Shimon the Robot Film Composer and DeepScore

Richard Savery and Gil Weinberg

Georgia Institute of Technology
{rsavery3, gilw} @gatech.edu

Abstract. Composing for a film requires developing an understanding of the film, its characters and the film aesthetic choices made by the director. We propose using existing visual analysis systems as a core technology for film music generation. We extract film features including main characters and their emotions to develop a computer understanding of the film's narrative arc. This arc is combined with visually analyzed director aesthetic choices including pacing and levels of movement. Two systems are presented, the first using a robotic film composer and marimbist to generate film scores in real-time performance. The second software-based system builds on the results from the robot film composer to create narrative driven film scores.

Keywords: Film composition, algorithmic composition, visual analysis, artificial creativity, deep learning, generative music

1 Introduction

Film composition requires a "connection with the film" and a deep knowledge of the film's characters[14]. The narrative arc is often tied together through key themes developed around characters; as Buhler notes "motifs are rigidly bound to action in film" [3]. Multiple authors have studied the link between musical themes and on-screen action [11][16]. Likewise, Neumeyer explores the relation of audio and visual, describing multiple visual and aural codes that link music and screen[15].

This deeply established practice emphasizing the relation between music, narrative and visuals, contrasts with existing computer film musical generative systems which focus on small form pieces, and do not include any video analysis. By including visual analysis and the film itself as intrinsic to the creation process, generative systems can begin to address the inherent challenges and opportunity presented in film composition. Analysis of film visuals also allow for a range of new approaches to generative music, while encouraging new musical and creative outcomes.

This research began by exploring what it means for a robot composer to watch a film and compose based on this analysis. With lessons learned from this implementation we were able to prototype a new software-based approach to film generation using visual analysis. This paper will explore the design of

both systems focusing on how visuals are tied to generative processes. The first system *The Space Between Fragility Curves* utilizes Shimon, a real-time robotic composer and marimbist that watches and composes for the film. This system acted as a catalyst to the work developed for *DeepScore*. The second system *DeepScore* is off-line and uses deep learning for visual analysis and musical generation. In both systems multiple visual analysis tools are used to extract low level video features that are then converted to meta level analysis and used to generate character and environment based film scores.

2 The Space Between Fragility Curves

This project was built around the concept of Shimon acting like a traditional silent film composer. The system was created for a video art piece called *The Space Between Fragility Curves* directed by Janet Biggs, set at the Mars Desert Research Station in Utah. A single video channel version premiered on the 17th of May 2018, at the 17 Festival Internacional de la Imagen in Manizales, Colombia. The two channel version premiered on the 14th of November, May 2018 at the Museo de la Ciencia y el Cosmos, in Tenerife, Canary Islands. The final film includes footage of Shimon interspersed amongst the footage from the Mars Research station.

Fig. 1. Shimon Practicing The Space Between Fragility Curves *(Photo by Janet Biggs)*

2.1 Shimon

Shimon (see Figure 1.) is a robotic marimba player, developed by the Robotic Musicianship Group at Georgia Tech, led by Gil Weinberg [8]. Shimon's body

is comprised of four arms, each with two solenoid activators striking mallets on the marimba. Shimon has toured worldwide and is used as a platform for many novel and creative musical outcomes.

2.2 Visual Analysis

With a set film given at the beginning of the project, visual analysis focused on extracting key elements that a composer may track within this specific film. Custom analysis tools were built in MaxMSP's Jitter, reading a JSON file generated by Microsoft's Video Insights. The JSON file included identified faces and their time and location on screen. It also included objects and a location analysis, defining if the scene was indoors or outdoors as well as the surrounding landscape. Jitter was used to extract director aesthetic choices, defined as conscious film choices that set the tone, style and pace of film. These tracked aesthetic choices were panning, zoom, similarity between camera angle changes, character movement and coloration. Parameters were then used to create a meta analysis of the overall pacing.

2.3 Musical Generation

A musical arc was set by the director, dividing the film into four minutes of character and object driven musical generation, with two segments given independent algorithmic processes. At the beginning of each run, four melodies are generated using a Markov model trained on melodies from Kubrick film scores. A Markov chain is a probability based model that bases future choices on past events. In this case we referred to three past events, using a third generation Markov Model for pitch and a separate fifth generation model for rhythm, both trained on the same data. Two melodies are assigned to characters, with the other two melodies set for indoor and outdoor scenes. Melodies are then used throughout the film, blending between each one dependent on what is occurring on screen. Melodies are varied based on movement of the chosen characters on screen, their position and external surroundings.

The first of the separate sections was centered inside an air chamber with director requesting a claustrophobic soundtrack. For this scene an embodied approach was used with Shimon. Shimon's design encourages the use of semitones, as both can be hit without the arms moving. Chords were then built around a rule set featuring chords built on these intervals. The second section was the conclusion of the film, which uses Euclidean rhythms [17], commonly used in algorithmic composition. The scene features one of the main characters riding an ATV across the desert. The number of notes per cycle is set based upon the movement of the ATV and position on screen.

2.4 Lessons Learned

For this system we did not conduct any user studies. We considered it a prototype to generate ideas for a complete system. As previously mentioned the

film is premiering in the Canary Islands and has multiple other showings lined up around the world indicating a certain level of success. Comments from the film director also demonstrated the importance of the link between visuals, "I love how clear it is that Shimon is making choices from the footage" [1]. Follow up emails also suggested that the generative material was effective however the overall musical structure could be improved "I watched the demos again and am super happy with the opening segment. The only other note that I would give for the second half is to increase the rhythm and tonal (tune) quality" [2].

After positive feedback from the director and informal viewings we reviewed the key concepts that were beginning to emerge. Extracting director aesthetic choices such as movement on screen and panning allowed an instant level of feedback that helped align the score with visuals. To some degree this naturally creates musical arcs matching the film's arc, however with only this information the music is always supporting the visuals and never complementing or adding new elements to the film. Like-wise character based motives were very successful from our small viewing sessions yet without intelligently changing these based on the character they also fell into a directly supporting role. Most significantly we came to believe that there was no reason for future systems to work in real-time. Shimon composing a new version for the film through each viewing provides a level of novelty but in reality the film will always be set beforehand and live film scoring is a vastly different process to the usual work flow of a film composer.

3 DeepScore

In contrast to Shimon acting as a real-time composer, *DeepScore* was created to be used off-line for a variety of films. This encouraged a significantly different approach to visual analysis parameters and methods used for musical generation. Where Shimon as a film composer focused on building a real-time system for one film, *DeepScore* aims to instead use more general tools to enable composition for multiple films.

3.1 DeepScore Backgound

Film Score Composition A successful film score should serve three primary functions, tonally matching the film, supporting and complementing the film and entering and exiting when appropriate[7, p.10]. From as early as 1911 film music (then composed for silent films) was embracing Wagner's concept of the leitmotif [3, p. 70]. The leitmotif in film is a melodic gesture or idea associated with a character or film component [3, p. 42]. While not the only way to compose for film, leitmotif's use has remained widespread most prominently by John Williams, but also by many other composers[2].

[1] Biggs, Janet. "Re: Demos" Message to Richard Savery, 26 October, 2017. Email
[2] Biggs, Janet. "Re: Demos" Message to Richard Savery, 6 November, 2017. Email

Deep Learning for Music DeepScore's musical generation and visual analysis rely on deep learning, a subfield of machine learning that uses multiple layers to abstract data into new representations. Deep learning has recently seen widespread adoption in many fields, driven by advances in hardware, datasets and benchmarks, and algorithmic advances [4, p.20]. In this paper we use Recurrent Neural Networks (RNNs) for music generation. RNNs are a class of neural networks that are used for processing sequential data. An RNN typically consists of one or more nodes (operating units) which feed their outputs or hidden states back into their inputs. In this way, they can handle sequences of variable length, and allow previously seen data points in a sequence to influence the processing of new data points and are thought to have a sort of memory. Standard RNNs suffer from a variety of issues which make them difficult to train, and so most applications of RNNs today use one of two variations known as Long Short Term Memory (LSTM) RNNs and Gated Recurrent Unit (GRU) RNNs. In each of these variations, the standard recurrent node is replaced with one which parameterizes a memory mechanism explicitly. An LSTM recurrent has three gates: the input gate, cell gate, and output gate, which learn what information to retain and what information to release. RNNs have been used widely in music generation. Magenta (part of Google's Brain Team) have successfully used RNNs for multiple systems to create novel melodies [10][9].

Deep Learning for Visuals For visual processing we primarily rely on Convolutional Neural Networks (CNN). A CNN is a neural network which uses one or more convolutional layers in their architecture and specializes them for the processing of spatial data. A convolutional layer applies an N-Dimensional convolution operation (or filtering) over it's input. CNNs have been shown to have the ability to learn spatially invariant representations of objects in images, and have been very successfully applied to image classification and object recognition [6, p. 322]. CNNs have also been used for musical applications, notably in WaveNet a model that generates raw audio waveforms[18]. WaveNet's model was itself based on a system that was designed around image generation, PixelCNN [19].

Film Choice - Paris, je t'aime A general tool for all films is far beyond the scope of DeepScore. We chose to limit the system to use on films where using leitmotifs for characters would be appropriate and emotional narrative arcs are present. Technical limitations of our visual system also restricted the system to films with primarily human main characters.For the purpose of this paper all examples shown will be from the 2006 film, *Paris, je t'aime* [5] a collection of 18 vignettes. The examples will use the vignette *Tuileries* directed by Joel and Ethan Coen. This film was chosen as it focuses on three main characters who experience a range of emotions.

3.2 DeepScore System Outline

DeepScore was written in Python, and uses Keras running on top of Tensorflow. Visual analysis is central to the system with the meta-data created through the analysis used throughout. Figure 2 demonstrates the flow of information through the system. Two separate components analyze the visuals, one using deep learning and the other computer vision. These two visual analysis units combine to create visual meta-data. The lower level data from the visual analysis is also kept and referenced by the musical generation components. Melodies and chords are independently created and annotated. With visual data extracted, the generated chords and melodies are separately queried to find those best fit to the characters in the analyzed film. After chords and melodies are chosen they are combined together through a separate process, using a rule-based system. These melodies and chord progressions are then placed throughout the film. After placement they are then altered with changes in tempo, chord and melody variations and counter melodies added, dependent on the visual analysis.

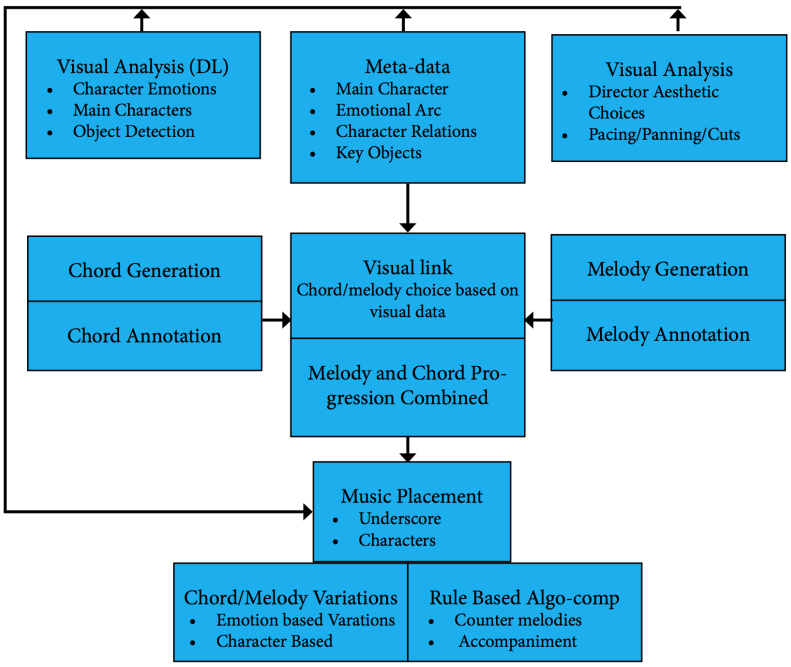

Fig. 2. DeepScore System Outline

3.3 Visual Analysis

The primary visual element tracked by *DeepScore* are the main characters and their emotions throughout the film. The three main characters are identified by which face appears most on screen. Emotions are tracked throughout the film using an implementation of an existing CNN [1] and the Facial Expression Recognition 2013 (FER-2013) emotion dataset[13]. FER-2013 was chosen as it is one of the largest recent databases and contains almost 36,000 faces tagged with seven expressions, happy, angry, sad, neutral, fear, disgust or surprise. Each frame with a face recognized is given a percentage level of each emotion (see Figure 3). Figure 4 shows the first 60 seconds of emotional analysis of Paris, je t'aime

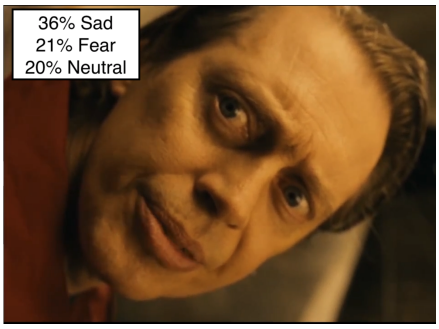

Fig. 3. Emotion Classification from *Paris, je t'aime*

Emotion was chosen for multiple reasons, at a simplistic level emotion is often considered a key component of both film and music. As previously discussed music should support the tone and complement the film, both characteristics relying on understanding the emotional content of a scene. Figure 4 demonstrates the emotional arc of the first sixty seconds.

In addition to emotions, the previously created custom analysis system for director aesthetics was used for DeepScore. The emotions were then combined with the director aesthetics into higher level annotations. These key points dictated when the musical mood would change and set the transition points from moving between characters.

3.4 Musical Generation

Chord Generation Chord progressions are generated using a character recurrent neural network, based on Karpathys[7] char RNN model and using the Band-in-a-box data set. The data set contains 2,846 jazz chord progressions. This type of RNN is often used to generate text. Character RNN is a recurrent neural network architecture which generates the next step in a sequence conditioned on

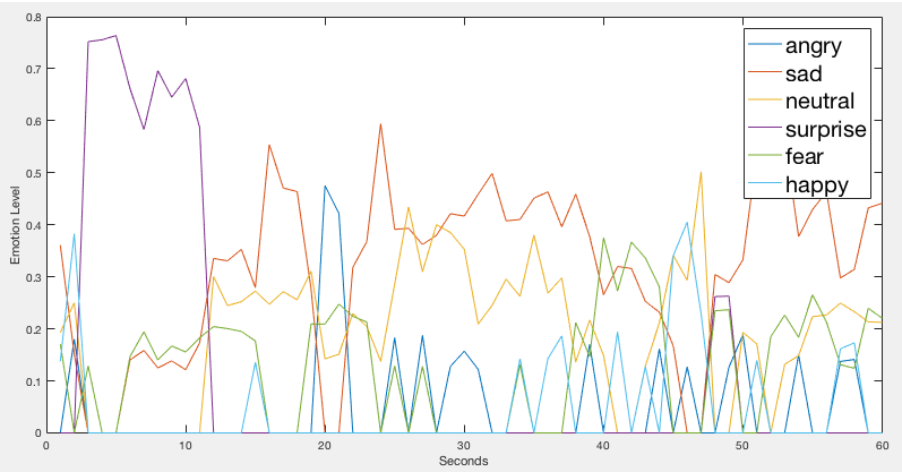

Fig. 4. Graph of Emotions

only the previous step. Running DeepScore creates 1000 chord progressions, all transposed to C. These are then annotated with consonance level and variation, both between 0 and 1. Consonance consists of how closely the chords align to chords built off the scales of either C Major, or C Minor. For example in C Major chords such as D minor 7 and E Minor 7 are given a 0 for consonance, D7 is given a 0.5 and Db Minor would be given a 1.0. The variation level refers to how many different chords are within a progression.

Melody Generation Melodic ideas are also created by an RNN in this case an LSTM and uses the Nottingham dataset, a collection of 1000 folk tunes. We tested multiple datasets, but found the folk melodies in this dataset worked best for their ability to be easily rearranged post creation and still retain their melodic identity. Each created melody is 8 bars long. Melodies are then annotated based on rhythmic and harmonic content using custom software written in python, combined with MeloSpy from the Jazzomat Research Project. Table 1 shows each extracted parameter. Annotation is based on principles on a survey of research into the impacts of musical factors on emotion [12].

3.5 Adding Music to Image

With chords and melodies annotated the system then uses the visual analysis to choose an appropriate melody for each character. This melody then becomes the leitmotif for the character. Starting with the character most present throughout the film three melodies and three chord progressions are chosen that align with the emotional arc of the main character. Two other characters are then assigned melodies primarily choosing features that align with their emotional arc, while contrasting that of the main character.

Musical Feature	Parameters Extracted
Harmony	Consonance, Complexity
Pitch	Variation
Interval	Size, direction and consonance
Rhythm	Regularity, Variation
Tempo	Speed range
Contour	Ascending, Descending, Variation
Chords	Consonance, Variation

Table 1. Musical Parameters for Emotional Variations

At this point the chord progression and melody have been independently created and chosen. To combine them a separate process is used that alters notes in the melody, while maintaining the chord progression. Notes that do not fit within each chord are first identified and then shifted using a rule based system. This system uses melodic contour to guide decisions, aiming to maintain the contour characteristics originally extracted. Figure 5 shows two melodies and chord progressions after being combined. Both were chosen by the system for one character.

Main themes are then placed across the film, with motifs chosen by the dominant character in each section. In the absence of a character a variation of the main characters theme is played. After melodies are placed, their tempo is calculated based on the length of the scene and the character's emotion. Dependent on these features either a 2, 4 or 8 measure variation is created.

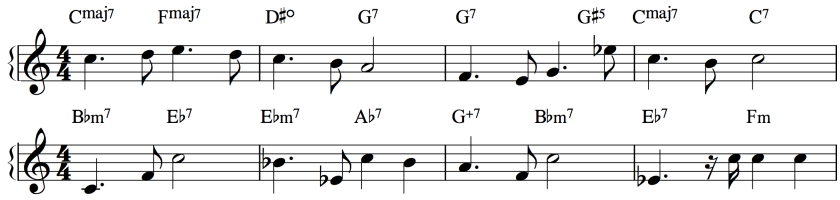

Fig. 5. Two Alternate Melodies Created for the Main Character

3.6 Counter Melodies and Reharmonization

Referring back to the visual analysis meta-data each section of the film is then given a counter melody or reharmonization dependent on the emotional characteristic in the scene. These melodies are developed by a rule based system using the same parameters presented in table 1. All emotions are mapped to different variations based on a survey of studies in the relation between music and emotional expression[12, p.384-387]. In addition to those parameters counter

melodies use volume variation and articulations (staccatos and tenutos). The interactions between characters are also considered, such as when one character is happy while another is fearful.

3.7 Output

The system's final output is MIDI file and a chord progression. This file doesn't contain instrument designations, but is divided by octaves into melody, countermelody, bass-line and chordal accompaniment. In the current iteration this is then orchestrated by a human. The aim in future work is to include orchestration in the system.

4 Evaluation

4.1 Process and Participants

For the evaluation we created three, one minute long clips of music generated by the system. The first clip used *DeepScore* to generate material, but did so based on emotional keywords representative of the short film as a whole, and not at a frame by frame level. This eliminated the immediate visual link to the film and was used to gauge a reaction to the generated music. Considerable work was done to ensure that this first clip was indistinguishable in terms of quality and processes to that of music generated purely by visuals. The other two clips used visual analysis as described in this paper on two different scenes. The three clips were randomly ordered for each participant. For each clip participants were asked three questions and responded with a rating between zero and ten. These questions were: *How well does the music fit the film's tone, how well does the music complement the on-screen action and what rating would you give the music considered separately to the film?* They were then given an option to add any general comments on the score. After answering questions on the generated material they were presented a brief overview of a potential software program implementing visual analysis. Participants response's were anonymous and they were not given any background information on how the music was created, or that it was computer generated.

We surveyed five film composers and five film creators. The film composers were all professional composers having a collective composing experience of over 700 publicly distributed films. Film creators were directors, editors or writers or often a combination of the three. Film creators had a combined experience of over 100 publicly distributed films. While only a small sample group, we chose to focus our evaluation on leading industry participants to gauge just how effective the system is to those with the greatest insight into the area.

4.2 Quantitative Findings

Figure 6 and 7 present the results from the survey asking about tone, if the music complements the scene, and the general rating of the music. The versions that

included visual analysis were rated better across all categories, despite using the same generative system. Composers rated the music of the non-visual analysis particularly low. There was also a consistent gap between the ratings of the film creators and film composers.

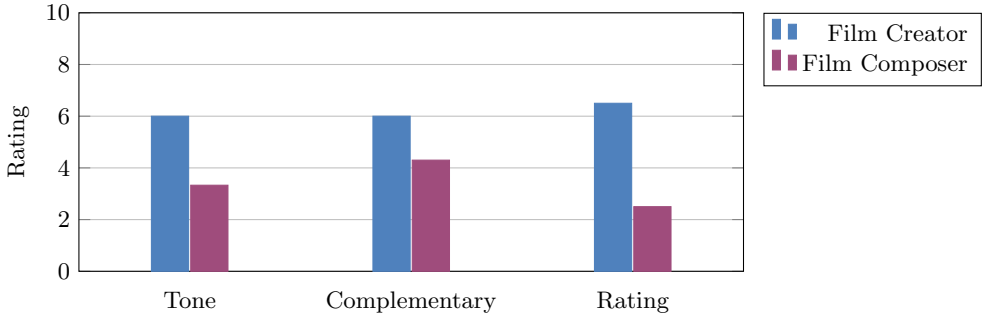

Fig. 6. Ratings for Music Generated with Keywords

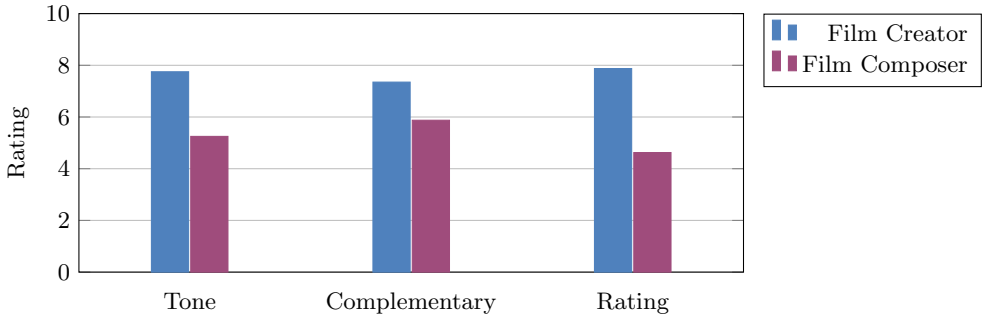

Fig. 7. Ratings for Music Generated with Visual Analysis

4.3 Qualitative Findings

From the general comments to the scores we had many unexpected insights and found significant variation between creator's and the composer's responses. For the version generated with keywords and no visual analysis one director noted: *"The music as such is good to listen. But I don't think it is in sync with the actions on the screen"*. Another director described, *"Simple score but is not taking me emotionally anywhere"*. The composers also described the lack of relation to the emotions on screen, *I think that the music is in the right direction*

but lacking the subtlety of action and emotion on the screen. and it *"doesn't seem to be relevant to what's happening on the screen in terms of tonality"*. Participants had no previous insight into our analysis of character's emotions in other scenes.

As expected from the quantitative ratings, comments were generally more positive for the versions using *DeepScore's* visual analysis. One director noted that the music was *"appropriate for the scene"*, a composer described *"it generally fits very well"* and another composer mentioned *"the music does well with changing action and emotion"*. Two composers did notice that the transitions that occurred as the scene changed could be *"a bit jarring"* and *"the sudden change is too extreme"*.

A key finding stemmed from the impact the music was having on the interpretation of the film. One director said that they *"feel like the music is very much directing my feelings and understanding of what is going on. If music had different feel the action could be interrupted differently"*. This related to one of the most common observations from composers, that the score *"plays the action, although it didn't add to more than what we already see on screen"* with another composer describing they would have led the scene in a different direction *"to either something more comical or menacing"*.

From the qualitative findings we drew three main conclusions on how the system operated. By closely following the emotions, changes can be too significant and draw attention to moments on screen that are already very clear to the audience. This leads on to a larger problem that the current mapping of emotions to the musical generative system one dimensionally scores the film. It is currently never able to add new elements and contrasting emotional analysis to the scene. Finally, the systems music and mapping to the screen dictates a mood onto the scene. This in itself isn't necessarily a negative or positive but restricts the applications of *DeepScore*.

4.4 Potential Applications

To conclude the survey we presented a short video demonstrating a potential implementation of the system for general use (see Figure 8). This tool showed three characters from the film, and allowed the user to choose a melody for each one (from a list of three generated melodies) and then hear the film with those melodies. One composer outright dismissed using any assistance to compose, believing composing should be a personal humanistic activity. The other composers were more open minded to the tool, two in particular proposed being able to use the system to augment their own work and create new variations. Another composer mentioned they would use it to create quick demos to test ideas, but not to compose an entire project. Directors were generally interested in using the tool however the main concerns were the tool had to be easy to use and simple to modify, but most importantly better than the temp tracks.

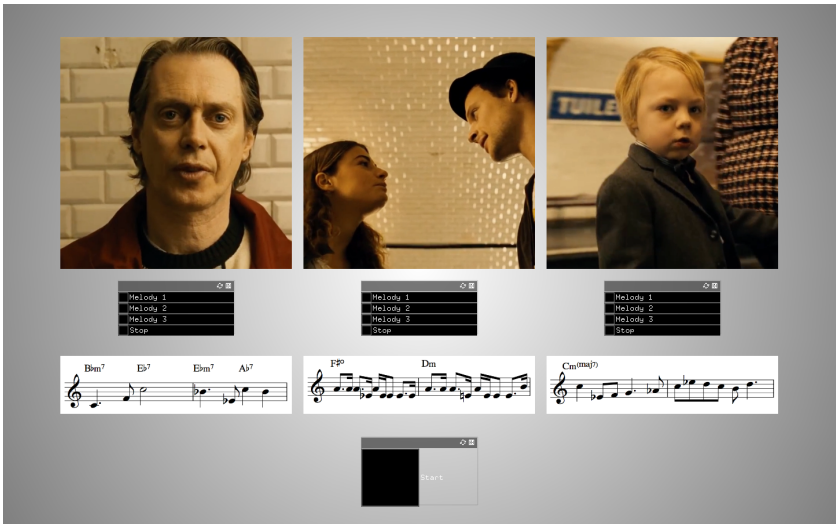

Fig. 8. Program Demonstrating Potential Application

5 Conclusion

In this paper we have described a software based film composer, built from successes and lessons learned while creating a robot film composer. There have been consistent links to visual analysis improving the relation of generated materials to the film. Linking to visuals not only improved the connection to the film, but also improved the rating and perception of the music's quality. General responses to the system showed that an emotional dialogue between score and visuals is central to connecting to the film. In future iterations this dialogue needs to become multiple dimensional, whereby emotions on the screen can be complemented in ways other than a direct musical response. Although only briefly analyzed, we also contend there are many applications for using visual analysis as a tool in film music generation for both composers and film creators.

References

1. S. Alizadeh and A. Fazel. Convolutional Neural Networks for Facial Expression Recognition. *CoRR*, abs/1704.0, 2017.
2. M. Bribitzer-Stull. *The modern-day leitmotif: associative themes in contemporary film music*, pages 255–300. Cambridge University Press, 2015.
3. J. Buhler, C. Flinn, and D. Neumeyer. *Music and cinema*. Published by University Press of New England for Wesleyan University Press, Hanover, NH, 2000.
4. F. Chollet. Deep Learning with Python. *Manning Publications*, 80(1):453, 2007.
5. J. Coen and E. Coen. Paris, je t'aime, 2006.
6. A. Goodfellow, Ian, Bengio, Yoshua, Courville and I. Goodfellow. Deep Learning. *MIT Press*, 2016.

7. A. Hill. *Scoring The Screen*. Hal Leonard Books, Montclair, NJ, 2017.
8. G. Hoffman and G. Weinberg. Gesture-based human-robot jazz improvisation. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 582–587, 2010.
9. N. Jaques, S. Gu, D. Bahdanau, J. M. H. Lobato, R. E. Turner, and D. Eck. Tuning Recurrent Neural Networks With Reinforcement Learning. 2017.
10. N. Jaques, S. Gu, R. E. Turner, and D. Eck. Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning. In *Deep Reinforcement Learning Workshop, NIPS*, 2016.
11. B. E. Jarvis. Analyzing Film Music Across the Complete Filmic Structure: Three Coen and Burwell Collaborations. 2015.
12. P. N. Juslin and J. A. Sloboda. *Handbook of Music and Emotion: Theory, Research, Applications*. 2010.
13. Kaggle. Challenges in Representation Learning: Facial Expression Recognition Challenge.
14. F. Karlin, R. Wright, and ProQuest (Firm). On the track a guide to contemporary film scoring, 2004.
15. D. P. Neumeyer. *Meaning and Interpretation of Music in Cinema*. 2015.
16. A. Simmons. Giacchino as Storyteller: Structure and Thematic Distribution in Pixar's Inside Out (. *Music on Screen*, (July), 2017.
17. G. Toussaint. The Euclidean Algorithm Generates Traditional Musical Rhythms. *BRIDGES: Mathematical Connections in Art, Music and Science*, pages 1–25, 2005.
18. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Arxiv*, 2016.
19. A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. *CoRR*, abs/1606.0, 2016.