Framing through Music: A Pilot Study

Stephen James Krol
Monash University
Melbourne, Australia
sjkro1@student.monash.edu

Maria Teresa Llano
Monash University
Melbourne, Australia
Teresa.Llano@monash.edu

Cagatay Goncu
Monash University
Melbourne Australia
Cagatay.Goncu@monash.edu

Abstract

We present a system that automatically generates music from visual art based on the perceived emotion of the given input. We propose the generated music as a framing device that can enhance the aesthetic experience of people viewing Computational Creativity (CC) outputs. In this paper, we carry out a first study to test this by comparing the aesthetic experience of viewing paintings generated by CC systems accompanied by either textual framing, our proposed musical framing or both. We evaluate our system by means of qualitative user evaluations, which require participants to rank their aesthetic experience from best to worst. The results from the study demonstrated that the musical framing generated by our system provided a better aesthetic experience for users compared to the textual framing. Furthermore, the results suggest that with more work, a combination of textual and musical framing could be used to further improve the aesthetic experience for people viewing visual CC art.

Introduction

Framing is an important element of creative work. In some cases, the framing of a creative output can increase its value and perceived creativity (Charnley, Pease, and Colton 2012). Traditionally, framing has been defined as describing how a creative process works. This is done using text as it is an effective medium to convey the narrative associated with a creative artefact. However, Gross et al. (Gross et al. 2014) expanded this definition to define framing as including anything created with a creative artefact that aims to change how the work or creator is perceived. This allows for other mediums to be used as framing devices, such as music.

In this paper we introduce a system that aims to automatically generate emotive music from visual art. It does this by predicting the perceived emotion of the painting using two CNNs in the Valence-Arousal scale, then using this to find a piece of music from the VGMIDI database (Ferreira and Whitehead 2019) that elicits a similar emotion. This piece of music is then taken as inspiration to generate music in a similar style using Magenta's Music Transformer (Huang et al. 2018). We present this system as a framing tool that aims to improve the aesthetic experience of viewing CC artwork by automatically generating musical framing that conveys the emotion of the painting.

A survey was conducted to test the system and utilised artwork and framing text generated by The Painting Fool (Colton, Valstar, and Pantic 2008). The survey asked participants to rank their aesthetic experience when viewing artwork with text or music, or both, as a framing device. The results of the study suggest that musical framing generated by our system creates a better experience for users compared to viewing the painting with only the framing text. The results also indicated that with some improvements to the framing text, our system could be used with The Painting Fool to create a better experience for viewers.

The paper is organised as follows: first a background section covering previous work in framing for CC systems is presented. Then, a detailed description of the system and the methodology applied for the study are described. This is followed by a summary of the results and a discussion highlighting the insights from the study. Finally, we conclude and outline future work involving our system.

Background

The concept of framing was first introduced in CC by Colton, Charnley and Pease in (Colton, Charnley, and Pease 2011) as "a piece of natural language text that is comprehensible by people, which refers to a non-empty subset of generative acts". Simply put, framing is a device that has been used in CC to provide a description of how a program works, to explain its inputs or outputs and to provide insights about intrinsic factors behind the creative process behind it. Since its introduction, most works in CC that have used framing have done so in the form of textual commentaries attached to the creative output describing intentions, motivations and sources of inspiration that have guided the creative process towards the accompanying output.

However, novel approaches to framing have also been proposed within the community. An illustrative example is the approach proposed by Cook and Colton (Cook and Colton 2018) for ANGELINA, a computationally creative game design system, in which framing is used to communicate the design process over a period of time, allowing people to be involved in the "development and growth during creation not just after the fact", as put by the authors. In this case ANGELINA documents the design process in a lot of detail with information such as lists of tasks and projects, version history, notes on success or failure, etc. and uses this

Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)
ISBN: 978-989-54160-3-5

information to frequently inform its users about it through twitter and through the system's blog.

Additionally, initiatives on the use of different framing devices has also been put forward. For instance, Gross et al. (Gross et al. 2014) described how a computational process of poetry generation was framed by means of an abstract visualisation and then turned into paintings by an artist, while in (Cook et al. 2019), the authors shift the emphasis of framing from creative acts onto the audience that is engaging with the work, giving rise to a revised definition of framing as follows: "'Framing' refers to anything (co-)created by software with the purpose of altering an audience or collaborator's perception of a creative work or its creator", and the authors specifically highlight how this revised definition does not only refer to natural language as the only mechanism for framing.

In our work we follow this revised definition and propose the use of music as a method for framing visual art. We argue that the textual framing used currently by CC systems serve an informative purpose but fail at providing an aesthetic experience that is intrinsic to the act of experiencing and engaging with art. As described in (Charnley, Pease, and Colton 2012), the textual framing usually attempts to answer very practical questions about generative acts, particularly "why did you do X; how did you do X; and what did you mean when you did X?". In our work, we propose the use of music as an alternative (or complementary) framing device for visual art and argue that music can be used to convey meaning as well as to provide the audience a more engaging experience, which is ultimately one of the purposes of framing as described in (Cook et al. 2019).

System description

A diagram of our system can be seen in *figure 1*. The current version of our system aims to generate music in the following stages:

1. Attempt to classify the perceived emotion in the Valence-Arousal scale.
2. Find music annotated with similar Valence-Arousal values.
3. Use this music as a primer for the Music Transformer in Magenta's library (Huang et al. 2018)

Emotion Classification

The first stage of the system involves classifying the perceived emotion of the input painting using the Valence-Arousal model. To achieve this, we trained two convolutional neural networks (CNN) using the WikiArt dataset (Mohammad and Kiritchenko 2018). WikiArt is a dataset containing 4000 pieces of art that have been emotionally annotated by various observers. To the best of our knowledge, no work has been done in training a CNN to predict valence and arousal values associated with a painting using WikiArt. As suggested by the authors, we used the AG4 WikiArt dataset in our system. This dataset attributes an emotion to a painting when more than 40% of the annotators have applied it. An important step in preparing the dataset for training was mapping the categorical emotions to their respective values in the Valence-Arousal model. This involved analysing psychology literature and determining a general consensus for each emotion. Notable works used in these mapping are (De Bruyne, De Clercq, and Hoste 2020), (Hussain et al.), (Jin and Wang 2005), (Sellers 2013) and (Wang et al. 2021).

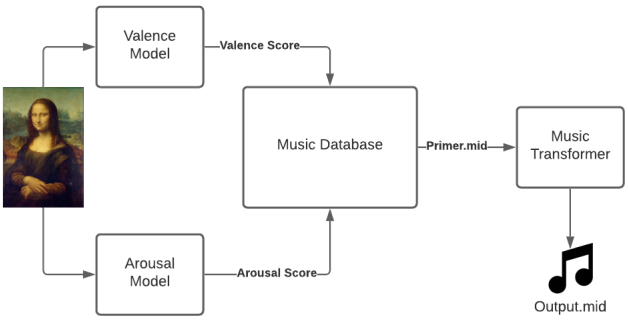

Figure 1: System Diagram. The painting is passed into both the valence model and the arousal model. The predictions are then used to find an appropriate primer to generate new music from.

To train the CNNs we used Keras version 2.24-tf with a Tensorflow 2.1.0 backend. Both models utilise transfer learning off the InceptionV3 (Szegedy et al. 2015) network pretrained on ImageNet. The architecture of both models can be seen in figure 2. We decided to train two separate models: one to predict valence and the other to predict arousal. This was because the WikiArt dataset is biased towards positive valence and high arousal emotions. Splitting the model into two networks simplified the unbalanced problem allowing us to effectively use undersampling to ensure balanced training. The tanh activation function was used in the final layer of both networks to ensure that outputs were in the desired (-1, 1) range. The networks were trained using a GTX1080ti with CUDA 10.1. Both networks used the Adams optimiser with a learning rate of 0.01 for 50 epochs with batch size 64 using the MSE loss function. The training and testing losses can be seen in table 1.

	Valence Model	**Arousal Model**
Training Loss	0.12	0.09
Testing Loss	0.13	0.11

Table 1: Model Loss Metrics

Annotated Music

The VGMIDI annotated database (Ferreira and Whitehead 2019) was used as inspiration for the music transformer. VGMIDI contains 95 tracks from various video games in midi format. These tracks have been emotionally annotated with the Valence-Arousal model by 30 participants, with the study using a total of 1425 annotaters. Each annotated track is split into different measures allowing participants to annotate different parts of the songs. The authors clustered the annotations into three groups: positive, negative and noise. The cluster with the most variance was considered noise

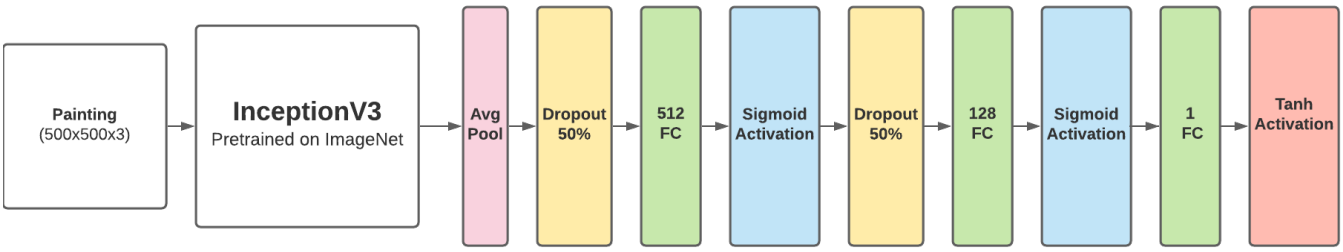

Figure 2: Emotion Detection Model Architecture. Model utilises an average pool layer on the InceptionV3 output the passes this through a series of dropout and fully-connect layers.

while the best cluster was the one with the most annotations. The best cluster was then used to generate the median valence and arousal score for the different measures in each track. We match our predicted valence and arousal scores to these median annotations to find a piece of music to prime the music transformer. The aim is for the music transformer to produce a piece of music with similar valence and arousal scores as the primer; therefore, the resultant music would convey the perceived emotion of the painting.

Music Transformer

Once an appropriate primer was found, Magenta's music transformer was used to generate music in a similar style to the primer. The music transformer is an attention-based neural network that attempts to generate music with coherent long-term structure. Compared to previous models from Magenta, such as performanceRNN (Simon and Oore 2017), the music transformer generates music that is more likely to play in a similar style to that of the primer. Maintaining a consistent style to its' input is necessary to ensure the generated music has a similar valence-arousal values as its primer; which in turn has a similar score to the artwork. We utilised a ported script (Bao 2020) from the Google Music Transformer notebook with melody_conditional_model_16 weights, to automatically generate midi files using the Music Transformer.

Pilot study

Methodology

In order to compare the experience of viewing a painting with textual framing versus musical framing, we prepared a pilot study. The study involved participants completing a small survey that contained eight paintings. The paintings and associated framing text were all generated by The Painting Fool (Colton, Valstar, and Pantic 2008). In two cases, framing text was unavailable so the authors manually created the framing text in the same style as The Painting Fool.

Four paintings conveyed positive valence while the other four conveyed negative valence. Two example paintings used in the study, one for each valence category, are shown in Figure 3. For each painting, our system generates music that attempts to convey the general emotion of the painting. We then use the painting, framing music and framing text to create three different experiences: Experience 1 is the painting combined with the framing text. Experience 2 combines the painting with the generated music. Finally, experience 3 combines both the framing text and music with the painting. The participants are then asked to rank the options in order of best experience. A link to a playlist containing the examples can be found in the appendix.

To control for order bias, three different surveys were created. The order of the paintings in each survey was shuffled using the Fischer-Yates algorithm which creates unbiased permutations. Participants were then randomly assigned one of the three surveys. We also shuffled the order of framing options for each painting.

Participants were recruited by responding to a call for participation posted in social media groups as well as sent to various email lists.

Results

A total of 21 participants responded to the call for participation and completed the survey. Participants had general knowledge in music and visual art; however, none were experts in either fields. It is worth noting that three of these participants had low vision.

Generally, participants voted that the music and painting provided the best experience. The second best experience was the combination of painting, music and text. Finally, the text and painting combination was generally voted as the least favourite experience. The results from the survey can be seen in figure 4. Low vision participants ranked the combination of painting, music and text as providing the best experience 66% of the time.

One participant stated that "the artist's description was bland, judgemental and contained a lot of useless information". Another participant stated that the "artist's captions need to explain more about the art or the emotion you receive from it." These comments suggest that the automatic framing generated by The Painting Fool could be enhanced with additional information or that alternative ways of framing may be more effective at communicating certain aspects of the creative process.

Discussion

The results from the survey demonstrated how the music generated by our system provided a better aesthetic experience for the users compared to viewing the painting with the automatically generated framing text. While this illustrates the benefit of using our system as a framing device, it also highlights improvements that are necessary when using text as the framing device. The Painting Fool's framing

(a) Example of a Painting Fool image that conveyed a positive valence and link to generated music: https://bit.ly/3jnpoyg.

(b) Example of a Painting Fool image that conveyed a negative valence and link to generated music: https://bit.ly/3s5xaRt.

Figure 3: Example of paintings from the Painting Fool used in the study.

text is not emotional but rather informational, even though it tries to convey its intention based on its perceived mood. In comparison, the music does attempt to invoke an emotional response in the user. Furthermore, while the painting and music combination was generally voted as the best experience, it was closely followed by the painting, text and music experience. This suggests that with some improvements to the framing text, our system could be utilised with The Painting Fool to create a better experience for users. This would be preferred over just using music to frame the artwork as there would be framing information (such as the artist's inspiration), that would not be effectively communicated through music.

An interesting application of our system would be utilizing it as a method to improve the accessibility of visual art, and visual CC outputs, for people with vision impairments. Although framing text provides useful information, it does not effectively convey an aesthetic experience associated with the art. This aesthetic experience is one the

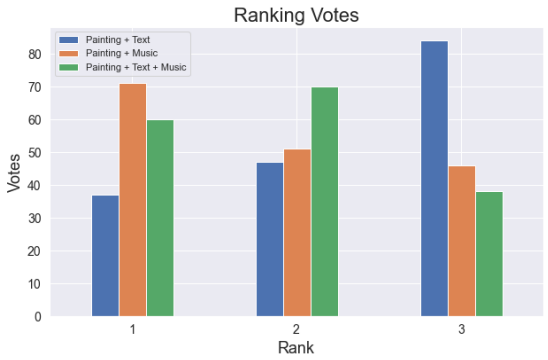

Figure 4: Survey Results: Demonstrates how the participants ranked the different experiences. In this study, 1 was the highest rank and 3 was the lowest.

main reasons sighted people view visual art and it should also be available to the visually impaired. Low vision participants ranked the combination of painting, text and music as the best experience 66% of the time. We hypothesise that the combination of framing text and music provides a better overall experience for the visually impaired as it allows the users to access more information relating to the artwork. Users can access both framing information and an aesthetic experience.

Conclusions and Future Work

In this study we introduced a system that could generate emotive music from visual artwork. The system classifies the perceived emotion of the painting using two CNNs and then utilising both the VGMIDI database and Magenta's music transformer, generates new music that conveys the emotion of the input painting. We proposed that this system could be used as a framing device for CC systems that create visual outputs rather than just using traditional text based framing. A study was conducted to test how the musical framing generated by our system affected the aesthetic experience of the participants viewing visual art generated by The Painting Fool. The results suggested that the musical framing generated by our system provided a better aesthetic experience than viewing the paintings with just the framing text provided. Furthermore, the results indicate that with some improvements to the framing text, our system could be combined with The Painting Fool to create a better experience associated with the artwork.

We see the benefits of alternative forms of framing going beyond its current use; for instance in order to make CC outputs more accessible. Future work will look into using this system to improve the accessibility of visual art for the visually impaired. This will involve adding more features to the system, such as including ambient sound effects of objects detected by the system within the painting. Compared to just using textual framing, music can create an aesthetic experience associated with the artwork for the visually impaired. Combining our system with framing text could significantly improve the accessibility of visual art for people with vision

impairments. This would also add a dimension of explainability to the generated musical framing, a feature that has been identified as important for CC systems (Llano et al. 2020). By including information of the objects detected by the system in the music, audience members can better understand what the system 'sees' from the paintings and as a result, better understand how the system works.

In this study, we did not evaluate the music directly. However, future work could also involve investigating the novelty of the generated music compared to its inspiration. Furthermore, investigating using features other than valence and arousal to generate music, such as velocity or tone, could allow the system to better frame its input artwork.

Acknowledgments

Cagatay Goncu is supported by the Australian Research Council (ARC) grant DE180100057. We thank the Faculty of Information Technology at Monash University for their support and Professor Simon Colton for providing the paintings and framing texts from the Painting Fool that were used in this work.

References

[Bao 2020] Bao, B. 2020. Piano transformer. `https://github.com/Elvenson/piano_transformer`.

[Charnley, Pease, and Colton 2012] Charnley, J. W.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *International Conference on Computational Creativity*, 77–81.

[Colton, Charnley, and Pease 2011] Colton, S.; Charnley, J. W.; and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. In *International Conferences in Computational Creativity*, 90–95.

[Colton, Valstar, and Pantic 2008] Colton, S.; Valstar, M. F.; and Pantic, M. 2008. Emotionally aware automated portrait painting. In *Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts*, DIMEA '08, 304–311. New York, NY, USA: Association for Computing Machinery.

[Cook and Colton 2018] Cook, M., and Colton, S. 2018. Redesigning computationally creative systems for continuous creation. In *International Conferences in Computational Creativity*.

[Cook et al. 2019] Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *International Conferences in Computational Creativity*, 156–163.

[De Bruyne, De Clercq, and Hoste 2020] De Bruyne, L.; De Clercq, O.; and Hoste, V. 2020. An emotional mess! deciding on a framework for building a Dutch emotion-annotated corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1643–1651. Marseille, France: European Language Resources Association.

[Ferreira and Whitehead 2019] Ferreira, L. N., and Whitehead, J. 2019. Learning to generate music with sentiment.

[Gross et al. 2014] Gross, O.; Toivanen, J.; Lääne, S.; Toivonen, H.; et al. 2014. Arts, news, poetry—the art of framing. In *International Conference on Computational Creativity*.

[Huang et al. 2018] Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; and Eck, D. 2018. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.

[Hussain et al.] Hussain, M. S.; AlZoubi, O.; Calvo, R. A.; and D'Mello, S. K. Affect detection from multichannel physiology during learning sessions with autotutor. In *Artificial Intelligence in Education*, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. 131–138.

[Jin and Wang 2005] Jin, X., and Wang, Z. 2005. An emotion space model for recognition of emotions in spoken chinese. In *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, 397–402. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Llano et al. 2020] Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational creativity. In *Proceedings of the Eleventh International Conference on Computational Creativity, ICCC 2020, Coimbra, Portugal, September 7-11, 2020*, 334–341. Association for Computational Creativity (ACC).

[Mohammad and Kiritchenko 2018] Mohammad, S. M., and Kiritchenko, S. 2018. An annotated dataset of emotions evoked by art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.

[Sellers 2013] Sellers, M. 2013. Toward a comprehensive theory of emotion for biological and artificial agents. *Biologically inspired cognitive architectures* 4:3–26.

[Simon and Oore 2017] Simon, I., and Oore, S. 2017. Performance rnn: Generating music with expressive timing and dynamics. `https://magenta.tensorflow.org/performance-rnn`.

[Szegedy et al. 2015] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015. Rethinking the inception architecture for computer vision.

[Wang et al. 2021] Wang, L.; Liu, H.; Zhou, T.; Liang, W.; and Shan, M. 2021. Multidimensional emotion recognition based on semantic analysis of biomedical eeg signal for knowledge discovery in psychological healthcare. *Applied sciences* 11(3):1338.

Appendices

Link to playlist containing examples used in the study: https://youtube.com/playlist?list=PLJXhSHZOX4QyIZPU-jEJf0Ajs0jj1xplO