Emotive Music Composition from Visual Sources of Inspiration

Dylan Lasher, Tyler Hedgepeth, Nickolas Nathan Taylor, Paul Bodily
Computer Science Department
Idaho State University
Pocatello, ID 83209 USA
{lashdyla, hedgtyle, taylnick, bodipaul}@isu.edu

Abstract

Cross-Domain creative systems can learn how to creatively map and/or transform semantic concepts from remote domains to their own. This project sets out to explore how a computational system relies on what it has learned from analyzing images to produce creative work. Specifically, the system presented here extracts semantic content and themes from an input image and uses this knowledge to creatively generate emotionally-appropriate musical artifacts to accompany the image. An independent experiment was conducted to provide a performance assessment and direct future work.
Source: `github.com/DLasher95/TransEmote`

Introduction

A hallmark of creative minds is the ability to apply knowledge from between creative domains. The concept of "cross-domain systems" allows us to take substantive inspiration from one domain and transfer it to another. Our system exists in the musical domain and extracts its inspiration from the space of visual arts. We computationally bridge the gap of these domains because, devoid of words, both music composition and imagery portray vivid and emotional narratives.

Our system, called *TransEmote* for its ability to translate emotive expressions, analyzes the content of an input photographic image and uses artificial intelligence (AI) to creatively generate emotionally-equivalent music to accompany the image. The system extracts psychology-based emotions, derived from the semantic content and colors present, and forms an emotional profile for the input. This way, TransEmote is able to extract multiple, complicated emotional themes and map them into semantically-appropriate music to accompany the image. There exist many benefits in pursuing this potential line of research:

- An aid to the visually impaired, helping to convey the emotional magnitude of photos, and illustrations.
- A rudimentary approach to developing dynamic soundtracks for large media, such as movies or games.
- A tool for emotional industries, like mental health clinics, to generate calming and/or happy music without having the domain knowledge to compose music themselves.

The challenge in composing music, just as in photography, is the magnitude of choices. We present a novel approach to extract visually-portrayed emotions, as well as a number of mapping rules to generate various elements of music, such as tempo or the major/minor key. Our goal with this project is to present an early framework by which the music and the visual arts can be further connected through creative AI.

Related Works

Cross-domain creative systems are few within the community. Intelligent music composition has seen more thorough treatment. We consider each in turn.

Cross-Domain Systems Text-to-music currently has difficulties with reliably mapping text to changes in sound. For this reason, prior text-to-music systems tend to rely on shallow features of text to direct generation rather than semantic context. Rangarajan (2015), for instance, proposed three methods of mapping text to music.

1. Mapping letters to notes, and frequencies to duration.
2. Mapping vowels to notes and note duration.
3. Mapping vowels and their respective uses to notes.

The *Transpose* project set out to generate music that captured emotion in literature by studying changes in the distribution of emotional words (Davis and Mohammad 2014). Transpose assigned major keys to novels with more positive emotions and minor keys with more negative emotions. It also connected the frequency of words with the tempo. Their approach laid the groundwork for cross-domain music generation.

Intelligent Music Composition There has been a significant amount of work done to map emotions to discretized parameters, such as tempo indicating happiness (Hunter, Schellenberg, and Schimmack 2010) and melody indicating calmness. The resources concerning emotional music generation is scarce, however.

It is important to note that these parameters are largely common in most cultures. However, an individual's experiences and developmental environment may influence their perception of the musical interpretation of emotions (Morrison and Demorest 2009).

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

Methods

TransEmote generates music according to the emotional content in images. It does so in several steps in Fig. 1.

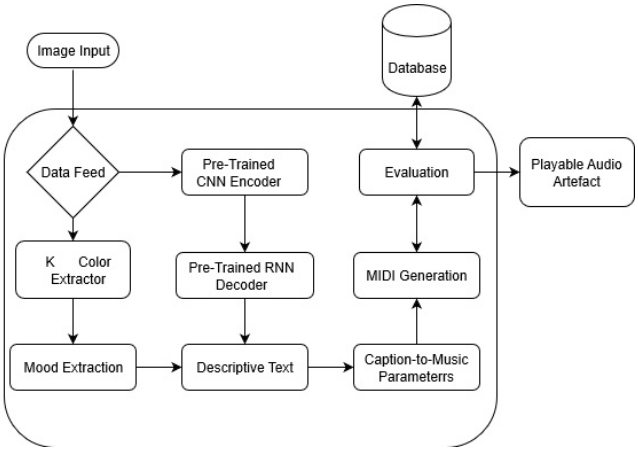

Figure 1: A system overview showing the flow of TransEmote.

1. Analyze input images and generate content descriptions.
2. Analyze the colors and use a color-to-emotion mapping to derive an emotional description.
3. Append the descriptions from (1) and (2) into an emotionally-descriptive profile.
4. Feed the distilled profile into a Natural Language Processing (NLP) section that maps the words to music-writing parameters.
5. Cycle the potential artefact through an intelligent quality check until it is satisfactory.
6. Output a playable audio file artefact.

Automated Content Extraction

The content of images are strong indicators of an emotional profile. Whether it be the happiness felt at a wedding or the anger depicted between two people arguing, TransEmote is tasked with taking into account objects/actions taking place in images. In this stage, TransEmote extracts the content of images for later analyses. We set out to develop an application which combines both computer vision and natural language processing to create accurate and comprehensive captions from provided images.

The system achieved this through two methods: A convolutional Neural Network (CNN) for extracting features out of the image and a Recurrent Neural Network (RNN) for translating the extracted features into natural sentences. We trained our system on a massive captioned set of images from *image-net.org*. Our approach implements the Tensorflow library and Facebook's *PyTorch*, which is an open-source machine learning library based on the Torch library. These are standard libraries used for applications such as computer vision and natural language processing.

A CNN is used for feature extraction and can produce a rich representation of input images by embedding it into a fixed-length vector (Vinayals et al. 2014). This representation can be used for a variety of vision tasks, which makes it a natural choice to use a CNN as an "encoder" by using the last hidden layer as an input into an RNN. We utilized a VGG16 CNN architecture because of its overwhelming preferred use in the field of machine vision, which is outlined in Simonyan and Zisserman's (Simonyan and Zisserman 2014) proposal of the system.

The images are then sent through the network and encoded into an array. The feature vector is linearly transformed to have the same dimension as the input to the RNN network. For our RNN model, which will "decode" the CNN output vector into readable text, we chose the standard long short-term memory (LSTM) network. This is a special type of RNN which is capable of learning long-term dependencies (Hochreiter and Schmidhuber 1997). For the RNN decoder training phase, the pre-trained CNN extracts the feature vector from a given image. After a linear transformation, the LSTM network input is the same as the CNN output. The decoder's source and target texts are predefined. Using these source and target sequences with the feature vector, the LSTM decoder is trained as a language model conditioned on the feature vector.

Color Analysis

The next step of TransEmote's emotional profiling abilities lie in the color analysis of each input image. From the developmental range of childhood, emotions are commonly associated with colors. In turn, we grow up to associate colors with emotions (e.g., *anger* is "red", *happy* is "yellow"). Extracting prominent colors in an image will help in establishing what emotions the user is experiencing. Fugate and Franco's (Fugate and Franco 2019) research into English speakers yielded a guide for major RGB values:

- Anger: Red (255, 0, 0)
- Calmness: Light Blue (0,128,255), Turquoise (0, 255, 255), White (255, 255, 255)
- Disgust: Sickly Green (204, 204, 0), Orange-Brown (204, 102, 0)
- Fear: Black (0, 0, 0), Red (255, 0, 0)
- Happiness: Yellow (255, 255, 0), Turquoise (0, 255, 255)
- Sadness: Navy Blue (0, 0, 255), Gray (160, 160, 160), Gray-Blue (0, 102, 204)

To extract the dominant colors, TransEmote uses a k-means clustering approach. We worked with three colors for this study so that we may focus results while still accounting for multiple emotional expressions.

We use a powerful machine learning library, scikit-learn, for our k-means clustering analysis. The RGB color scale forms a 3-dimensional vector space with orthogonal components. We can think of each pixel as lying somewhere in the 3D color vector space. After running this algorithm on our input image, we are able to derive a portrait of the 3 clustered, dominant colors in Fig. 2.

Once we have the RGB coordinates of our three dominant colors, we simply need to calculate the Euclidean distance to

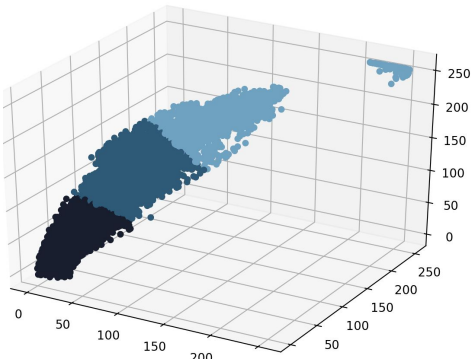

Figure 2: Example pixel colors extracted from the image in Fig. 4 shown in vector space.

Figure 3: Profiles of the three most dominant colors for the input image shown in Fig. 4.

each of the Fugate-Franco (2019) emotional colors. When TransEmote finds the nearest emotions to our dominant colors, we finally have our emotional color profile in Fig. 3.

The rest of the major colors, such as violet, were added without labels in order to avoid color values being matched to unreasonably distant emotional colors. This approach allows for the expression of multiple complex, and even disjoint, emotions that are commonly expressed through visual expressions (e.g. fear and calm). We simply have to append the emotions to the caption we previously generated.

Caption Distillation

Now the system needs to hand off this semantic information to the part of the system that maps it to musical parameters. First, we need to convert the captioned output into a friendlier format: an array of key words. To do this, TransEmote removes all duplicate and non-keyword strings from the generated caption. In doing so, we can better extract the key nouns and verbs that influence the image's emotional tones in Fig. 4.

Music Generation and Critique

Semantic Evaluation

Semantic evaluation is critical to the way the system interprets and manifests the previous results. TransEmote uses a word distance approach (e.g. Word2Vec) as a generally intuitive, yet human-like way to transform input strings into meaningful musical context. Using a dictionary as our data structure, parameters can be added and [0, 1] scores retrieved, which can then be used to obtain real

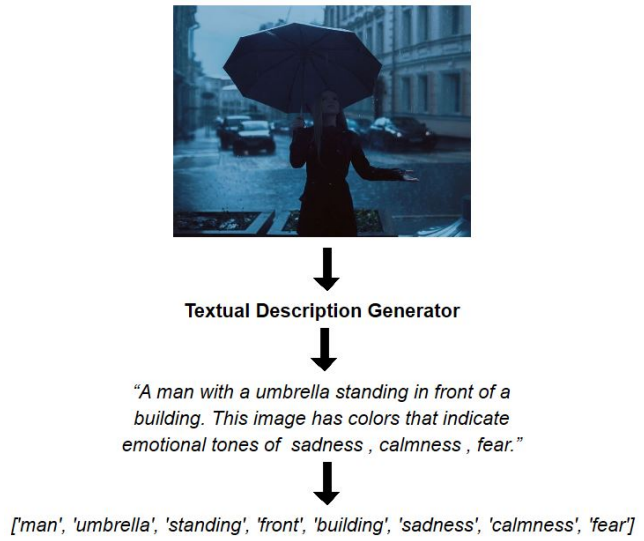

Figure 4: Example outputs of caption distillation. Full sentences are dissected for the central objects and emotions.

musical values such as *pitch*, *range*, and *BPM*. To improve on this approach, antonyms may be calculated (*slow* vs. *fast*) with a normalized score between them.

Instrumentation is also acquired using this approach. MIDI recognizes 128 unique instruments ranging from pianos, strings, brass, and synths. Each instrument is given a string representation and an average word-distance score is calculated for each input string. The instruments with the highest scores are used in the composition.

Rhythm

The length of a beat (quarter note) is established, then other rhythmic values are obtained through multiplication and division operations. TransEmote generates a rhythm by first providing it with a length, for example, *16 * q* for a four measure rhythm. A sequence of note lengths are generated which add up to the length of the section, critiquing and validating each beat by constraining the options based on parameters discussed above. This establishes a baseline for rhythmic generation upon which tonal ideas are built.

Tone

Chord progressions provide a backbone to the melodic structure of music. TransEmote defines a progression in two parts: *rhythm* and *intervals*. A rhythm, as discussed above, is used to generate a pseudo-random sequence of intervals. Based on the obtained musical and global parameters, constraints are imposed on the random generation to improve the sense of intentionality, quality, and consistency. These include preferring 4/5/6 cadences, ending on the tonic, and not resting on unstable (e.g. diminished) intervals. With a chord progression in place, melody is laid over it in a similar constrained psuedo-random way. Rhythms are constructed to fit the length of each chord in the progression, with its notes defined by the chord itself.

Writing to MIDI

TransEmote uses the Mido Python library to handle the creation and editing of MIDI files. The aforementioned rhythms, chords, melodies, and instruments are converted into MIDI message format, and subsequently exported to MIDI, where it can be converted or used in a variety of ways.

Results

Method We tested the success of TransEmote with a Google Form[1] sent to universities and online AI communities. Our survey had eight examples in total with a five-point Likert scale asking "how well does the music convey the emotion of the image?". This was a blind study consisting of randomly organized test and control samples. The four test samples had the musical artefact created by TransEmote paired with their images. The control consisted of randomly generated artefacts paired with random images.

Analysis Our survey received 109 full responses; which are ample data for effective analysis. We sum the times an artefact was given a four or five on each sample which count as successes. The results between the two groups are averaged. We found that 60.1% of test artefacts and only 20.4% of controls were rated by participants as 'accurately portraying the emotions of the images'.

To corroborate this, a two population t-test suggests a highly significant ($p < 0.0005$) difference between the two groups. We found no significant differences between responses between genders.

Evaluation of Creativity

Beyond the ability to generate emotionally-relevant music, it is important that our system also engages in creative behavior. Prior work in computational creativity suggests that creative composition systems should be able to generate music that is novel, high quality, intentional, and typical of the domain (Boden 2009).

The creativity of TransEmote exhibits these attributes. As far as we know, the system is the first of its kind to extract semantic information from visual mediums and effectively translate it into musical audio. This reinforces the argument for **novelty**. Also, the emotional expression in music is a major portion of composition, and our system is able to bridge the non-verbal gap between these domains. Regarding **quality**, creativity doesn't exist in a vacuum. We must involve the audience/community (Colton, Charnley, and Pease 2011). According to our survey results, the community at large does indeed ascribe value to the artefacts produced by TransEmote. The system also exhibits **typicality** because it utilizes existing domain rules of music composition and emotional expression. Lastly, our survey results prove the system is **intentional**. TransEmote sets out to extract visual semantics and uses them as an inspiration seed to generate author-guided, semantically-equivalent audio files. There is also an intelligent screening process to avoid artefacts that have already been generated.

[1] shorturl.at/svMQU

Conclusion

We present a cross-domain system that extracts Western emotional profiles from images and translates them into music. Based on our survey, TransEmote has empirically demonstrated that it is able to usefully and creatively map and transform emotions between these domains.

For future steps towards improving the system, we would like to experiment extracting more information from images. Additionally, we believe that utilizing evolutionary algorithms would advance the emotional expression of musical generation. Training the thematic extractor on non-photographic images would also broaden its descriptive range. Lastly, we believe that TransEmote can be thought of as a preliminary support tool for music generation. Here, we implemented general models of musical expression but users have the ability to specify further musical parameters. Users would be able to integrate rules of jazz, for instance, in order to exclusively express semantic jazz music. We assert, in short, that TransEmote is computationally creative. The artefacts TransEmote produces are not mere generations, but thoughtful expressions of emotion. It is the authors' hope that the next steps toward TransEmote-inspired systems will help to further enrich inspired music generation.

References

Boden, M. 2009. Computer models of creativity. *AI Magazine* 30(3):23.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. *Proceedings of 2nd International Conference on Computational Creativity*.

Davis, H., and Mohammad, S. M. 2014. Generating music from literature. *EACL 2014* 1–10.

Fugate, J., and Franco, C. L. 2019. What color is your anger? assessing color-emotion pairings in english speakers. *Frontiers in Psychology* 10:206.

Hochreiter, S., and Schmidhuber, J. 1997. Very deep convolutional networks for large-scale image recognition. *Nueral Computation* 9. 1735-80.

Hunter, P. G.; Schellenberg, E. G.; and Schimmack, U. 2010. Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. *Psychology of Aesthetics, Creativity, and the Arts* 4(1):47.

Morrison, S. J., and Demorest, S. M. 2009. Cultural constraints on music perception and cognition. *Progress in Brain Research* 178:67–77.

Rangarajan, R. 2015. Generating music from natural language text. *IEEE* 85–88.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv* 1409.1556.

Vinayals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2014. Show and tell: A neural image caption generator. *arXiv* 1411.4555:23.