

# TIME LAPSE VIDEO SONIFICATION; WATCHING AND LISTENING TO EVENTS UNFOLDING

**Karakonstantis Xenofon**

Electrical and Computer Engineer-  
ing (NTUA)  
xkarak@gmail.com

**Bakogiannis Konstantinos**

Electrical and Computer Engineering  
(NTUA) / Music Studies (UOA)  
konbakog@gmail.com

**Georgaki Anastasia**

Music Studies (UOA)  
georgaki@music.uoa.gr

**Cambourakis George**

Electrical and Computer Engineering  
(NTUA)  
gcamb@ece.ntua.gr

## ABSTRACT

Sonification is a constantly evolving field, with many implementations. There is a scientific need to adopt alternative methods of analysis, especially nowadays when the amount of data and their complexity is growing. Moreover, contemporary music relies often on algorithms, whereas there is an open discussion about the nature of algorithmic music. After Xenakis' works, algorithmic music has gained great reputation. In the contemporary Avant Garde scene more and more composers use algorithmic structures, getting advantage of the modern powerful computers. In that project we aim to create music that accompanies time-lapse videos. Our purpose is to transform the visual informational content into music structures that enhance the experience and create a more complete audio-visual experience. For our application we use digital video processing techniques. Our concern is to capture the motion in the video and we focus on the arrangement of the dominant colours. We relate the background of the video with a background harmony and the moving items that stand out against the background with a melody. The parameters of the music rhythm and video pace are taken into consideration as well. Finally, we demonstrate a representative implementation, as a case study.

## 1. INTRODUCTION

"Imagine listening to changes in global temperature over the last thousand years. What does a brain wave sound like? How can sound be used to facilitate the performance of a pilot in the cockpit?" With these questions, "Sonification Handbook", a book of great importance, introduces the idea of "sonification". Sonification is "the technique of rendering sound in response to data and interactions". [1]

*Copyright: © 2018 Karakonstantis X. et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License 3.0 Unported](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

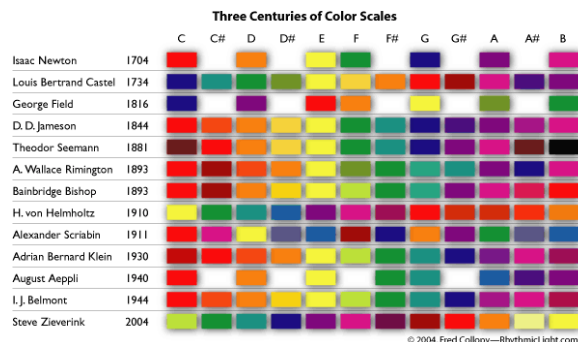
Despite the scientific essence of the term, it is not used only for academic purposes. On the contrary, everyday life offers plenty of examples. Many 'non-sonic' events convey their information through sound. The 'tic-tac' of a simple watch is an acoustic indication of a time pass of one second, a car alarm of a break-in and the beam of an ECG of the patient's heart rate. Information is a useful key to examine all the steps of a music procedure, from its creation and its transmission to its perception. [2]

Audition is one of the basic human senses. Therefore, it is expected to be used as a vehicle which transfers information. Nowadays, people can hear colours [3] and listen to gravity waves<sup>1</sup>. "Applications (of sonification) range from topics such as chaos theory, bio-medicine, and interfaces for visually disabled people, to data mining, seismology, desktop computer interaction, and mobile devices, to name just a few".

One of the most common fields is that of Image and Video Sonification. [4] The search for connections between images and sounds lead many scientists, musicians and artists (Isaac Newton, Herman von Helmholtz, and Scriabin, to name but a few) to relate certain notes with colours (Figure 1). A very significant correlation of colours and sounds is related to a neurological phenomenon, called 'synaesthesia'. This condition results in merging of senses that aren't normally connected. Therefore, some people, can literally hear colours and see sounds. Many famous musicians and composers have been inspired by their neurological 'gift' (from Franz Liszt to Olivier Messiaen). Apart from them, the joined inspiration of image and music has been met in many cases. Many painters have attempted to depict music, sounds and noise, or have been inspired by them, such as Marcel Duchamp, Wassily Kandinsky, Paul Klee, Henri Matisse, Joan Miro, Pablo Picasso, Arnold Schoenberg etc. In addition, many composers have created music from visual stimuli [5]. Tibetan monks used visual motives of nature to create music [6], Iannis Xenakis developed UPIC, a digitized tablet that transforms lines into music and Anestis Logothetis invented a novel graphic notation system based on sound attributes. Abstract animation, although it is a visual art, is often called *visual music*, due to the structural

<sup>1</sup>The official site of LIGO team, which won the Nobel Prize in Physics in 2017. <https://www.ligo.caltech.edu/video/ligo20160211v2>.

base similarities to that of absolute music and because it can induce emotional responses that are more related to the form of the expression than to the content of the imagery [7,8]. Furthermore, there are many artists, such as Malevich, Pollock, Rothko, Kline, Francis, who try to achieve a common hybrid form of audio and visual. [9]



**Figure 1.** Proposals for mapping of colours with notes<sup>2</sup>.

Actually, the compositional interconnection of arts and music – which led to the field of “image sonification” – is quite common in the late 19<sup>th</sup> and in the 20<sup>th</sup> century [9,10]. In the last decades there are plenty of approaches that involve appropriate computer music software [11]. Our project is a part of that ongoing procedure. Moreover, in our implementation we are interested in transforming visual information in music structures, not just sounds. For that purpose, we use some basic composing principles, so as to achieve an easy relation of the audio outcome of the algorithm with what an average listener perceives as music. [12]

## 2. GOAL AND METHODOLOGY

Our project aims to turn a video into music. Through a digital processing of a video data are collected. After an algorithmic procedure, these data are translated into audio structures resulting in a music that accompanies the video. The moving parts of the video against the static background are related to the parameter of the rhythm of the music, whereas the arrangement of colours transforms both the melody and the harmony (“the musical background”).

The project has been implemented with the use of appropriate software, i.e. Python and the libraries of OpenCV regarding video processing, analysis and sampling. [13,14] The Mingus library has been used in order to create MIDI files. The audio output occurred after inserting these MIDI files into a Digital Audio Workstation (DAW, such as Ableton, Cubase and Reaper), using their digital sound libraries, which include deep-sampled acoustic instruments and synthetic sounds.

This project has been a part of an interdisciplinary research of students (both under and postgraduate) of two collaborating departments, that of Electrical and Computer Engineering (NTUA) and Music Studies (UoA).

Therefore, the goals have been set in order to fulfill mainly academic criteria, rather than aesthetic or commercial. Its implementation has been a practical indication of how the stored information of a system (here, the data of video processing) can be translated into information of another system (here, MIDI / notes and sonic events). The first concern of such a transformation is which information to use. A digital video is a collection of moving images (frames), each one of them including many pixels. In most videos, the frames follow one another in a rate of at least 25 frames per second (hereafter fps). Although this value is crucial so as to give the sense of visual continuity, as far as music is concerned it is extremely high. For example, if each frame was a click of the metronome, 30fps means a tempo of 1800 crotchets per minute. Not only there is no musician that can play that fast, but for our audio perception is redundant. Therefore, a sampling is needed, with a sampling rate that can have musical meaning.

In order to achieve a better understanding of the video it is essential to look not only at each frame but also at its parts. For example, supposing we get the information that half of the frame is black and half white. That is just statistic knowledge, but it doesn’t reveal a lot about the informational content of that frame, i.e. where the black and white parts actually are. That is connected strongly with the concepts of entropy / information. [2] Therefore, in order to be able to describe more accurately a video, a segmentation procedure is necessary. However, we could say that a digital frame is already been segmented, in pixels. But in a typical digital video each frame consists of at least 320x640 pixels. It is easily understood that it is quite a great amount of information that cannot provide for any musical outcome. Group of pixels, in other words frame segments, should be used.

Regarding the musical video sonification, an algorithmic interpretation of the data from video processing is, for sure, not enough. The goal is to create music that accompanies the video, interacting with that, enhancing the viewing experience. Therefore, the music should be in connection with what the video presents. For example, the music to accompany a robbery should be quite different from that of scuba diving. We believe, therefore, that a universal algorithm, that could accompany whichever video, is, at least, quite a task, if not impossible. Each algorithm should be designed taking into consideration the individual parameters of each type of video.

For our implementation, we chose to work with time-lapse videos. The way an event is unfolding creates very interesting visual impressions. Moreover, the background stays, more or less the same, therefore, resemblance can be found with musical harmony. It is the basis, the static substrate where the more moving part occurs, that of melody. The melody can be parallelized with the depicted unfolding of the event. For us, it has been quite a challenge to achieve musically the ambient relaxing character of a time-lapse video, to connect the music with the impression of the moving image, but also to create an understood transformation of “what happens” while the video goes on.

<sup>2</sup>Screenshot from Rhythmic Light Website:  
<http://rhythmiclight.com/archives/ideas/colorscales.html>

### 3. PRESENTATION OF THE ALGORITHM

Quite shortly, our algorithm samples the video with the rate of 1 fps. After that, the frame is segmented in 8 pieces. The comparison of each segment with its previous and with its neighboring segments as well, reveals information about the movement. When movement is noticed the melody evolves. In each segment a colour analysis takes place. Some clusters of predominant colours are connected with some notes, according to a colour wheel. The pitch and the volume of the note are derived from the HSV chromatic model values. The same procedure takes place in the whole frame.

It is time to describe the procedure more thoroughly.

The time lapse videos which we used in order to design the algorithm didn't present noticeable changes in short time intervals. Therefore, we chose a sample rate of 1 fps. That rate captures adequately the changes in the video. We tried also windows with different size and overlap percentage but no significant improvement occurred. The fps value we chose captures sufficiently the visual information that is needed in order to create music and it ensures fast analysis and computation as well.

After sampling, the frame is segmented in eight pieces (2 rows and 4 columns, as shown in Figure 2). When motion is captured in a segment a note is triggered. The frame is scanned serially. First the first row, from left to right, and then the second row. Therefore, if a note is to be triggered, that takes place at the appropriate moment regarding that scanning. It provides the time index. Despite scanning methods, we tried to implement probing methods as well, influenced by Yeo and Berger [15]. The free, improvisatorial characteristic of probing resulted in a fuzzy outcome. The audiovisual perception lost its coherence. On the contrary, the strict following of the same scanning path of each frame made clearer the connection between sound and image.

The tempo we have chosen is a very common one, that of 120 beats (quarter notes / crotchets) per minute. Therefore, with sampling rate 1 fps, every frame corresponds to two crotchets. The frame is segmented in eight pieces, therefore the time a frame lasts – i.e. 1 sec – is divided in eight parts. Therefore, every segment corresponds to a sixteenth note (semiquaver). We can see that a segmented frame acts as an alternative notation system.

In order to capture the moving the frame is converted to gray scale. The absolute difference of each pixel between two serial frames is zero (black) if there is no change between them, or 255 (white) if there is a very significant change (from total black to white and vice versa). Of course, we have all the values in-between. Therefore, the negative of the image in grayscale is produced. The great differences between the frames lead to bright pixels (close to 255) whereas the low in dark (close to 0). We have, consequently, to set a threshold so as to decide if the difference is significant enough to indicate change in each pixel. Our threshold has been set to 180. Therefore, from the negative picture, still in a gray scale domain, we

create a black and white picture (pixels with value less than 180 become 0 and more than 180 become 255). Then, the black and white picture has been blurred with a Gaussian function, in order to smooth corners, to make the image less sharp, to reduce image noise and to erase some unwanted discrete points (Figure 2). With that way, the remaining image comprises the contour of moving objects. After that procedure, we still have to decide if there is enough change in the picture so as to regard motion. For our implementation, if the objects inside the contours are more than 1500pixels, then we have enough moving area so as to consider motion. This procedure happens in every one of the eight segments of each sample.



**Figure 2.** The absolute difference between the frames (gray scaled and blurred) and the segmentation.

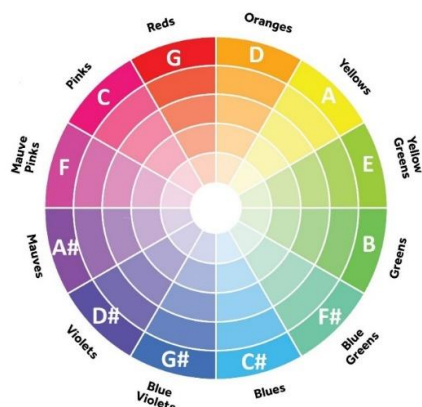
Till now we described the part of the algorithm which provides the time index for a note to be triggered, after deciding if motion is captured and in which segment, given the sampling frame rate and the music tempo. After deciding “when to play”, it is time to describe the other part of the algorithm which decides “what to play”.

Time lapse videos are basically constructed by a background and an unfolding event. Therefore, shapes or object tracking is not necessary. The change in colours, both in the whole frame and in each segment, can provide for the basic information. The basic concern of the algorithm is to correspond notes with colours. The chromatic model we use is the HSV. The reason is that its values can be directly related to musical parameters. Hue for the choice of notes, saturation for the pitch (the certain octave to which the note belongs) and value with the volume. In order to be consistent with the appropriate terms, hue is mapped with chroma and saturation with height.<sup>3</sup>

The algorithm, therefore, turns the rgb model into hsv and looks for the predominant colour. HSV is basically a cylinder. The relation of notes (chromas) with hues becomes in the slide where saturation and value are maximized. Then, that cycle is divided in 12 parts, creating a colour wheel. Every part corresponds to a certain note (chroma), with the use of the circle of fifths (C-G-D-A

<sup>3</sup>For example, the pitch class set of C notes (...C<sub>2</sub>,C<sub>-1</sub>,C<sub>0</sub>,C<sub>1</sub>,C<sub>2</sub>...) share the same chroma, but each one in different heights.

and so on). In that way a correspondence between the colour [16] and music harmony is achieved (Figure 3).

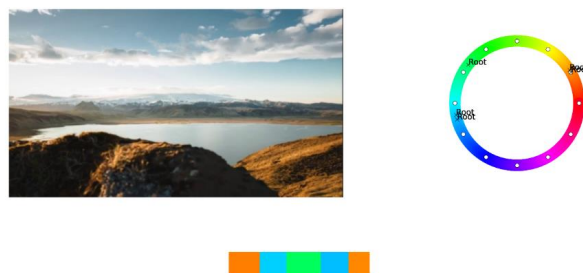


**Figure 3.** The colour wheel which depicts the correspondence of colours and notes.

We have described before how the motion in video is captured in the segments, and by that the notes are triggered. We have stated that motion in video is related to the melody. Therefore, through the procedure which is described above, in each segment we look only for the one predominant colour, which corresponds to a certain note, so as a monophonic melody occurs. Although the same concept is kept, there is a slight differentiation regarding the analysis of the whole frame. The whole frame corresponds to harmony. Therefore, more than one predominant colour should correspond to more than one note, so as to gain a cluster of notes, which acts like a chord. For that reason, we chose to create a cluster of five colours, therefore a 5 note chord (Figure 4). In order to conclude to these clusters, the k-means method is used. Generally, the application of k-means organizes all the  $n$  individual points ( $n$  corresponds to the number of pixels) into  $k$  groups. The points in each group share a high level of similarity, in comparison with those of others. Every cluster has a 'centre', a value that is very close to that of all of each members. This value is then assigned to all the members of the cluster. The reason that we concluded in 5 clusters is that after many trials we noticed that in the cases of more than 5 clusters, some not so statistically important colours have been considered as important. Therefore, insignificant colours have triggered certain notes in the melody, leading to an inconsistency in the transformation of visual information to sound. Respectively, in the cases of less than 5 clusters some important visual information has been lost.

Regarding the connection between pitch and saturation, we related low saturation with low frequency area and vice versa. The reason is that images with low saturation look weak, pale and without intense. We believe that this visual impression finds its musical analogy in low frequency sounds. On the contrary, the more intense the colour the higher the pitch.

Finally, the value dimension describes the lightness and brightness, which is connected with the volume of the note (the brighter the colour, the louder the sound).



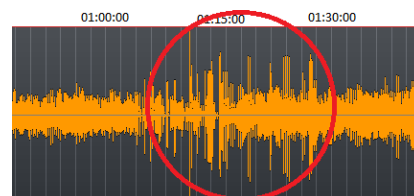
**Figure 4.** A representative frame analysis, where the cluster of 5 colours is presented below the image, whereas their position in the colour wheel, in the right of the image.

#### 4. PRESENTATION OF CASE STUDY; EYLENDA / ICELAND 4K

All the subjective parameters and thresholds have been chosen after plenty of trials with various time lapse videos. The final outcome provides for an adequate aesthetic result that is connected with the informational content of the video and it doesn't demand for heavy computations, unless there is a need.

We consider as the most representative example the EYLENDA | Iceland 4k, a creation of Marcus Sies and Flo Nick, students of Audiovisual Media at the Stuttgart Media University in Germany<sup>4</sup>. This video is a visual journey at the stunning landscape and wildlife of Iceland.

Due to the content of the video, many complementary hues of red (ground, soil, rocks etc), of blue (sky, sea) and of green (trees, grass etc.) have been noticed, because of the different landscapes of Iceland. Moreover, the static, motionless shots lead to notes of great duration. Furthermore, the low light and the natural lighting of the shots, as well as the low contrast of the colours of the landscapes of Iceland, resulted in frames with low saturation and therefore the melody of the sonification consists of many low pitch notes.

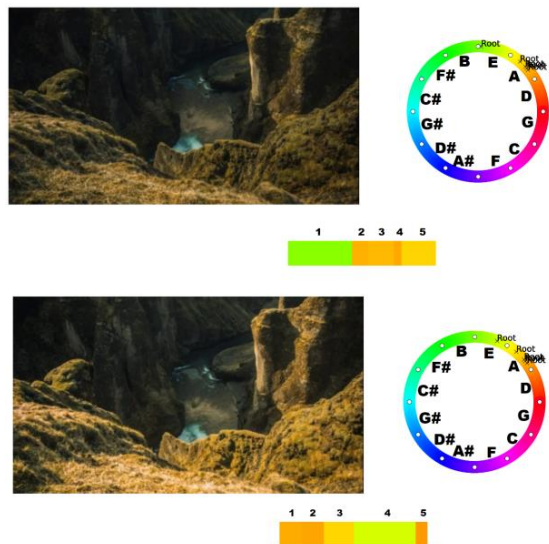


**Figure 5.** The waveform of a part of our implementation

<sup>4</sup>Sies, M., & Nick, F. (Directors). (2015). EYLENDA | Iceland 4k [Motion Picture]. [https://www.youtube.com/watch?v=U3r62Np\\_pxY](https://www.youtube.com/watch?v=U3r62Np_pxY)

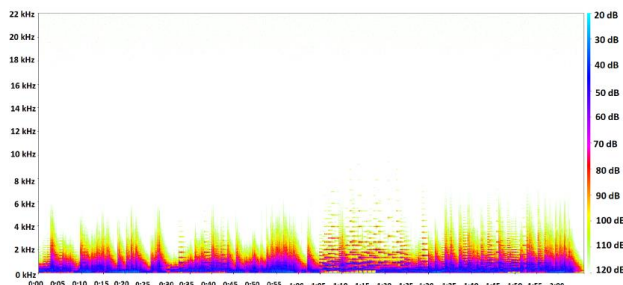


Figure 5 presents the waveform of the created music. The amplitude of the graph increases abruptly between 01:15 and 01:30. This graphical information reveals that the music is loud. The reason is that the brightness (value) of the colours is accordingly increased during that time, in consistency with our mapping.



**Figure 6.** The analysis of the frames that correspond to the 1:09, where an abrupt increase in the loudness of the music is noticed (as seen in Figure 3). It is obvious that the second frame is brighter, therefore with higher saturation, which leads to a higher volume correspondingly.

In the following spectrogram (Figure 7) we can notice that between 1:05 and 1:17, low frequencies are diminished. On the contrary middle frequencies seem to be quite intense. The reason is the high saturation values in the frames of that time. We can also notice that throughout the spectrogram the low frequencies are high due to the lighting and the low saturated images of the landscape of Iceland.



**Figure 7.** The spectrogram of our implementation. In left y axis there are the frequencies, in the right y axis the dBs, whereas in x axis the time.

The harmony varies throughout the piece. In some parts the resulted harmony is similar to a typical classical, but the result is not restrained by any musical form. For example, in the next picture, the cluster of notes consists of four notes (A, C#, G#, B), very similar to A major chord with the addition of the 7th and 9th note, whereas the next frame doesn't vary chromatically and just one colour is dominant, and hence only one note (A) is heard.



**Figure 8.** The analysis of two random frames in which we notice the difference outcomes of the music harmony corresponding their chromatic content.

## 5. CONCLUSION

According to our criteria the final outcome fulfils its purpose. The musical outcome of the algorithm is an adequate informational transformation of the time lapse video. It creates a more complete perception, combining audio and visual senses. The music accompanies sufficiently the video, concerning both the narration and the aesthetics. Finally, it has an independent aesthetic value. However, some parameters can evolve the implementation. The algorithm captures sufficiently the changed that happen gradually, such as the passing of a cloud, the succession of day and night and the blooming flowers. But it is not so decent in capturing very fast changes of small items in a quite static background, such as passengers passing by far in the background. Moreover, colour theory can be used in order to enhance the effectiveness of the transformation. For instance, there are some approaches that describe their “harmony” [17], such as complementary, analogous and triadic colour schemes. Their correspondence with harmony in music is not so profound, but its research may result in remarkable outcomes. Moreover, the algorithm can be enhanced and include more

modern techniques, such as machine learning. In any occasion, “A formal language of representation is needed in order to give composers the opportunity to fully explore the possibilities offered by computer music sound reversibility. If some consistent mathematical rules, and aesthetic values, were applied to create a formal environment, a collaborative and dialectic engagement with this new medium would become the source of many interesting approaches to composition”. [11]

## 6. REFERENCES

- [1] T. Hermann, A. Hunt. The sonification handbook. JG. Neuhoff, editor. Berlin: Logos Verlag, 2011
- [2] K. Bakogiannis, G. Cambourakis, “Information Physics; Towards a new conception of ‘Musical Reality’,” in International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS), 2017, 4(1), pp. 7-15
- [3] N. Harbisson, Painting by ear. Modern Painters, The International Contemporary Art Magazine. 2008 Jun:70-3.
- [4] WS. Yeo, J. Berger, “Application of Image Sonification Methods to Music,” In Proc. ICMC, 2005.
- [5] N. Τσινίκας, Αρχιτεκτονική και Μουσική. University Studio Press. 2009, pp. 24-27
- [6] PHU. Kaempfer, S. Pinkel, “The Earth Music of Thamkrabok Monastery,” In Leonardo, 2004, 37.1, pp. 25-30.
- [7] B. Evans, “Foundations of a visual music,” In Computer Music Journal, 2005, 29.4, pp. 11-24.
- [8] T. DeWitt, “Visual music: Searching for an aesthetic,” In Leonardo, 1987, 20.2, pp. 115-122.
- [9] Α. Γεωργάκη, “Προς μία κατηγοροποίηση του εικαστικού ήχου στη τέχνη του 20ου αιώνα: από τη χρωματική ακολουθία στη διαδραστική μουσική,” In Πολυφωνία, 2016.
- [10] Θ. Βελένη, Εικαστικές τέχνες και μουσική (τέλη 19ου και 20ός αιώνας), συναισθητικοί πειραματισμοί και οπτικοακουστικές εφαρμογές στην τέχνη του 20ού αιώνα: από τη συναισθησία στην πολυαισθητηριακή συνέργεια (Doctoral dissertation, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ). Σχολή Φιλοσοφική. Τμήμα Ιστορίας και Αρχαιολογίας. Τομέας Αρχαιολογίας και Ιστορίας της Τέχνης).
- [11] E. Lemi, A. Georgaki, and J. Whitney, “Reviewing the transformation of sound to image in new computer music software,” In Proceedings of the 4th Sound and Music Computing Conference, 2007.
- [12] K. Bakogiannis, G. Cambourakis, “Semiotics and memetics in algorithmic music composition,” In Technoetic Arts, 2017, 15.2, pp. 151-161.
- [13] J. Howse, OpenCV computer vision with python. Packt Publishing Ltd, 2013.
- [14] V. Pisarevsky, Opencv, the open computer vision library (2008).
- [15] WS. Yeo, J. Berger, “Application of Image Sonification Methods to Music,” In Proc ICMC, 2005.
- [16] KE. Burchett, “Color harmony,” In Color Research & Application, 2002, 27(1), pp. 28-31.
- [17] Z. O'Connor, “Colour harmony revisited,” In Color Research & Application, 2010, 35(4), pp. 267-73.