Real-Time Synaesthetic Sonification of Traveling Landscapes

Tim Pohle
Dept. of Computational Perception
Johannes Kepler University
Linz, Austria
tim.pohle@jku.at

Peter Knees
Dept. of Computational Perception
Johannes Kepler University
Linz, Austria
peter.knees@jku.at

ABSTRACT

When travelling on a train, many people enjoy looking out of the window at the landscape passing by. We present an application that translates the perceived movement of the landscape and other occurring events such as passing trains into music. The continuously changing view outside the window is captured with a camera and translated into midi events that are replayed instantaneously. This allows for a reflection of the visual impression, adding a sound dimension to the visual experience and deepening the state of contemplation. The application can both be run on mobile phones (with built-in camera) and on laptops (with a connected Web-cam). We present and discuss different approaches to translate the video signal into midi events.

1. INTRODUCTION AND CONTEXT

Looking out of the window and watching the landscape passing by is a common thing to do on train journeys. Another popular activity is listening to music on a mobile player. However, the impressions from these two tasks – although they are often performed simultaneously – are not corresponding with each other, i.e. there is no synaesthetic experience. In this paper, we present an application that aims to create these synaesthetic experiences by capturing images of the outside and translating them to sounds that correspond to the visual impressions.

A major inspiration for this work was the music video for the track “Star Guitar” by “The Chemical Brothers” directed by Michel Gondry [1]. The video gives the impression of a continuous shot filmed from a passengers perspective on a speeding train. The train passes through visually rich towns, industrial areas, and countryside. The crux of the video is that all buildings and objects passing by appear exactly in sync with the various beats and musical elements of the track. While in this video the visual elements were composed based on the musical structure, in this work, we try to compose music in real-time based on the visual structure of the passing real-world landscape. For the resulting compositions, the elements surrounding the tracks can be considered the score which is going to be interpreted based on outside conditions such as weather and lighting, the speed of the train, and the quality of the camera. Thus, every journey will yield a unique composition.

In the past, several other approaches that aim at automatically composing music based on visual content have been presented. Most of them directly map the two dimensions of images onto two acoustic dimensions, i.e. the position of pixels on the y-axis is often interpreted as the pitch of the corresponding sound, while the x-axis is interpreted as time domain. A more sophisticated approach is presented in the work of Lauri Gröhn [2]. Based on a cell-automaton like concept, images are filtered by removing pixels in an iterative process. Different tracks for the compositions can be obtained by partitioning the image and different movements by applying slightly different graphical filters. A large number of impressive examples is made available on the Web site and the high number of on-line visits suggest that also that a wide audience is considering the results to exhibit some sort of synaesthetic correspondence. With our approach presented in this paper, we try to go even one step further by not only sonifying static images but real-time video instantaneously captured with a camera.

2. GENERAL IDEA

The general idea of our application is to capture the passing landscape with a camera and transform the visual data to sound, probably even to music. The landscape can either be recorded with the built-in camera of a mobile phone or with a Web-cam connected to a laptop. The data captured by the camera is given as a series of images (frames). Each frame is an array of pixels. From each captured frame, we take the middle column of the pixels and use this data to create and modify an audio stream.

The user interface of the application is divided into two main areas (cf. Figure 2). In the right half of the screen, the current frame is shown as delivered by the video camera. The left half contains a kind of history of the middle column of the picture. This history is updated at a constant rate. Every time a new frame is processed, the data contained in the left part is copied one column to the left. The (now empty) rightmost column closest to the red line is assigned the values from the last frame’s middle column.

3. TRANSFORMATION APPROACHES

For the suggested application, we tried out several approaches to use the video data for sound creation. These are presented in this section after briefly discussing the used color space models.

In this work, two color representations are used. First, in the (r, g, b) representation the red, green and blue components of each pixel are measured independently, and represented each as a value in the range [0..1]. The values measured by the camera are given in this representation. For a perceptually more meaningful representation, the (r, g, b) representation can be transformed into the (h, s, v) representation, where *hue* (i.e., color), *saturation* (i.e, color intensity, ranging from 0 which is white / gray / black to 1 which is "screaming" color) and *value* (related to perceived brightness, e.g. sun / shadow) are given independently.

3.1 Filter Approach

In the first approach to transform video data into sound, the (r,g,b)-values in the middle column of the current video frame are transformed to grayscale by taking their mean, so that each pixel has only one value associated instead of three. These values then are interpreted as the characteristic of an audio filter. The band associated with the bottom pixel (index $i = 0$) has a center frequency f_0, and the bands associated with the other pixels have center frequencies of integer multiples of f_0 (i.e., band i has center frequency $i \cdot f_0$).

Such a filter can be implemented by applying an inverse FFT (iFFT, inverse Fast Fourier Transformation) on the pixel values. The output values of the iFFT then are used as taps in a FIR filter (Finite Impulse Response). We use this filter to impose the desired spectrum on pink noise.

3.2 Piano Roll Approach

The second approach we tried is quite common. It is based on interpreting the middle column of the current video frame as a short fragment of a piano roll [5]. The piano roll was invented at the fin de siecle. It allows for operating player pianos without a pianist being present.

In our straightforward approach, the top pixels of the video frame column are associated with high pitches, and bottom pixels are associated with the lowest notes. To generate music, the interpreted data is sent to a MIDI instrument. The brightness (v-value) of a pixel is interpreted as the volume, while its color (h-value) is mapped to the available midi instruments (MIDI *program* number). To come closer to human perception, connected regions of similar color (or alternatively, similar brightness) are treated as one entity. To this end, we applied edge detection and region merging algorithms. If sound and pitch do not change in consecutive frames, no new MIDI event is generated.

3.3 Color-based Approach

Some synaesthetes have color associations when listening to sounds or tones. The Russian composer and pianist Alexander Nikolajewitsch Skrjabin was a synaesthete who created a mapping between piano tones and colors (cf. Fig. 1). His composition technique sometimes is referred to as a precursor of the twelve tone technique [4].

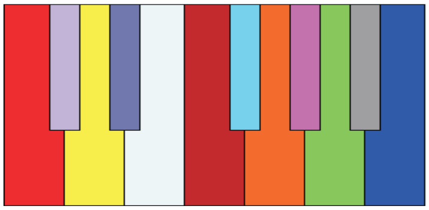

Figure 1: Tone-to-color mapping of Scriabin's Clavier à lumières (cf. [4,6])

We adopt this color mapping in the following way: As the range of the given mapping is only one octave, the middle pixel column of the current video frame is divided into $n = 4$ parts of equal height. Each of these parts then is used to generate tones played in a different octave. The pixels of a part are transformed to a pitch by calculating the cosine distance of each contained pixel's (h,s,v)-value to all of the twelve colors of the color piano. These values then are subsumed into a twelve-binned histogram over all pixels. The fullest bin is the pitch that is played in this octave. Additionally, if the second fullest bin is nearly as full as the fullest (cutoff value 0.75), then this tone is also played. The velocity is taken from the maximum v value of all pixels. The corresponding notes are played on a midi sound generator with a piano sound.

In many cases, colors do not change significantly between consecutive frames. To avoid repetitively playing such notes at every frame, these notes are held if the change is below a certain threshold. However, the piano has a sound that decays and vanishes after some time. Thus, this could result in a situation where all sound is gone, for example when the train stops at a station, or when the passing landscape does only change slightly. Therefore, if a note is constantly held for more than $m = 7$ frames, it is repeated. In some cases, this results in repetitive patterns that are perceived as musical themes. To avoid dominance of such patterns over the resulting overall sound, notes repeated this way are played with less and less velocity, until a minimum velocity value is reached, which is used for all consecutive repetitions.

4. DISCUSSION

In our implementation, the filter approach did not produce convincing results. Probably the most important thing is that the resulting sounds are noisy and whistling. Such sounds are associated with trains anyway, so producing them by technical means does not add so much to the experience already available without any equipment. Also, the implementation seemed not to be sufficiently fast for a real time usage on mobile devices since calculation of an iFFT for each frame together with the transitions between the frames turned out to be too computationally expensive.

The Piano Roll Approach is more promising. However, creating algorithms that reasonably map the regions perceived in the landscape by humans to sounds (both in the x- and y-Dimension) turned out to be a task beyond the scope of this work. Although we tried to reduce the amount of notes and note onsets by region finding algorithms and by holding non-changing notes, the resulting sounds are very complex even for landscapes with a very simplistic appearance.

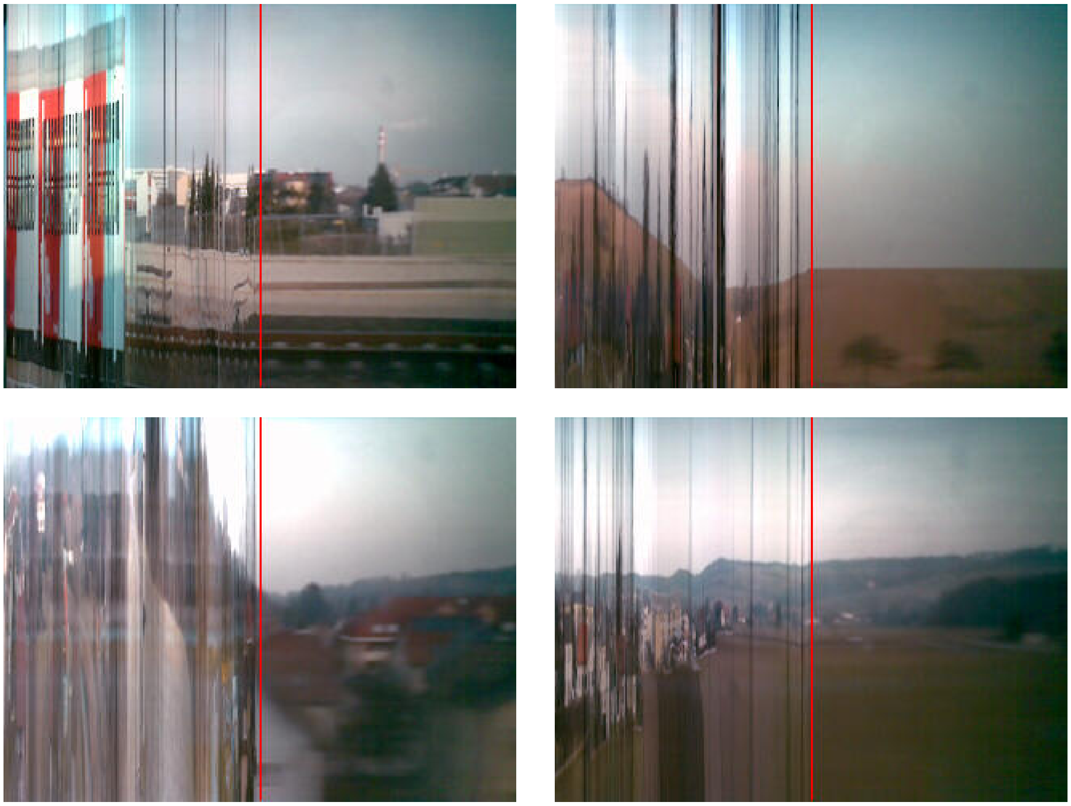

Figure 2: Four example screenshots taken from the mobile version of the software running on a Nokia 6120. The right half of the screen displays the current image taken from the camera. The left half consists of the sequence of recently sonified pixel columns. The left part also exhibits some interesting effects caused by the movement of the train. Since frame rate and position of the camera are both static, proximity of objects and slope and velocity of the train result in characteristic visual effects. For example, objects that "move" at high speeds are displayed very narrow, whereas objects filmed at low speeds appear stretched. Note that similar effects can also be observed using the tx-transform technique by Martin Reinhart and Virgil Widrich (cf. [3]).

The Color-based Approach yields in our opinion by far the best results. Due to a steady rate of seven frames per second, there is a clearly noticeable basic rhythm pattern in the music, which the listener may associate with the steady progression of the train. Depending on the landscape, notes in some bands are played in fast repetition or movements, while in other bands they sound only sporadic. The resulting harmonies are quite pleasurable, which might be a result of the color distribution in the mapping from colors to pitches. Also, a changing landscape is reflected in the resulting music, while the overall feeling remains the same. An example video of a sonified train journey sequence can be found at `http://www.cp.jku.at/people/pohle/trainpiano.wmv`.

5. ACKNOWLEDGMENTS

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number L112-N04.

6. REFERENCES

[1] Michel Gondry. Music video for "The Chemical Brothers – Star Guitar", 2002.

[2] Lauri Gröhn. Sound of Paintings. URL: `http://www.synestesia.fi/` (last access: 08-Feb-2008)

[3] Martin Reinhart and Virgil Widrich. tx-transform. URL: `http://www.tx-transform.com/` (last access: 08-Feb-2008)

[4] Wikipedia (English). Alexander Scriabin. URL: `http://en.wikipedia.org/wiki/Alexander_Scriabin` (last access: 08-Feb-2008)

[5] Wikipedia (English). Piano roll. URL: `http://en.wikipedia.org/wiki/Piano_roll` (last access: 08-Feb-2008)

[6] Wikipedia (German). Alexander Nikolajewitsch Skrjabin. URL: `http://de.wikipedia.org/wiki/Alexander_Skrjabin` (last access: 08-Feb-2008)