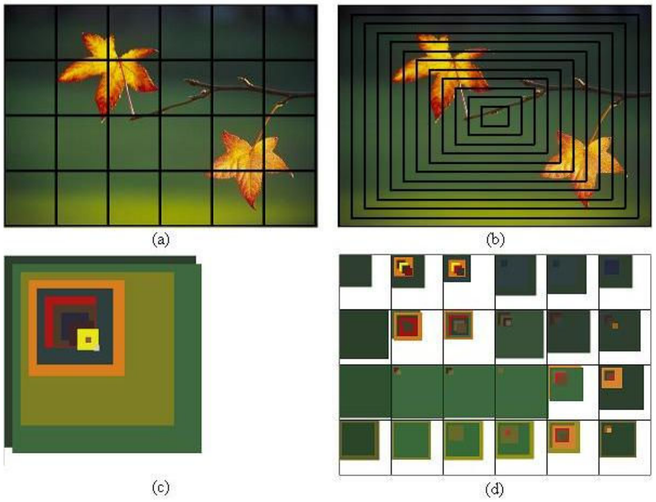

Figure 2: (a) Divide image into equal-size blocks; (b) Centric partitioning; (c) Most representative colors of the image are shown as overlapped squares of their specific sizes; (d) Most representative colors in blocks of (a).

Another way of partitioning and sequencing is to read image from center to edge, as illustrated in Fig. 2(b). This seems more consistent with human cognitive process as we normally "see" the center of an image first and then the remaining part of it.

3.2.2. Mapping

Once the blocks are obtained, they are mapped to one or more musical notes. To test the possible associations of color and pitch, a block's average hue, saturation and luminosity are mapped to pitch respectively. Similar to contour mappings, both the uniform chromatic scale and the non-uniform 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.

Similar to contour mapping, average color attributes can give pleasant results in many cases, especially when using the Major scale. In this case, the generated music uses only 7 tones of the Major scale. That causes the results simpler and easier to listen to, than when using all 12 tones of the chromatic scale.

Comparing the three attributes, the luminosity mapping generates more intuitive results than the other two. Listening to the generated music with the image in front, it is easier to observe the brightness 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.

3.2.3. Tonal hierarchy

It has been noticed that using uniform chromatic scale in mapping does not necessarily give pleasant results. This is because tones are not perceived uniformly and are not of equal importance. Instead, they are differentiated into a hierarchical structure, called *tonal hierarchy* [12]. Fig. 3 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 4 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 [12]. Though further research is needed to fully discover the tonal hierarchies of different kinds of music, it is evidently clear that tonal hierarchy forms the basis of most music.

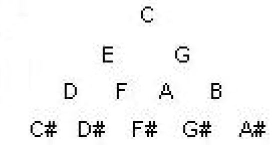

Figure 3: Tonal hierarchy in C Major context. From top to bottom, the tones are considered more and more unstable.

The existence of tonal hierarchy suggests that a proportional mapping scheme may not suit the nature of music. If we can get some hierarchical structure from image, this structure can be used as a tonal hierarchy to guide music generation. In our attempt this hierarchical structure is called Global Color Hierarchy (GCH).

GCH is the list of 12 most representative colors of the image, sorted in the descending order of their size. Here size means the number of pixels one color occupies in the image. The most representative colors are obtained by analyzing the image color histogram to find the biggest clusters. For example, the GCH of image in Fig. 2(a) is shown in Fig. 2(c). The 12 colors in GCH are mapped to the tonal hierarchy described in Fig. 3, such that the biggest color maps to C, the second biggest maps to E, and so on. When a block is read, its most representative colors are computed and mapped to the closed colors in the GCH. For example, if a block has 3 most representative colors, their matches in GCH will give 3 corresponding tones. In such a way, the generated notes will follow the guidance of the tonal hierarchy such that lower level tones will appear less frequent than those high level ones. Fig. 2(d) shows the most representative colors of the blocks in Fig. 2(a).

As each block could generate more than one note, they can be played concurrently or sequentially. If played concurrently, chords can be formed. In this case the results are more pleasant than those of single notes, because of richer sounds. More importantly, when multiple notes are sound together our brains can automatically choose the most suitable one to form the main melody.

3.2.4. Melodic anchoring

If playing the notes in the same block sequentially, the playing order remains a question. Several possibilities exist. One method, called *melodic anchoring* is used to sort the notes in proper order.

Melodic anchoring is the music phenomenon that notes at a lower level of the tonal hierarchy will have a high inclination to resolve to a note at a higher level in the tonal hierarchy. For example in the context of the C Major scale, if the 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 proposed the principle of proximity in melodic anchoring: the distance between the two notes should be as close as possible [10].

Based on the melodic anchoring principle, an algorithm is used to sort the block notes such that melodic anchoring happens for each note at level 3 or 4 if possible. The duration of each note is proportional to the corresponding color size. The results show a better smoothness in the melody than a random sequence.

3.2.5. Some issues

Block size affects the generated music largely. Smaller block size leads to more notes. However, the level of musical change does not necessarily increase. When the size goes down, adjacent blocks are getting more indistinguishable from each other, especially at areas with large uniform color. To solve this problem, a non-uniform grid is implemented such that areas of more content are divided into more blocks.

The tonal hierarchy used previously is in the context of the Major scale. As tonal hierarchy exists in many kinds of music, it is possible to imitate certain music style by using its particular tonal hierarchy. More study is required to test this assumption.

Further more, if a new tonal hierarchy is defined, a very different kind of music could be generated. To test this point, an image-independent color-pitch mapping is tried. We divide the HSV space into 24 colors and map them to 24 pitches in 2 octaves. Each block's most representative colors are matched to pitches according to this fixed color-pitch mapping. Using this method, tones corresponding to more popular colors in the image will appear more frequently. Consequently, a different tonal hierarchy emerges in the generated music for every different image.

Besides color analysis, histogram is also tried to generate music directly by sounding 2D histogram (e.g. hue on y axis and luminosity on x axis) with pitch on y axis and time on x axis. However, this method does not work well in general as the mappings have no apparent perceptual meaning.

3.3. Texture

To explore the usage of structure-based texture in music generation, an interactive method is implemented in this study. First, the user selects the main area within a texture element (*texel*) of interest from the image. Second, an algorithm is run to analyze the color 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 leave and the algorithm will find other similar leaves. Finally, the obtained texels are mapped to musical notes in a way similar to the first contour mapping. Using the X axis as the time line, the y coordinate of the center of each texel is mapped to pitch, while the size of the texel is mapped to duration. If two texels have very similar x coordinates, they will be sound together.

As this method is similar to the first contour mapping, the generated results and analysis are also similar. An interesting point of this method is that different timbres can be used for different textures. This remains a topic in the further study.

4. CONCLUSION AND DISCUSSION

This paper introduces several methods of generating music from image features: contour, color and texture. Test results show that interesting and pleasant music segments can be produced in many cases. Though the quality of the generated music is depending on the image used, this music composition approach has a high potential and is worth of further investigation.

Further works include investigating the use of different tonal hierarchies in style creation. As the results obtained at this stage are only music segments, it is desirable to integrate them to create more complete pieces. As both image and music are able to arouse emotional response, it is also under investigation to apply emotional models. Another interesting extension is to apply these methods to video frames as video has a natural time line.

Evaluation of the generated music is mainly by the authors' observation at this point. A more systematic evaluation should be conducted to estimate the quality of generated music and more importantly, whether there is a strong association between the image and its generated music.

5. REFERENCES

[1] J. L. Caivano, “Colour and Sound: Physical and Psychophysical Relations”, *Colour Research and Application*, 19, 2, pp. 126-132, 1994.

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

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

[4] K. Giannakis, “A Comparative Evaluation of Auditoryvisual Mappings for Sound Visualisation”, *Organised Sound*, 11, 3, pp. 297-307, 2006.

[5] V. Lesbros, “From Images to Sounds: A Dual Representation”, *Computer Music Journal*, 20, 3, pp. 59-69, 1996.

[6] G. Papadopoulos and G. Wiggins., “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.

[7] R. W. Pridmore, “Music and Color: Relations in the Psychophysical Perspective”, *Colour Research and Application*, 17, 1, pp. 57-61, 1992.

[8] Metasynth, http://www.uisoftware.com/PAGES/index.html, software.

[9] M. Nayak, S. H. Srinivasan, M. S. Kankanhalli, “Music synthesis for home videos: an analogy based approach”, ICICS-PCM, 3, pp. 1556-1560, 2003.

[10] J. J. Bharucha, “Melodic anchoring”, *Music Perception*, 13, pp. 383-400, 1996.

[11] E. Kabisch, F. Kuester, and S. Penny, “Sonic panoramas: experiments with interactive landscape image sonification”, *Proceedings of the 2005 international Conference on Augmented Tele-Existence*, 157, pp. 156-163, 2005.

[12] C. L. Krumhansl, “Rhythm and pitch in music cognition”, *Psychological Bulletin*, 126, pp. 159-179, 2000.

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

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

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

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