sound/tracks: Real-Time Synaesthetic Sonification and Visualisation of Passing Landscapes

Tim Pohle, Peter Knees, and Gerhard Widmer
Department of Computational Perception
Johannes Kepler University Linz, Austria
tim.pohle@jku.at, peter.knees@jku.at, gerhard.widmer@jku.at

ABSTRACT

When travelling on a train, many people enjoy looking out of the window at the landscape passing by. We present sound/tracks, an application that translates the perceived movement of the scenery and other visual impressions, 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 is intended to be run on both mobile phones (with built-in camera) and on laptops (with a connected Web-cam). We propose and discuss different approaches to translating the video signal into an audio stream, present different application scenarios, and introduce a method to visualise the dynamics of complete train journeys by "re-transcribing" the captured video frames used to generate the music.

Categories and Subject Descriptors: J.5 [Computer Applications]: Arts and Humanities

General Terms: Algorithms

Keywords: Mobile Music Generation, Sonification, Train Journey

1. MOTIVATION 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 activities – although they are often performed simultaneously – are not corresponding to each other, visual stream and sound stream remain separate entities with no logical or aesthetic connection. The goal of the project presented here is to develop an artistic application that combines visual and acoustic channels to create a unified travel experience – a synaesthetic experience where visual and acoustic impressions correspond to (or allude to) and reinforce each other, and that, at the same time, reflects the uniqueness of every individual journey. The main idea is to capture the images of the outside with a video camera and to translate them into sounds (music, in the best case) that complement the visual impressions, in real time. This allows for a reflection of the visual experience and deepening of the state of contemplation. The resulting application should be aesthetically pleasing and inspiring, both in the visual and in the acoustic domain, and should be usable on-line (i.e., while travelling) as well as off-line (e.g., to enable public exhibits or private re-playing and re-experiencing of trips).

A specific goal is to make this work available to everybody to use and enjoy during journeys or afterwards. To this end, the software we have developed for various platforms (specifically, PC and mobile phone – see below) will be made freely available to the general public.

A major inspiration for this work was the music video for the song "Star Guitar" by "The Chemical Brothers" directed by Michel Gondry [3]. The video gives the impression of a continuous shot filmed from a passenger's 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 that video the visual elements were composed based on the musical structure, we aim at achieving the opposite in our project: our goal is to give the passing landscape an active role, to make it the central player that controls the real-time production of music. In a sense, the passing landscape with its fleeting visual impressions acts as the musical score which is going to be interpreted based on outside conditions such as weather, season, lighting, the speed of the train, and the quality of the camera. Every journey is different and will yield a unique composition (in contrast to, e.g., Michel Gondry's video, which is and remains fixed).

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 position along the x-axis is interpreted as point in time. A more sophisticated approach is presented in the work of Lauri Gröhn [4]. 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

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

Figure 1: sound/tracks on a PC / laptop. The video is displayed in real time in the top panel, the condensed history of the journey (the sequence of recently sonified pixel columns) slowly moves past in the form of a 'score image' in the lower panel.

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 suggests that a wide audience is considering the results to exhibit some sort of synaesthetic correspondence. *Bondage* by Atau Tanaka [12] is an installation including the transformation of a displayed picture into sounds. The displayed picture is altered based on a camera's images of the visitors. The *Monalisa* project [5] is a work that builds on transforming sound into image data and vice versa. For example, software plugins offer functionality to apply sound effects to image data, and to apply image effects to sounds.

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 SCENARIO AND USER INTERFACE

The general scenario is as follows. The passing landscape is captured with a video camera and transformed into sound (music) instantanenously. 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) from which 7 frames per second are selected for further processing. Each frame is an array of pixels. From each chosen frame, we take the middle column of the pixels and use this data to create and control an audio stream. The resulting music is played back via MIDI and selected instrument samples (piano sounds, in our current implementation).

The video content itself is also displayed on the computer screen. To be more precise, the user interface of the application is divided into two main areas (cf. Figures 1 and 2). In the top part of the screen (or the right half, in the case of the mobile version), the current frame is shown as delivered by the video camera (i.e., the recorded video is played back in real time). The bottom (left) part 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, that is, 7 times per second, the history image is moved one column to the left, and the current central frame column is added to the right. The effect is a condensed history of the journey that slowly moves to the left. The details of this process, and the resulting effects, are explained in more detail in section 6. The video available at the project web site (*http://www.cp.jku.at/projects/soundtracks*) gives a more direct impression.

The technical part of this paper is divided into two parts, according to the two major components of the system: the real-time transformation of video to sound (see Section 3) and the creation and display of the condensed history – the 'score image' (Section 6). Section 7 then gives some technical details about the realisation and required hardware and software. Section 8 shows various modes of using and presenting the application, and Section 9, finally, offers a critical discussion of the current state of the project.

3. TRANSFORMING VIDEO TO SOUND

A central challenge in this project was to devise a general algorithm for translating aspects of a live video into musical structures in a meaningful way, and in a form that is in some way comprehensible to the user. The desiderata of a good video sonification include the following properties:

- The sonification should reflect general characteristics (complexity, lighting, liveliness, ...) of the passing scenery
- The sonification should be consistent in the sense of mapping similar conditions to similar sounds
- The sonification should sound musical (perhaps including a sense of rhythm, which is crucial to the common music listener)

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 video taken from the camera. The left half is the condensed history – the 'score image'.

- Ideally, the sonification approach takes into consideration also results of psychological studies on how humans generally associate visual impressions with sounds
- The sonification technique should not require too complex calculations to permit fast execution and video processing in real-time also on mobile platforms.

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 colour space models and some findings from psychological studies concerning associations of visual aspects and sounds.

3.1 Colour Space Models

In this work, two colour 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., colour), *saturation* (i.e, colour intensity, which is 0 for e.g. white, grey, and black and 1 for "screaming" colour) and *value* (ranging from black to full colour/full white, and being related to perceived brightness, e.g. sun vs. shadow) are given independently.

3.2 Psychological Findings

Although investigating a concept like *synaesthesia* is in general a very challenging task since it has to study very elusive phenomena, there exist a number of publications on humans' associations between visual perceptions and sound. Experiments find that there is an association of *brightness* with musical scales, modes and pitch height [1]. Datteri and Howard find an association of the frequency of pure sine tones and *colour frequency* [2]. Rusconi et al. point out that some people associate pitch height with *spatial height* [11]. Thus, it seems that there are associations of pitch height with all of brightness, colour hue, and spatial height. For an application like the one we are building, there is the question of how these can be combined into an overall function that maps visual impressions (light) into (combinations of) sounds. Such a function may exist, or may not exist. As mentioned by Datteri and Howard [2], Marks finds that some participants associate an increase of brightness of grey surfaces with increase in loudness of pure tones, while others associate an increased loudness with a decrease of brightness, and suggests that most participants associate visual brightness with auditory brightness [7]. This diversity might be an indication that a universal function to translate visual impressions into sound that generally matches human perception may not exist.[1]

4. TRANSFORMATION APPROACHES

To transform the passing landscape into sound, visual information is extracted from the captured images. To keep the amount of information to process small and to add to the idea of "scanning" the passing landscape, we examine only one pixel column per frame. We chose to process the central column because it is the least distorted one (in terms of perspective and optical distortion due to the camera lens). To deal with various video stream formats from different camera models, i.e., streams with different frame rates, we capture all incoming frames but process images only at a constant rate of 7 frames per second. This also determines the sound's underlying beat structure. In the following, we describe the different transformation approaches. Schematic representations of all methods are depicted at the bottom of Figure 3.

[1]Interestingly, there are also indications of object-sound associations. For example the nonsense word *bouba* containing rounded vowels is associated with rounded shapes, while words with unrounded vowels such as *kiki* are associated with angular shapes [8].

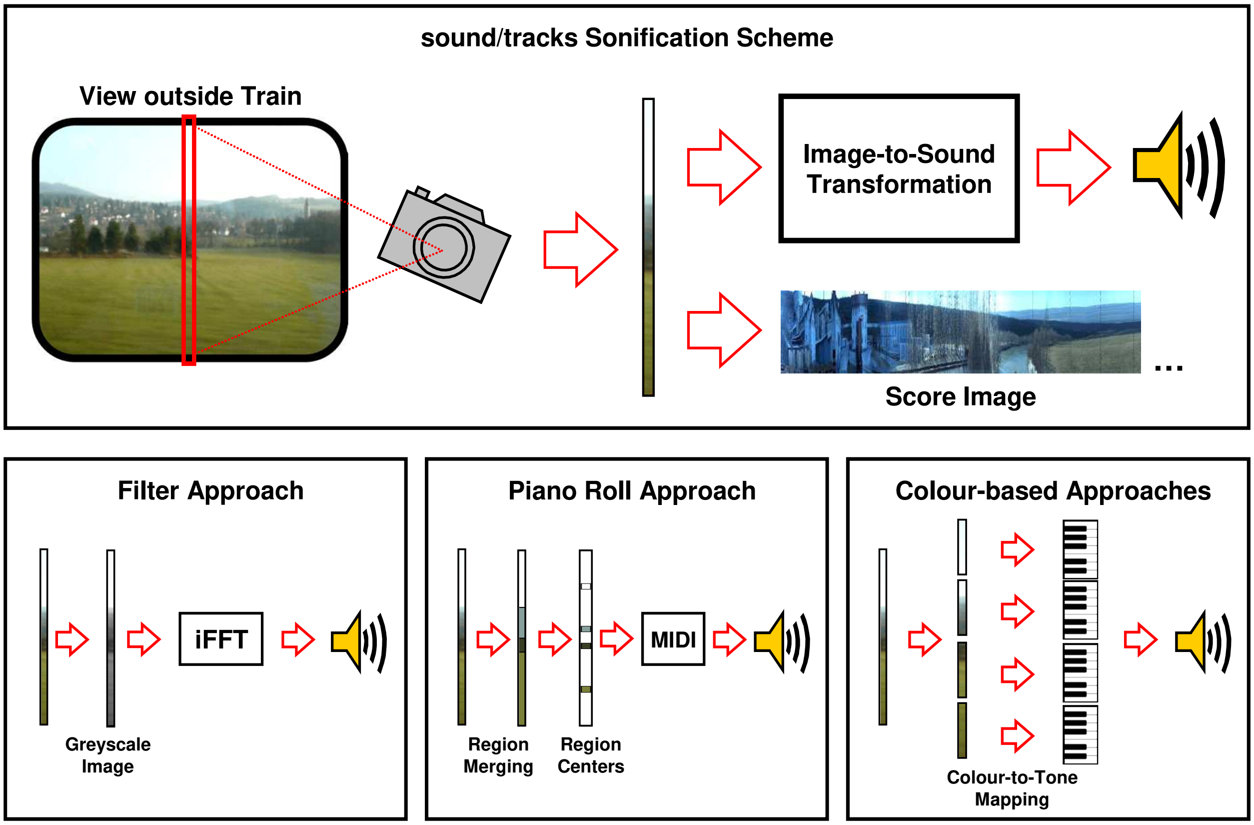

Figure 3: Top: Basic process of generating sound from the view outside the train. The central pixel column is extracted from the currently captured video frame. This pixel column is then used for two things: to generate sound using a translation function and to construct a panoramic scan of the journey by sequentially joining consecutive pixel columns. Bottom: Schematic overviews for the three image-to-sound transformation approaches.

4.1 Method 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 greyscale 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 $f_i = i \cdot f_0$). To better match the human ear's nearly logarithmic frequency perception, the frequency mapping is done in an exponential way ($f_i = f_0 \cdot exp(k_1 \cdot i + k_2) + k_3$, with k_1, k_2 and k_3 being constants that are chosen appropriately). We achieve this by rescaling the image data, interpreting each of the 128 initial input bands as one semitone. Band $i = 0$ is associated with A (27.5 Hz or 900 cent),

The filter is realized by applying an inverse Fast Fourier Transformation (iFFT) on the pixel values. The output values of the iFFT then are used as taps in a Finite Impulse Response (FIR) filter. We use this filter to impose the desired spectrum on white noise. With a sampling rate of 22050 Hz, the filter has 1605 taps. Transitions between consecutive frames are smoothed by interpolating between the associated filters. Due to the characteristic of the image data after transformation to the log domain the resulting sounds have a $\frac{1}{f}$ characteristic. $\frac{1}{f}$ noise (pink noise) is the type of noise frequently found in real-world phenomena. It is perceived as having equal magnitude in all frequency bands. A spectrogram of a sound generated with Method 1 is shown in Figure 4.

4.2 Method 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 (a common way to visualise MIDI events in sequencers). The piano roll was invented at the end of the 19th century. 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 colour (h-value) is mapped to the available MIDI instruments (MIDI *program* number). To come closer to human perception, connected regions of similar colour (or alternatively, similar brightness) are treated as

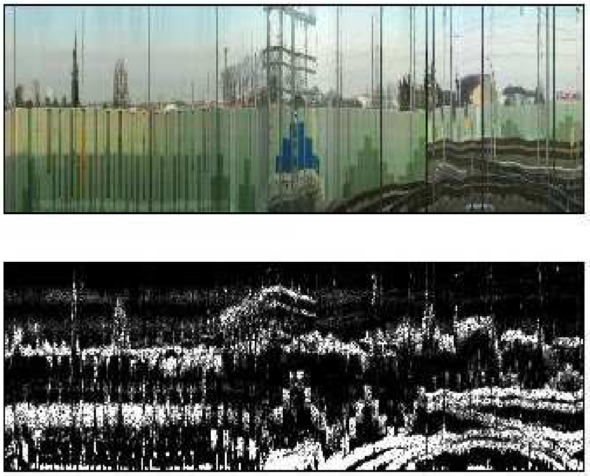

Figure 4: Filter approach. Top: Short excerpt of a journey as taken by the camera (all middle columns of the processed frames). Bottom: Spectrogram of a generated sound.

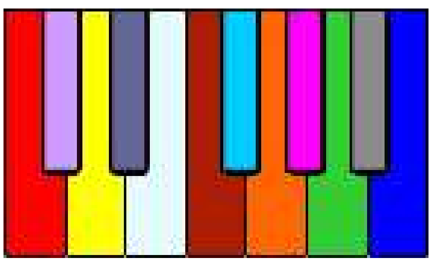

Figure 5: Tone-to-colour mapping of Skrjabin's Clavier à lumières.

one entity. To this end, we explored basic edge detection and region merging algorithms (e.g., based on the variance within each region). If sound and pitch do not change in consecutive frames, no new MIDI event is generated. The implemented algorithm is of rather general nature, allowing for trying out a variety of different settings.

4.3 Method 3: Colour-based Approaches

Instead of using the colour information to control timbre (i.e., to select a MIDI instrument), here, we transform colour information to pitches. We tried two variants: the simpler one is historically inspired, while the more elaborate approach incorporates findings from psychological studies on preferred pitch. Both approaches are discussed in the next sections.

4.3.1 Historically Inspired Approach

Some synaesthetes have colour associations when listening to sounds, tones, or chords. The Russian composer, pianist, and self-claimed synaesthete Aleksandr Nikolajevich Skrjabin created a mapping between piano tonalities and colours (cf. Fig. 5) [13].

We adopt this colour 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 (cf. rightmost scheme at bottom of Figure 3). The pixels of a part are transformed to a pitch by calculating the cosine distance of each pixel's (h,s,v) value to all of the twelve colours of the colour 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, colours 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.

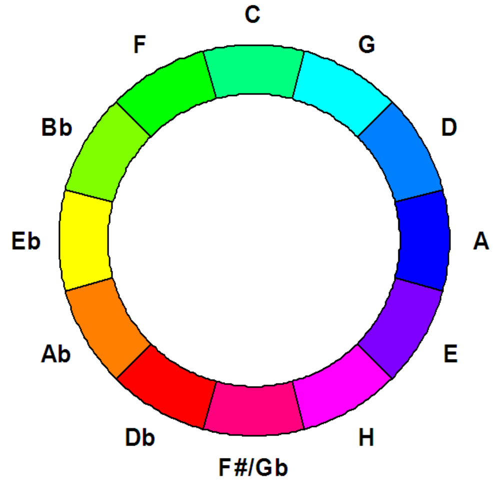

Figure 6: The basic colour mapping used in the perceptually motivated approach.

4.3.2 Perceptually Motivated Approach

When listening to the produced music, it becomes evident that the approach presented in the previous section sounds better when using the cosine *distance* between colours than using the (cosine or Euclidean) *similarity* (i.e., the inverse distance, which would correspond to Skrjabin's mapping). The latter produces rather random-sounding note sequences. To use a *similarity* measure and still obtain music that fits the way western listeners' hearing is conditioned, we superpose a tonality on each frame by determining a 'root node'. The outline of the process is as follows:

1. Use all pixels in the current frame's middle column to determine the root note.

2. Split the column into $n = 4$ parts, and create a 12-binned pitch activation histogram for each. More saturated colours should result in a higher tonal richness.

3. Finally, each of the $n = 4$ found histograms is additionally weighted with the pitch class prominence. This weighting indicates how well each pitch "fits" the root node and the chosen mode (major or minor). We use pitch class prominence values from [6] (as reproduced by [9]) in minor mode.

Step 1: The frame's root note is determined by creating a histogram of pitches over all pixels in the frames's middle column. Each pixel p_i contributes amounts to all histogram bins based on the distance of the pixel's (h,s,v) values and the bin's (h,s,v) values, and additionally weighted by the pixel's s and v values: $w_i = \sqrt{s_i \cdot v_i}$. The amount a pixel p_i contributes to bin b_j thus is given by $a_{i,j} = w_i \cdot (1 - d(p_i, b_j))$, where d denotes the Euclidean distance between the (h,s,v) values of the pixel and the bin's (h,s,v) values in the hsv space represented as a colour cone. For each bin, all activations are summed up to an overall activation value A for bin j: $A_j = \sum_i a_{i,j}$. The histogram bin j with maximum activation A_j is assumed to be the current frame's root note. The weighting is motivated by the idea that "clearer" colours (i.e., more saturated and bright colours) should contribute more to the chosen note than dark and grey colours. To avoid strong unexpected root note transitions between consecutive frames, each frame's root note is only allowed to be one step away from the previous frame's root note on the circle of fifths. If the distance is larger, only one step on the circle is made per frame.

Step 2: To superpose a tonality on the frame, the *saturation* component of colour is mapped to harmonic *richness*. More saturated colours should create a wider spectrum of allowed pitches, while for completely unsaturated pixels (i.e., greyscale), only the root note pitch is allowed. Thus, in general, higher harmonic richness is associated with higher colour intensity. This is realised by first calculating the similarity $s_{i,j}$ of each pixel's value to each of the keyboard colours (cf. Figure 6) $s_{i,j} = 1 - d(p_i, b_j)$, where d again denotes the Euclidean distance between the (h,s,v) values of the pixel and the bin in the hsv space represented as a colour cone. The superposition of tonality is done by subsequently changing bin/pitch j based on the pixel's s value to a (in most cases different) bin k. j is changed to a pitch that is closer (more closely related on the circle of fifths) to the root note according to the factor determined by the pixel's s value. An s value of 1 (fully saturated colour) produces no change of the bin (i.e., $k = j$), while an s value of 0 (white / grey / black) always changes the bin to the root note bin (i.e., $k = \texttt{rootnote}$). For example, if the pixel's s value is 0.5, the root note is C, and the current bin j is orange (Ab, cf. Figure 6), then the bin k, where the calculated amount is added, is A#. Denoting this bin transformation function as $k = t(j, p_i)$, the final activation A of bin b_k in band / octave B is $A_{k,B} = \sum_{i \in B} \{s_{i,j} | t(j, p_i) == k\}$, where b is the current band ($B = 1..n$).

In *Step 3*, the tonality is further emphasised by weighting the histograms with perceptually based factors that reflect how well a pitch 'fits' into the current tonality. We use the well-known tonal pitch profiles that were determined in psychological experiments by Krumhansl and Kessler [6]. The final pitches to be played are selected after this weighting is applied to the histograms.

5. EVALUATION OF THE SONIFICATION ALGORITHMS

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 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 expensive computationally.

The Piano Roll Approach is more promising. However, creating algorithms that reasonably map regions as 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 that look very simple.

The Colour-based Approaches yielded by far the best results in our opinion. 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 sporadically. The resulting harmonies are quite pleasant, which might be a result of the colour distribution in the mapping from colours to pitches. Also, a changing landscape is reflected in the resulting music, while the overall feeling remains the same. For this reason, the colour-based approach was adopted for this project in the end. Several examples of sonified train journey sequences can be found at our project web site at *http://www.cp.jku.at/projects/soundtracks*. To get an impression of the different sonification methods and allow for comparison, we have prepared a dedicated web site at *http://www.cp.jku.at/projects/soundtracks/methods.html*.

6. VISUALLY SUMMARISING A JOURNEY: THE SCORE IMAGE

The scenery passed during a train journey can be considered the underlying 'musical score' of the composition generated by sound/tracks. Sonification of the journey is thus an interpretation of this score based on outside conditions such as weather and lighting, the speed of the train, the quality of the camera, and the degree of staining of the window glass. As mentioned, the landscape is scanned by sequentially analysing the central column of the captured video frames, at a constant rate of 7 frames per second. When re-transcribing these columns back to an image (i.e., generating an image by joining the sequence of analysed columns along the x-axis), one obtains a panoramic overview over the scenery passed so far. Such a panoramic image captures and displays the *dynamics* of the journey and 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 narrowly, whereas objects filmed at low speeds appear stretched. That becomes particularly apparent during stops at train stations:

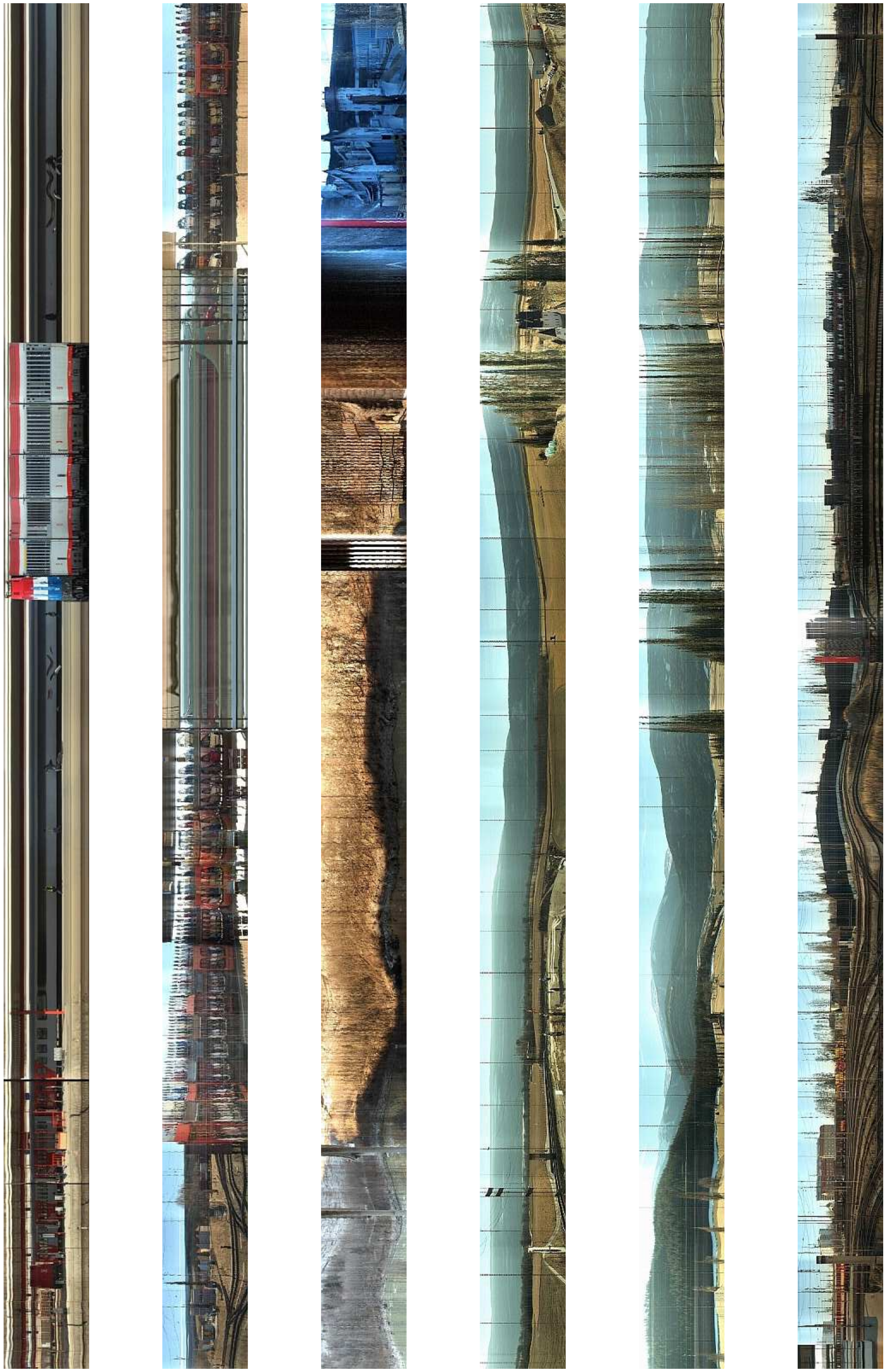

Figure 7: Six short excerpts from visual 'score images'.

the static background is blurred; only moving people and passing trains change the score and thus the panoramic image. Clear illustrations of that effect can be found in Figure 7, which depicts short excerpts of score images from several train journeys across Austria. The two top-most images nicely illustrate the stretching effects during a stop. The third image gives some impressions from changing light conditions during a trip. The fourth and fifth image show landscapes from the alpine region. The score image at the bottom, finally, visualises the departure from a train station and exhibits some interesting patterns caused by the numerous other rail tracks.

In the sound/tracks interface, a score image containing the recent score history is always displayed in addition to the currently captured video (cf. Figures 1 and 2). These generated score images permit to persistently capture and archive the fleeting impressions of the journey and also allow for re-experiencing the complete trip visually at a later point. Using the score image, it is also possible to regenerate the composed music from the journey. Hence, the score image allows also for re-experiencing the trip acoustically. Additionally, at least in our opinion, these score images have a visual appeal of their own and will make nice complements to demos when being displayed as printouts on long paper strips.

7. TECHNICAL REALISATION

We have implemented two versions of sound/tracks – one for laptops and one for mobile phones. The laptop version is implemented in Java for easy portability and uses Java's built-in MIDI Sound Synthesiser to generate the sound output. The mobile version is currently being developed and optimised for Symbian S60 3rd Edition enabled devices. The user interface is implemented in Python and builds upon C++ or Java to access the mobile device's MIDI synthesiser.

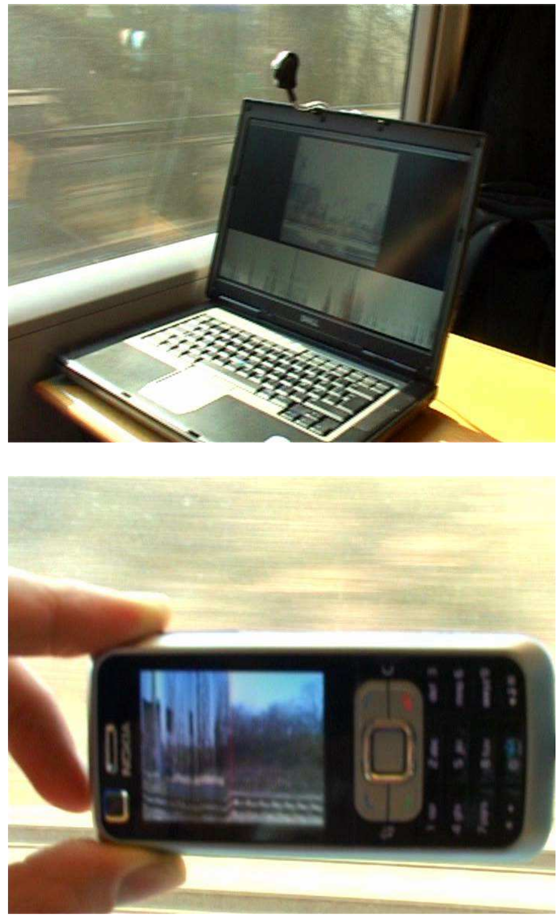

Figure 8: sound/tracks live on the laptop (top) and a Nokia 6120 mobile phone (bottom).

8. APPLICATION AND PRESENTATION SCENARIOS

In this section, we will highlight possible usage scenarios for sound/tracks, either for personal enjoyment during train journeys or for public presentations.

8.1 Private Enjoyment

The most immediate application scenario is (of course) to use sound/tracks directly when being on a train journey. Playing sound/tracks live and translating the observed scenery to music instantaneously yields the maximum "synaesthetical" effect. To this end, sound/tracks can be used during a journey – both on laptop with a camera attached (e.g., a high-quality Webcam), and on appropriate mobile phones (see Figure 8). sound/tracks can also be used to re-play and re-experience trips off-line, after the journey, by augmenting the recorded video playback with music and moving score image generated from the recorded video itself.

Note that we plan to make the sound/tracks software (both the PC and the mobile phone version) freely available to the general public, via our project web page.

8.2 Public Installations and Performances

Train journeys that have been recorded at high quality (e.g., with a DV camera) can be adapted for performances in public space, e.g. as a multimedia exhibit or even a visually augmented piano concert. In such a public demonstration, the recorded video and the generated score image could be projected on separate walls and sonification of the trip could be performed live and in high-quality on a computer-controlled grand piano instead of synthesised instruments. To get an impression of the sound quality of train journeys sonified via a grand piano, the reader is invited to listen to a selection of audio recordings, all performed by the computer on a Bösendorfer CEUS Computer-controlled Imperial Grand Piano, on the project web page. The outlined installation setting is illustrated schematically in Figure 9.

To complement the audio/video aspect of the exhibit, entire journeys in the form of printouts of score images could be glued around the walls of the exhibition room to document train journeys off-line. Figure 7 shows six short excerpts form such 'score images' from different journeys across Austria. Each of these covers 96 seconds of travelling. Thus, a two hour trip, for instance, will produce a visual summary of about 15 metres length (when printed out at 300 dpi).

9. DISCUSSION

We have presented an application that produces a soundtrack for a train journey in real-time based on the passing scenery outside the train. In order to find a method that accomplishes the translation from visual input to auditory output in a preferably intuitive and synaesthetic manner, we

Figure 9: A possible scenario for a public installation (sketch).

proposed and discussed several approaches. From our point of view, an approach that incorporates both spatial information and colour distribution properties of captured video frames to generate piano music, yielded the most pleasing results. In addition, we proposed a method to visualise and document the dynamics of a train journey. The resulting panoramic images can also be interpreted as the scores of the generated musical compositions and add a further dimension to the overall aesthetics of the project.

Several people that were confronted with resulting videos (i.e., recordings of train journeys played in sync with the generated music and the recent score image) enjoyed watching and listening for an unexpectedly long time. People seem to appreciate and enjoy the result of the sonification, which may even reflect synaesthetic aspects of perception to some extent (supposing that that is possible at all). However, the chosen approach represents just the current state of the project as we will further explore alternative image-to-tone translation approaches in the next development steps. Finally, with the planned release of the developed software for laptops and mobile phones, we aim at providing a novel and interesting gadget that adds another dimension to people's train journeys and makes them more enjoyable.

10. ACKNOWLEDGMENTS

Research underlying this work is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under project numbers L112-N04 and L511-N15. Special Thanks are due to L. Bösendorfer Klavierfabrik GmbH, Vienna for giving us the opportunity to record selected tracks on a CEUS Computer-controlled Imperial Grand Piano and for granting use of the piano image in Figure 9.

11. REFERENCES

[1] W. G. Collier and T. L. Hubbard. Musical scales and brightness evaluations: Effects of pitch, direction, and scale mode. *Musicae Scientiae*, 8:151–173, 2004.

[2] Darcee L. Datteri and Jeffrey N. Howard. The sound of color. In *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8)*, Evanston, IL, USA, August 3-7 2004.

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

[4] Lauri Gröhn. Sound of Paintings. URL: `http://www.synestesia.fi/` (last access: 24-Apr-2008).

[5] Kazuhiro Jo and Norihisa Nagano. Monalisa: "See the sound, hear the image". In *Proceedings of the 8th International Conference New Interface for Musical Expression. NIME '08*, pages 315–318, Genova, Italy, June 5–7 2008.

[6] C. L. Krumhansl and E. J. Kessler. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89:334–368, 1982.

[7] L. E. Marks. On associations of light and sound: The mediation of brightness, pitch, and loudness. *American Journal of Psychology*, 87 (1-2):173–188, 1974.

[8] Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. The shape of boubas: sound–shape correspondences in toddlers and adults. *Developmental Science*, 9:3:316-322, 2006.

[9] H. Purwins. *Profiles of Pitch Classes Circularity of Relative Pitch and Key - Experiments, Models, Computational Music Analysis, and Perspectives.* PhD thesis, TU Berlin, 2005.

[10] Martin Reinhart and Virgil Widrich. tx-transform. URL: `http://www.tx-transform.com/` (last access: 24-Apr-2008)

[11] Elena Rusconi, Bonnie Kwan, Bruno L. Giordano, Carlo Umiltà, and Brian Butterworth. Spatial representation of pitch height: the SMARC effect. *Cognition*, 99:2:113–129, 2006.

[12] Atau Tanaka. Bondage. URL: `http://www.xmira.com/atau/bondage/` (last access: 30-Jul-2008), 2004.

[13] Wikipedia. Alexander Scriabin. URL: `http://en.wikipedia.org/wiki/Alexander_Scriabin` (last access: 24-Apr-2008).