Generating Music from an Image

Conference Paper · October 2015

CITATIONS
0

READS
663

4 authors:

Gwenaelle Cunha Sergio
Kyungpook National University
10 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Rammohan Mallipeddi
Kyungpook National University
116 PUBLICATIONS **3,538** CITATIONS
SEE PROFILE

Jun-Su Kang
Kyungpook National University
15 PUBLICATIONS **56** CITATIONS
SEE PROFILE

Minho Lee
Kyungpook National University
333 PUBLICATIONS **2,444** CITATIONS
SEE PROFILE

Generating Music from an Image

Gwenaelle C. Sergio, Rammohan Mallipeddi, Jun-Su Kang, Minho Lee*
Kyungpook National University, Daegu, South Korea
gwena.cs@gmail.com, mallipeddi.ram@gmail.com, wkjuns@gmail.com, mholee@gmail.com

ABSTRACT
Images can convey emotion just like music. If that's so, then it might be possible that, given an image, one can obtain a music that can produce a similar reaction from the listener/viewer. The challenge lies in how to do that. In this paper, we analyze the image using the HSV color space model and assume that each one of the three components have a relation with basic music elements, like tone, pitch, rhythm and loudness. The image is then scanned from left to right and top to bottom in order to generate a sequence of notes. In the end, the emotional Mean Opinion Score (MOS) is used to evaluate the performance of the proposed method. This work could prove to be a very important contribution to the field of HCI because it can improve the interaction between computers and humans who are visually and/or hearing impaired. In the current work, we only consider two emotions; positive and negative.

Author Keywords
Music Elements; HSV Color Space Model; Emotion; MOS; GAPED

ACM Classification Keywords
G.4, H.5.5, I.4.8, J.5

INTRODUCTION

The need to communicate is intrinsic in any living organism, and as such, humans have always felt the need to communicate with each other. In fact, spoken language was originally just a sequence of sounds; and after that, when we felt a greater need to register history, written language started as ancient cave paintings.

That means that art, music and paintings, have always been present in society, and they're essential to any human interaction. If we can find a direct relation between those two arts, that would open a whole new range of possibilities. For starters, it would give an opportunity for visually and/or hearing impaired people to appreciate the field of arts they are unable to. In the engineering field, if we want computers to be able to effectively interact with humans, it is imperative to teach them arts.

There are currently some applications that generate music given an image, such as Photosounder [11], SonicPhoto [13] and Paint2Sound [12]. Photosounder is the "first audio editor/synthesizer to have an entirely image-based approach to sound creation and editing" and it can transform sounds into image as well as transform image into sounds. The second and third applications only transform pictures to sound, and not the other way around. SonicPhoto was inspired by the first application mentioned above. It doesn't have all the features as Photosounder but the creator claims that it has an automatic and convincing stereo and "a unique harmony filter to help create distinct and professional effects". In this application, time indexing increases from left to right and the pitch of the sound increases from bottom to top. The last application, Paint2Sound, synthesizes sine waves from each pixel row assuming that each color of the image pixel represents a frequency band and the brightness of the pixels represent the amplitude of the sound.

A common limitation between the applications mentioned above is that they do not consider emotion when generating the music. Thus, the goal of this paper is to implement an algorithm that takes into consideration that important aspect of music and images.

To do that, we will try to find a relationship between the HSV components of an image and music components such as tone, pitch, rhythm and loudness and analyze the performance of the proposed method using Mean Opinion Score (MOS).

This paper is organized in 5 sections. The first section is the introduction. The second one introduces the relationship between emotion and music and picture. The next section is the proposed method, and introduction of the relationship between music and science and the basic and major elements of music, the sciences of music and images, the implemented algorithm and the evaluation method. The fourth section introduces the results, and the last section is the conclusion.

EMOTION VS MUSIC AND PICTURE

Emotion Definition

Emotion is defined by the Oxford Dictionary [10] as "*a strong feeling deriving from ones circumstances, mood, or relationships with others*"; and by Dictionary.com [4] as "*something that causes such a reaction: the powerful emotion of a great symphony*".

Both definitions converge to one point: emotion is a feeling caused by something or someone else, either a person, a situation, a music, or a painting.

Emotion from Picture

Clive Bell [9], an English art critic, associated with the Bloomsbury Group (influential group of English writers, philosophers, intellectuals and artists, the best known members of which included writer Virginia Woolf and economist

HAI 2015, October 21 - 24, 2015, Daegu, Kyungpook, Republic of Korea

ACM ISBN 978-1-4503-3527-0/15/10 ...$15.00.
http://dx.doi.org/10.1145/2814940.2814978

John Maynard Keynes), said in his book [2] that "all sensitive people agree that there is a peculiar emotion provoked by works of art". He also defended that what makes a work of art is the way that lines and colors combined in a particular way have the ability to stir our aesthetic emotions. That quality is what he calls "Significant Form".

A few works have been done in the engineering field regarding the extraction of emotion from visual stimuli. Zhang [15, 16] aims to embed emotion into a machine that can analyze images and learn more complex emotions by interacting with humans by analyzing EEG signals and image features.

In [14], the authors extract texture information from images in the IAPS dataset, using Wiccest and Gabor features and uses those to train a SVM network to classify between different emotional valences. After the training, a collection of masterpieces is applied to the system, obtaining a performance a little higher than 50%, and showing that machines have a potential to derive emotion from paintings.

Emotion from Music

Daniel J. Levitin [8], an American award-winning neuroscientist, cognitive psychologist, musician, professor, author and record producer, wrote in his first best-selling book [7] that "the essence of music performance is being able to convey emotion" and that's the reason why most of us listen to music, for that emotional experience.

He emphasizes how music influences us by mentioning various situations where music has been used to harness emotions, such as advertisers attracting costumers, mothers soothing their babies and movie directors telling us how to feel in ambiguous scenes.

He also mentions a few more scientific aspects of music: in western music, major scales are associated with happy emotions and minor scales with sad ones; a fast tempo tend to be regarded as happy, whilst a slow one is regarded as sad; alert sounds are abrupt, short, loud sounds, whereas slow, long and quieter sounds tend to be seen as calming or neutral.

In the scientific field, Kim [5] extracts emotion indicators given a video, that is, given musical and visual components. It extracts Tempo, Melody and Loudness from the musical component and Hue, Saturation, Intensity and Orientation from the visual component. It also uses EEG signals as the Valence and Arousal indicators, and, in the end, it uses ANFIS to classify the emotion as positive or negative.

Lee [6] developed a system based on 3D fuzzy visual and EEG features that recognizes a person's emotional state while watching a movie clip (visual stimuli and EEG signals).

PROPOSED METHOD: PICTURE TO MUSIC

Science of Music

Sound is the way one perceives vibrations in the air around them, and it can be characterized by the following equation.

$$P = Asin(2\pi ft) \tag{1}$$

Where P is the pressure in Decibels or Pascal, A is the amplitude (loudness or volume of sound), f is the pitch in Hertz and t is the time in seconds.

Humans cannot hear above 22 KHz. So, according to the Nyquist Theorem below, it is safe to assume that an appropriate sampling frequency f_s for a music is 44 KHz. Usually, audio codecs use 44.1 KHz, so well be using that in this work.

$$f_s \geq 2f_1 \tag{2}$$

HSV Color Space Model

HSV stands for Hue, Saturation and Value, see Figure 1. According to [1], Hue is "the most obvious characteristic of a color", and it's divided into 6 main groups: red, yellow, green, cyan, blue and magenta. The Hue of a color is the angle from the initial point, red, to the point in the projection that represents the color in evidence, so it's range is from 0^o to 360^o.

Saturation represents the purity of a color, and ranges from 0 to 1. The more saturated a color is, the more rich it looks, and the less saturated it is, the more grayish it looks.

Value, also called luminosity or brightness, represents how light or how dark a color is. The color black appears when $V = 0$ and white when $V = 1$. So, the higher the value, the lighter a color is, and the lower the value, the darker it is.

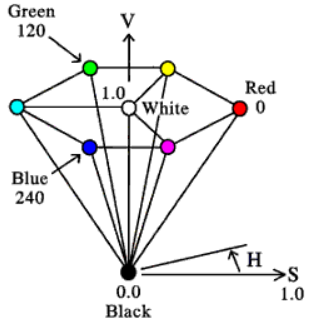

Figure 1: HSV - Pyramid Model

Even thought this model also requires three components to represent a color, like the RGB Model, it's much easier to agree on what color it is, independently of its purity and/or brightness.

Music and Color

As discussed previously, an image can convey emotions just like music. If that's so, the challenge is in how to obtain a music that can produce a similar reaction from a subject.

One of the main aspect of an image is the colors that compose it, so the Hue values will be used to decide which notes should be played at a given time. We will be assuming that the scope is western music. Dark and blue shades of color usually convey a sad feeling, and, as mentioned in the previous subsections, minor keys convey sad feelings in music. Also, chords that are usually used to represent sad tones are C and D. Light colors on the other hand convey happy feelings, and so do major keys.

We'll also be considering that the duration t of the note, or rhythm, is related to the Saturation; and the loudness or amplitude A of the note, is related to the Value. Remember that these variables are related to Equation 1.

Now, the frequency belonging to each note was obtained according to the harmonics in a piano and this relation and all of the above are grouped together in Table 1.

ID	Color	Note	Harmonic (Hz)	Key	Emotion
1	black	$C\#$	34.648	minor	sad
2	grey	D	36.708	minor	sad
3	yellow	E	41.203	major	happy
4	green	F	43.654	major	happy
5	cyan	G	48.999	major	happy
6	blue	C	32.703	minor	sad
7	magenta	A	27.5	major	happy
8	red	$A\#$	29.135	major	happy
9	white	B	30.863	major	happy

Table 1: Relationship between Colors and Music

Algorithm

1. Transform image from the RGB to HSV model
2. Calculate the mask size to scan image
   $M_s = floor\left(\sqrt{\frac{image_{dimension}}{music_{duration}}}\right)$
3. For each patch of image of size M_s x M_s
   (a) Decide whats the dominant color
4. In the end, a color map of the image is obtained
5. From this color map, obtain a notes map
6. Use the Formula 1, changing f according to their respective notes.
7. Music is obtained.

Evaluation Method

The emotional Mean Opinion Score (MOS) is used to evaluate the performance of the proposed method. For simplicity, the emotion will be classified into positive or negative according to the emotion axis in Figure 2. The subjects will be shown a music or an image and they will be asked to choose a number between 1 and 9, where 1 represents a very positive feeling and 9 represents a very negative one.

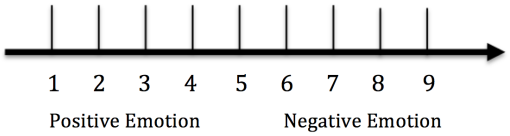

Figure 2: Emotion axis

EXPERIMENTAL RESULTS

GAPED

The Geneva Affective PicturE Database (GAPED) [3] consists of 730 pictures and it was "created to increase the availability of visual emotion stimuli". Eight images were used from the dataset: 4 representing negative emotions (Figure 3) and 4 representing positive emotions (Figure 4).

(a) Spider (Sp149) (b) Spider (Sp160)

(c) Skulls (H081) (d) Snake (Sn103)

Figure 3: Images representing negative emotions

(a) Newborns (P005) (b) Seal (P079)

(c) Puppies (P114) (d) Christmas Deco (P058)

Figure 4: Images representing positive emotions

MOS

The selected images and their respective musics were presented separately to 9 people and the emotional Mean Opinion Score from each image and music were plotted in the graph in Figure 5. The graph is plotted Images (1-8) vs emotional MOS, and the images are mixed up before presenting to the participant, that is, we don't present first the negative images and then the positive ones, they are presented randomly as following: Sp149, P005, Sp160, H081, P058, P079, Sn103, P114 (check Figures 3 and 4 for the labels).

The correlation between the two curves was high, 0.86, because they follow a similar pattern. However, the images produced more extreme results, where the participants really liked an image or really hated. As for the musics, they varied less, meaning that even though this work was successful

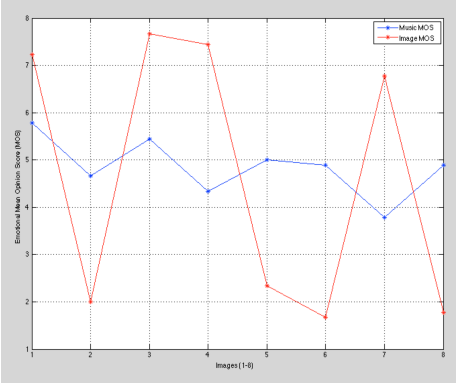

Figure 5: MOS Graph: Images (1-8) vs emotional MOS

in obtaining high correlated curves, it still has to find methods to improve the difference between extremely positive and negative emotions.

CONCLUSION

It can be concluded that this work successfully established a relation between image and music by relating the Hue, Saturation and Value of an image to the tone, pitch, rhythm and loudness of a sound, as can be seen with the MOS evaluation. However, this work still has a lot to improve in the future.

Future Works

In the present work, we scan the image from left to right, top to bottom. However, we believe that a better way to do that would be to scan the image like a person would, considering the human eye movements, instead of just raster scanning the picture, and also considering EEG signals. Another possible addition to the method would be to separate the background from the foreground before the transformation.

Another work that should be done in the future is to consider a wider range of emotions, which will also require a more profound study of music theory. In parallel to that, higher-order music concepts should also be considered when generating music from an image. Also, more research should be done regarding methods to solve the problem of the drop that occurs in the music when changing from one note to another, resulting in various sharp sounds throughout the music.

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning, under grant number 2013R1A2A2A01068687.

REFERENCES

1. Adamson, J. C. Hue, saturation & value : The characteristics of color, 2012. Retrieved May 13, 2015 from The Muser Physics & Physiology of Color, `http://www.greatreality.com/color/ColorHVC.htm`.
2. Bell, C. *Art*. New York Frederick A. Stokes Company Publishers, 1913.
3. Dan-Glauser, E. S., and Scherer, K. R. The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods 43*, 2 (2011), 468–477. Downloaded May 21, 2015, `http://www.affective-sciences.org/system/files/webpage/GAPED_2.zip`.
4. Dictionary.com, 2015. Retrieved May 20, 2015, `http://dictionary.reference.com/browse/emotion`.
5. Kim, T. Emotion classification by acoustic, visual and eeg signals using fuzzy clustering and anfis, 2014.
6. Lee, G., Kwon, M., Kavuri, S., and Lee, M. Emotion recognition based on 3d fuzzy visual and eeg features in movie clips. *Elsevier: Neurocomputing 144* (2014), 560–568.
7. Levitin, D. J. *This Is Your Brain on Music: The Science of a Human Obsession*. Dutton Penguin Books Ltd, 375 Hudson Street, New York, NY, USA, 2006.
8. Levitin, D. J. Dr. daniel j. levitin: Neuroscientist, musician, author, 2015. Retrieved May 20, 2015, `http://daniellevitin.com/publicpage/`.
9. of Encyclopdia Britannica, T. E. *Clive Bell (or Arthur Clive Heward Bell)*. Encyclopdia Britannica, 2014.
10. Press, O. U. Oxford dictionaries: Language matters, 2015. Retrieved May 20, 2015, `http://www.oxforddictionaries.com/definition/english/emotion`.
11. Rouzic, M., 2008. Retrieved May 20, 2015, `http://photosounder.com/`.
12. Singh, J. F., 2012. Retrieved May 20, 2015, `http://flexibeatz.weebly.com/paint2sound.html`.
13. White, D., 2011. Retrieved May 20, 2015, `http://www.skytopia.com/software/sonicphoto/`.
14. Yanulevskaya, V., Gemert, J. v., Roth, K., Herbold, A., Sebe, N., and Geusebroek, J. Emotional valence categorization using holistic image features. *15th IEEE International Conference on Image Processing (ICIP)* (2008), 101–104.
15. Zhang, Q., Jeong, S., and Lee, M. Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system. *Elsevier: Neurocomputing 96* (2012), 33–44.
16. Zhang, Q., and Lee, M. Emotion development system by interacting with human eeg and natural scene understanding. *Elsevier: Cognitive Systems Research 14* (2012), 37–49.