

numediart

Research Program in Digital Art Technologies - QPSR Vol. III - No. 2

Quarterly Progress Scientific Report

Vol. 3, No. 2, June 2010

T. Dutoit, B. Macq (Editors)



Published online by:

Université de Mons (UMons)

Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS)

<http://tcts.fpms.ac.be>

Université Catholique de Louvain (UCLouvain)

Laboratoire de Télécommunications et Télédétection (TELE)

<http://www.tele.ucl.ac.be>

ISSN:

2032-5398 (printed)

2032-538X (online)

Credits:

Editors: Thierry Dutoit (UMons/TCTS), Benoît Macq (UCL/TELE)

Cover photo: Christian Graupner

\LaTeX editor: Christian Frisson (UMons/TCTS), using \LaTeX 's confproc class (by V. Verfaillie)

All copyrights remain with the authors.

numediart homepage: <http://numediart.org>

Contact: contact@numediart.org

Preface

numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631). Its main goal is to foster the development of new media technologies through digital performances and installations, in connection with local companies and artists.

numediart is organized around three major R&D themes:

- **HyFORGE** - Hypermedia Navigation: Information indexing and retrieval rely classically on constrained languages to automatically describe contents and allow formulating queries, respectively. This approach becomes hardly applicable for multimedia contents such as music or video because of the disparity between computable low-level descriptors and desired high-level semantics - the so-called semantic gap. Alternatively, HyFORGE investigates human-in-the-loop approaches and innovative tools for structuring and searching multimedia contents. Along with audio and image processing, HyFORGE builds up on self-organizing models to derive enhanced views of multimedia collections and provide users with efficient browsing interfaces.
- **COMEDIA** - Body & Media: COMEDIA is named from a French contraction between body and media or stage director and media, which nicely sums up the main objective of this axis: giving to bodies the means to be their own artistic director! Hence based on position on stage or choreography between multiple artists for the inter-relationship and gestures or voice for the intra-relationship, COMEDIA aims at creating interactivity between performing artists and the multimedia context around. Event description, low-level feature analysis, pattern recognition, heterogeneous sensor fusion, robustness against lighting and real-time are our keywords in 1D, 2D and 3D signal processing to reach these goals.
- **COPI** - Digital Instruments Design: COPI aims at developing a software/hardware toolbox for creating innovative digital musical instruments, from scratch or by augmenting existing instruments with new interactive channels. The main challenges for this R&D axis are to produce expressive instruments which maintain a close, embodied relationship with the musician. Our approach is to produce new sound design architectures using a large database of pre-recorded signals while maintaining real-time control of the design process. Our scientific work therefore implies three main axes: the development of expressive production models (audio signal processing), followed by the design of gestural control systems for their synthesis parameters, coupled with statistical modeling of this dynamic control.

numediart is the result of collaboration between Polytech'Mons (Information Technology R&D pole) and UCL (TELE Lab), with a center of gravity in Mons, the cultural capital of Wallonia. It also benefits from the expertise of the MULTITEL research center on multimedia and telecommunications. As such, it is the R&D component of MONS 2015, a broader effort towards making Mons the cultural capital of Europe in 2015.

This tenth session of numediart projects was held from April to June 2010. The session ended with a public presentation of the results (with demonstrations) in the newly inaugurated numediart Room, on Friday, July 2nd, 2010.



The video clip of [Ghinzu](#)'s latest single, "Cold Love", is just out. Directed by the "Satisfaction" collective, composed of John Israel, Anthony Collard and Francois Jacques, the video contains a ground breaking sequence, in which the artists literally catch fire during their performance. These images have been made possible thanks to the involvement of the NUMEDIART Institute, who provided support for motion capture. The animation, by [F. Jacques](#), required 5 computers working 24/7 for three months...

Projects

Session #10 (Apr-Jun 2010)

- 25 Project #10.1: IVISIT: Interactive Video and Sound Installations Tools. Application to the BorderLands project
Todor Todoroff, Xavier Siebert
- 33 Project #10.2: AudioGarden: towards a Usable Tool for Composite Audio Creation
Christian Frisson, Cécile Picard, Damien Tardieu

IVISIT: INTERACTIVE VIDEO AND SOUND INSTALLATIONS TOOLS – APPLICATION TO THE BORDERLANDS PROJECT

Todor Todoroff¹, Xavier Siebert²

¹ Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS), Université de Mons (UMons) and ARTeM, Belgique

² Laboratoire de Mathématique et Recherche Opérationnelle (MathRo), Université de Mons (UMons), Belgique

ABSTRACT

This project aims at providing tools to develop interactions with a collection of prerecorded videos and reactive sounds. These tools allow the detection of similar frames among a collection of video, the detection of visitor’s behavior in a given setup, and the use of the latter information to trigger or govern video navigation and audio interaction. This project results from an ongoing collaboration with the media art project "BorderLands" [2] by Christian Graupner.

KEYWORDS

Video Analysis, Sensors, Sound, Installation

1. INTRODUCTION

This project aims at solving several scientific or technological challenges, with a direct application to the “BorderLands” project. The first goal of this project is to find similar frames (also called “intersections” or “branching points” in this context) between videos, as represented in Fig. 1. Christian Graupner calls this collection of interleaved video sequences the *Movie Time Space* or, in short, MTS.

These intersections are used in the associated artistic project (BorderLands [2]) where a video is played in an installation, switching from one sequence to the other depending on the visitor’s behavior. The branching points are used to make these transitions seamless.

The second goal of this project is to detect visitor’s movements using sensors, and to further use this information to navigate among the video sequences and to trigger or control sounds interactively.

2. RESULTS

2.1. Video Analysis

Several features were used to compare two video frames and to find intersections between them, as detailed below. Depending on the specific requirements of the artistic project, any combination of features could be used.

2.1.1. Scale-Invariant Feature Transform

Several algorithms were developed over the last decade to detect interest points in images, notably SIFT [8] and SURF [1]. We used an implementation of the SURF algorithm based on OpenCV (*find_obj.cpp*), which consists of an extraction of the interest points (or key points) followed by a matching of these key points.

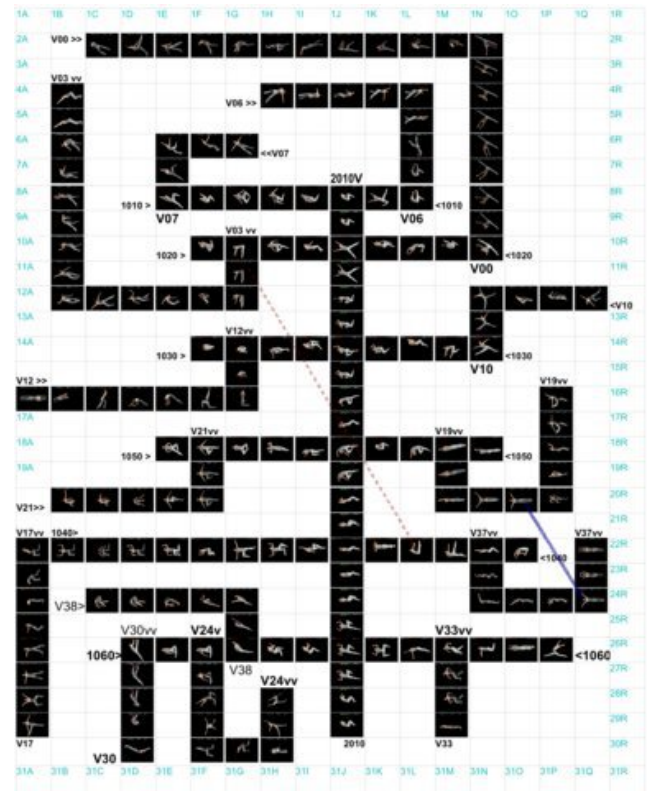


Figure 1: Schematic of interleaved video sequences, the *Movie Time Space*, with branching points at each crossing. Additional lines are needed to show them all, as a 2-D space does not allow to display them all as crossings.

Fig. 2 shows that similar images have indeed more matching points than dissimilar ones, confirming the validity of this technique. However, the algorithms to extract and then compare features are time-consuming, making the comparison between all pairs of frames from all videos not tractable as such. Future work in this direction could include GPU implementations [5, 19] or pre-segmentation of the videos, which will probably be tackled in subsequent numediart projects.

2.1.2. Hu Moments

Another widely used technique to detect features independently of scale, translation and rotation of an image are the Hu moments [6]. We calculated the Hu moments (based on their implementation in

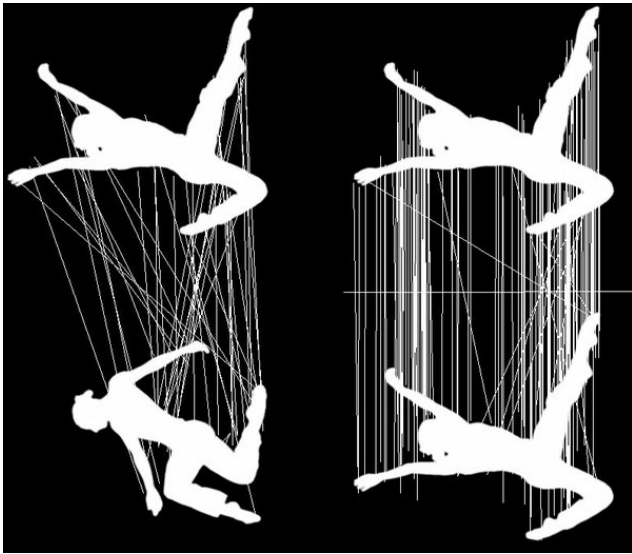


Figure 2: Matching of interest points between two frames using the SURF features. Similar images (right) have more corresponding points than dissimilar images (left).

the OpenCV library [4]) of all frames in two videos. Then we calculated the distances between two images A and B as follows:

$$d(A, B) = \sum_{i=1 \dots 7} \frac{|m_i^A - m_i^B|}{|m_i^A|} \quad (1)$$

where

$$\begin{aligned} m_i^A &= \text{sign}(h_i^A) \cdot \log h_i^A \\ m_i^B &= \text{sign}(h_i^B) \cdot \log h_i^B \end{aligned} \quad (2)$$

and h_i^A, h_i^B are the Hu moments of A and B , respectively.

The resulting distance matrix between pair of frames from two videos (labeled 10151 and 20102, respectively) is shown on Fig. 3.

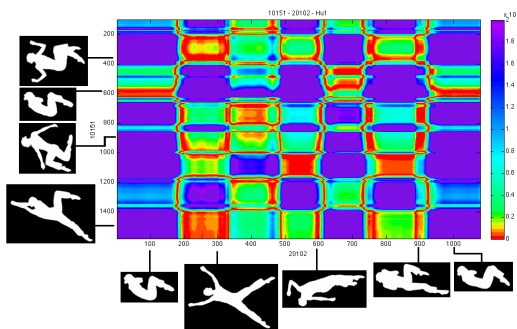


Figure 3: Distance matrix representing the similarities between two videos in terms of the first Hu moment. The colors indicate the similarity, as indicated in the distance scale. For instance, red color represents similar frames (small distance). Some representative frames are shown in black and white, in the margin.

Fig. 3 reveals that the Hu moments allow the detection of some frames which are indeed very similar. In the following, let us note

(x, y) the x^{th} frame of video labeled 20102 and the y^{th} frame of video labeled 10151. Among the couples detected as similar (red zones) we find some matches that appear as correct, for example $(100, 600)$ or $(1000, 600)$, with the dancer in a fetal position. However, couples such as $(300, 280)$ are also detected as similar, although it is more difficult to understand why considering the difference between the position of the dancer in these frames. These results led us to investigate another descriptor based on Fourier methods.

2.1.3. Fourier descriptors

Fourier methods have proven to be efficient for shape-based image retrieval [20]. Here we used a Fourier transformed in polar coordinates, that yields results that are translation-independent. Due to the properties of the Fourier transform, frames related by a 180 degrees rotation are also going to be considered similar. Rotation independence could be incorporated using the Fourier-Mellin transform (see for example [21] and references therein).

The resulting distance matrix between pair of frames from two videos (labeled 10151 and 20102, respectively) is shown on Fig. 4. The results appear to be qualitatively better than those obtained by the Hu moments approach. Indeed, the frames detected as similar (red zones on Fig. 4) generally correspond to images that are fairly similar, either directly, as in $(100, 550)$, or after a 180 degrees rotation $(100, 830)$.

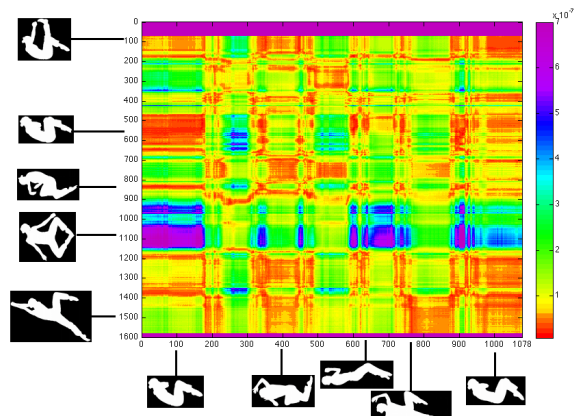


Figure 4: Distance matrix representing the similarities between two videos in terms of the Fourier Polar moment. The colors indicate the similarity, as indicated in the distance scale. For instance, red color represents similar frames (small distance). Some representative frames are shown in black and white, in the margin.

2.1.4. Software

A software application was developed to conveniently analyze the intersections between videos. The core video analysis algorithms were developed within the mediacycle framework [14]. The Graphical Interface was developed using Qt's Phonon module¹. A snapshot of the beta version of the software is shown on Fig. 5.

¹<http://doc.trolltech.com/4.6/phonon-module.html>

2.2. Generation of sound from movements inside the video sequences

As described before, video sequences were analyzed and features were extracted frame by frame. Each frame was then annotated with the following parameters: center of gravity, bounding box width and height, contraction index (the number of active pixels within the bounding box), bounding box ratio and Hu moments.

In the context of the installation, video sequences can be played back at different speeds as well as reversed, and different branching points will be chosen depending on the interactions with the visitors. Therefore the derivatives of the features between subsequent images cannot be imbedded in the annotation of the video but need to be computed live. We also computed the smoothed derivatives as seen in Fig. 6.

A selection of those features was mapped to prerecorded sound parameters as well as to synthesis models in order to generate sounds evolving synchronously with the movements of the dancer in the video.

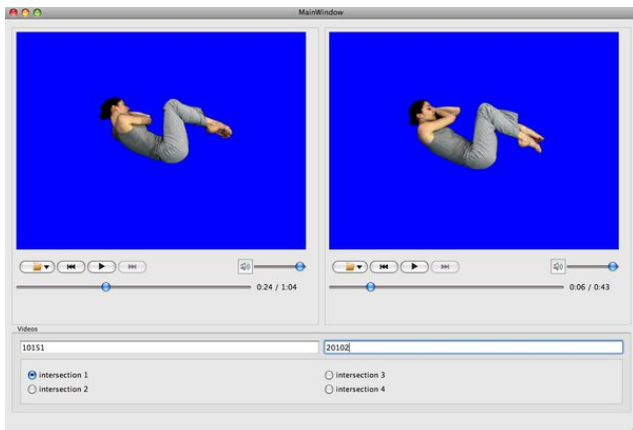


Figure 5: Snapshot of the software to analyze intersections between videos.

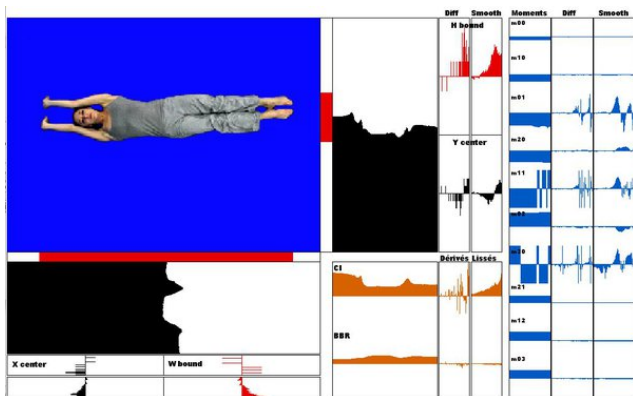


Figure 6: Snapshot of the Max patch showing annotated analysis parameters: center of gravity (black), bounding box (red), contraction index and bounding box ratio (orange) and Hu moments (blue), as well as their respective derivatives and smoothed derivatives.

2.3. Visitor behavior sensing

As the setup for BorderLands is to be some kind of pit, with the image at the bottom and the visitors interacting with their hands on and around the balustrade (see Fig. 7), we had to find a system whereby the hand positions of the visitors could be monitored while keeping the system hidden for aesthetic reasons.

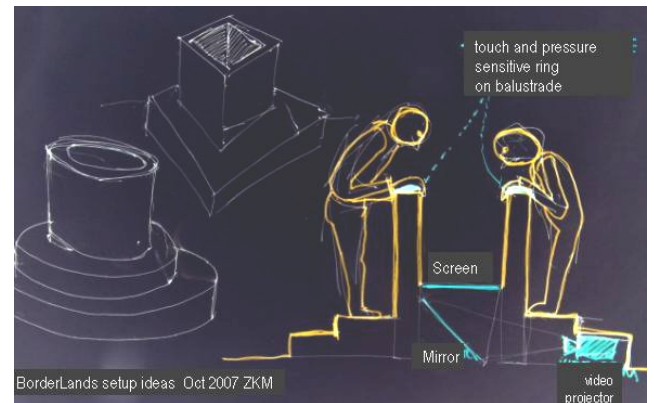


Figure 7: Sketch of the BorderLands setup by Christian Graupner.

We first envisioned to use a combination of theremin antennas and pressure sensors on the balustrade to monitor the movements and pressure of the visitor's hands. But the extent of the surface of interaction makes it difficult and expensive to cover with pressure sensors. A possible solution would have been to create a rigid frame on top of the balustrade that would rest solely on pressure sensors on its four corners in order to sense the centre or gravity of the visitor's hands pressure. We used a similar system for a 3m by 3m interactive surface for a dancer in *Mes Jours et mes Nuits* [12]. Or to have a piecewise frame, each side of it being mechanically independent from the others and resting on pressure sensors at the extremities, something we used in the performance *Twelve Seasons* [11]. This would give the center of gravity of the interaction on each side independently from the others. All those solutions are feasible, though the rigidity of the top construction needed in order to avoid the pressure to be passed to the main structure somewhere in the middle of one side (if the flexibility of the top panel would allow it to touch the structure) suggests using a metal frame that would get in the way of Theremin antennas.

And the Theremin principle, as we explain further, allows to "feel" the hands even before they touch the surface, which is a strong argument in their favor. Moreover, as the sensitivity of the Theremins increases as the hand reaches closer to the surface, we could add a layer of medium-density foam on top of the antennas. As pressure on the foam shortens the distance between the hand and the antenna, we can effectively create a pressure sensor of arbitrary shape at a low cost. We therefore decided to go for a mix of theremin antennas. As a dedicated microprocessor controlled electronic circuit for an eight antenna Theremin had been developed at ARTeM for previous installation projects, we adapted it for BorderLands.

We had thought at some point of using a 3D camera to detect the height of the hands, but decided that we had already enough information of that kind from the Theremins. Additional data made indeed no sense for the video navigation, due to the limited amount of decisions that can be made at a given time: continue the same video sequence, go backwards, bifurcate left or right. And the

use of the physical model layer we describe below gives us plenty additional parameters to drive the sound generation in a continuously interesting way. This choice gives the additional advantage of having a totally self-contained structure that could be installed anywhere without the need to suspend a camera.

2.3.1. Theremin principle

The Theremin holds a special place in the history of music instruments, as the first instrument ever to be played without touching it, and as one of the first electronic ones. It is named after its Russian inventor, Lev Termen or Léon Theremin. He built a high frequency oscillator at the Physical Technical Institute in Petrograd in order to measure the dielectric constant of gases with high precision. When he added circuitry to generate audio tones, he discovered that the pitch was changing as he moved his hand around. He named his instrument the *etherphone* but it was later renamed after him.

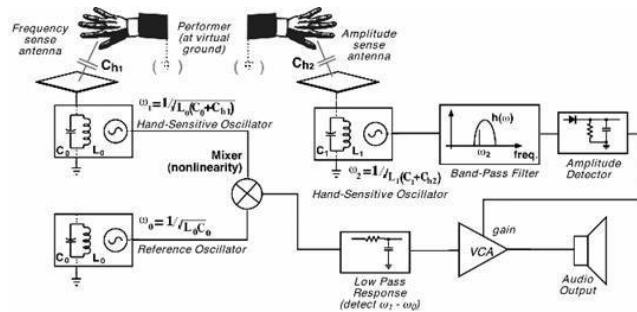


Figure 8: Heterodyne principle of the Theremin pitch antenna leads to high precision (picture from [13]).

A variable oscillator is controlled by the performer's hand distance from the antenna as the hand acts as a grounded plate of a variable capacitor in the oscillator circuit. The influence of the hand position, and hence its distance from the antenna, could be measured as a deviation of the oscillator frequency, but the percentage of the deviation compared to the oscillator center frequency would be very small and hence difficult to measure with precision. In order to amplify the influence of the performer's hand, Leon Theremin used the heterodyne principle: the difference frequency between that variable oscillator and a fixed oscillator tuned around the same frequency (350 kHz in the original design) is the audio frequency heard by the listener. That way, even a 0.05% change in the variable oscillator can be substantial at audio frequency, which is enough, with good design, to give a range of five octaves to the instrument.

The original heterodyne Theremin principle is shown in Fig. 8, taken from the Theremin Center webpage [16], lead by Andrey Smirnov [15] and located at the Moscow State Conservatory.

2.3.2. Theremin design used for the project

In our case, we were not looking for visitors to directly generate Theremin-like sounds with their interactions. We rather wanted to have distance information from each antenna and process that information to control the sound and the navigation within the MTS in more flexible ways. The difference frequency therefore needs not be in the audio frequency domain; it just needs to be measured by a microprocessor. We use a hardware frequency divider to bring

the frequency measurement back to the measurement of the time between two successive low to high transition of a square wave signal. This represents in fact the period of the signal which can be easily measured thanks to the capture detection available on most microprocessors. We tune the variable oscillator associated to each antenna so that the difference frequency increases when the hand approaches, and so that the maximum time between the transitions is 1 ms in the absence of the hands. As the hand approaches, the frequency rises and the period decreases, insuring that an antenna measurement never takes more than 1 ms to be completed, so that the maximum latency is under control. We use a 16-bit timer, reset at the first transition and reaching its maximum count after 1 ms. We can thus measure a period ranging from 0 to 1 ms with a 16-bit resolution. A multiplexing scheme is used to measure the frequency of each antenna sequentially. Only the variable oscillator of the antenna under measurement is activated, leaving the others turned off to reduce the risks of mutual influences.

The perfect tuning of the variable oscillators may be quite difficult to reach, even with precision multi-turn adjustable potentiometers. As a precise tuning is the key to the maximum sensitivity of the antennas, we replaced the fixed oscillator in our design with a digitally controlled oscillator, allowing for a software fine-tuning of the fixed oscillator frequency, independently for each antenna. Fig. 9 shows the ARTeM 8-antennas Theremin electronics.

2.3.3. Communication between Max and the Theremin hardware

Data transfers between the microprocessor and the host computer are done by means of UDP packets. Max receives the eight Theremin difference periods every 10 ms, using a custom protocol decoded by a Max external. Fig. 10 displays the evolution of the 8 antenna difference periods over time. Max also sends configuration data to the hardware. Fig. 11 shows the tuning patch for the fixed oscillator frequency associated to each antenna.

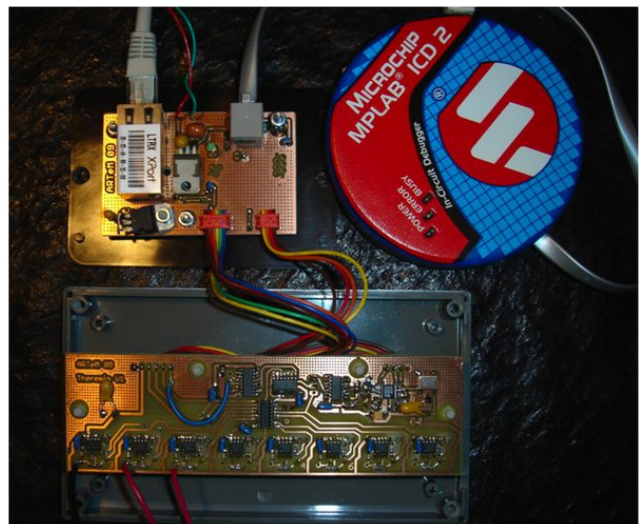


Figure 9: Prototype of the microprocessor board with Ethernet Interface, Theremin circuitry with eight variable oscillators (at the bottom of the PCB), digitally controlled oscillator and multiplexers, shown with the Microchip Pic programmer.

2.3.4. Antenna design

In order to obtain the most useful information about the visitors hand gestures over the installation, we tested various antenna designs as shown on Fig. 12. The antennas were made by cutting aluminum foil in various patterns. We chose to have two antennas on each side and tested left versus right and periphery versus center designs. More interleaved designs led to cross-influence between the antennas as the capacity between them increased. If we want to detect the left-right or periphery-centre position, the optimal design is a compromise between more interleaved antennas and wider gaps between them to limit the capacity and the mutual influence between them. As we choose for a virtual physical

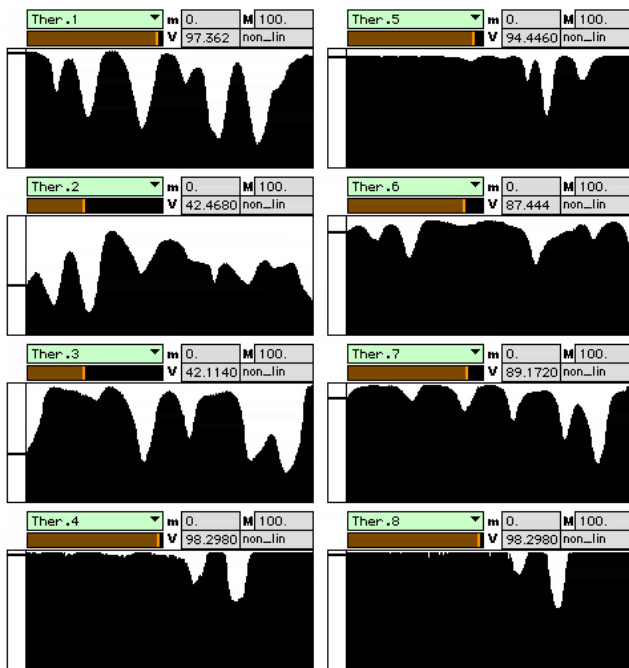


Figure 10: The evolution of the values of the measured period for each of the 8 antennas, as they appear in Max/MSP. Antennas are paired on each side of the prototype: 1 with 5, 2 with 6, 3 with 7 and 4 with 8.

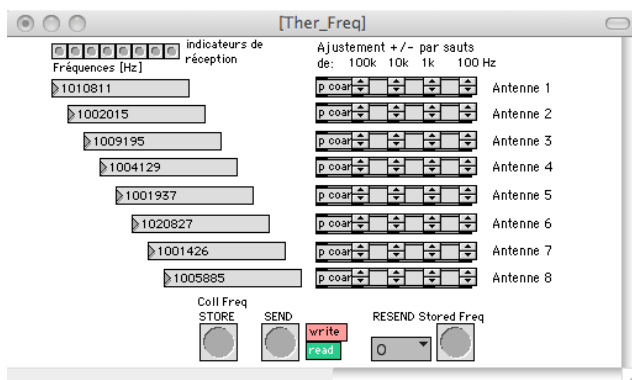


Figure 11: Max patch to fine-tune the digitally controlled fixed oscillator individually for each antenna, in order to reach the maximum sensitivity.

model layer, it became less necessary to be able to find the position in between the pair of antennas, rather than creating a clearly perceived difference when the hand moves.

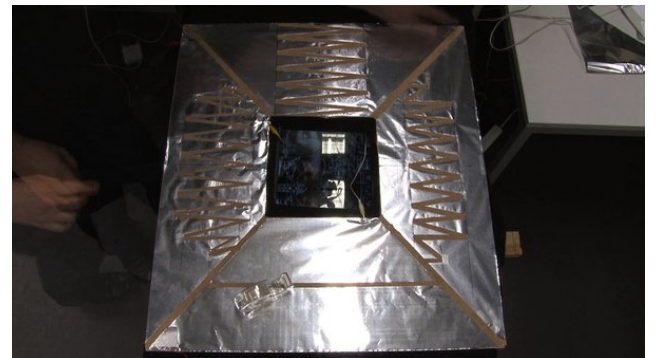


Figure 12: Test of different pairs of antenna designs, aimed at detecting either movements from left to right or movements from the periphery to the centre.

2.4. Physical model layer

For the project, it is important to devise a system that allows the user to feel instantly that he is interacting, thanks to the immediate response in the sound. But at the same time we wished to avoid a completely repeatable interaction scheme, where going back to the same hand positions would result in exactly the same sound output. We aimed at avoiding such a "fader-like" approach of the interaction in order to add more dynamical complexity to the response and to allow for contextual changes. We have also to avoid that one visitor keeping his hands on the balustrade would prevent other visitors to interact in a meaningful manner. The way we choose to achieve that was to insert physical models as an intermediate layer between the Theremin sensing devices and the sound mapping and video navigation.

The immediacy of the sound response is important for the visitors to feel his control over the system, because the principle of the video navigation, that imposes to finish the display of a video sequence until the next branching point before a bifurcation choice can be made, is by nature not immediate. Another layer in the image though, the NET as shown on Fig. 13, that reacts to the shape and movements of the dancer within the video sequences, can and will also respond immediately to the movements of the visitors. A physical model layer also helps the NET to behave in a way that will be perceived as more fluid and natural.

box2d Physics Engine [3] is a widely used open source engine for simulating rigid bodies in 2D. Initially developed on Windows using Visual C++ by Erin Catto, from the Massachusetts Institute of Technology, it is available on many platforms and languages like C#, Java, Python, AS3. It has interesting features like continuous collision detection and there are ports for other popular programming environments like Flash, Max/MSP or Processing. We used the library of Max objects developed by Charles Bascou and Mathieu Chamagne [9], also available for PureData.

We chose a topology of four bodies, corresponding each to one of the four sides of the interactive surface and receiving opposite forces from the corresponding pair of antennas shown in Fig. 12. The visitors apply increasing forces to the bodies as their hands move closer to the surface.

Those virtual bodies are attached to the ground by mean of elastic joints, so as to bring them back to a rest position when the visitor's hands move away from the Therman antennas. In order to have a system where there is a mutual influence between the movements of the hands of several visitors at one time, so that one visitor cannot completely "steal" the effect associated to a pair of antennas by keeping his hands on the surface, they are also attached through elastic joints to an additional "center" body as can be seen on Fig. 14. This is done in a way that induces rotations of the "center" body. This rotation is furthermore amplified by the adjunction of 2 satellite bodies (in green). This system moves progressively from fluid movements as the hands are moved slowly to more hectic behaviors as the hands move fast with brusque changes of direction. These fast movements generate collisions between bodies that can be monitored on a per-body base.



Figure 13: BorderLands NET layer reacts to the dancer video sequences and to the visitors.

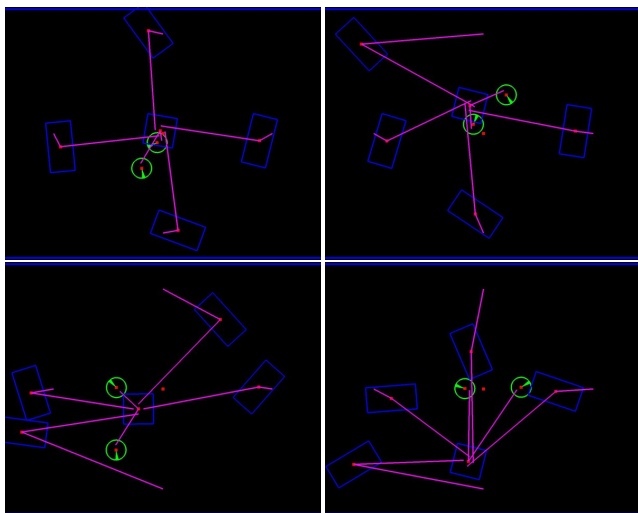


Figure 14: 4 views of a box-2D model where each pair of therman antennas send opposite forces to one of the 4 virtual bodies on the corresponding side. Clockwise, (a) the system close to rest, (b) one of the antenna sends a force on the blue top body, (c) collisions between the 2 green satellites bodies with 2 bodies attached to pairs of antennas and (d) collision of two blue bodies.

2.5. Interactive sound

We have eight channels of direct data from the visitors movements in the form of the periods of the eight Therman antennas. They can be used as such if needed, for example in moments when we want the visitors to be able to scrub through a video sequence. But we have a lot more information from the virtual physical bodies: position, angle, speed as well as their eventual collisions. This later data is, by nature, not static: after a movement by the visitor, the system moves and oscillates until it finds a new equilibrium with the forces corresponding to the new hand positions. It creates a dynamic flux of data that can be efficiently mapped to the sound algorithms using the techniques described in the *Dancing Viola* project (Numediart #04.2) [18] [17]. Collision detection on specific bodies, a threshold on the collision count during a defined time period or the exceeding of a speed threshold are all possible and valid ways to provoke a context change, in which sounds and the way to control them is modified. This allows to navigate between various states of interaction in a rather naturally perceived manner. Sonifying the collisions also creates subtle or dramatic effects depending on the chosen sounds. Mapping the angles of the satellite bodies on sonic properties create a feeling of evolution that is influenced indirectly but naturally by the movements as the speed of change is dependent on the level of activity of the visitor. Finally, those context changes can also apply to the sound layer generated from the movements found in the video sequences as explained in section 2.2, a sound layer that exists even in the absence of visitor interaction.

2.6. Navigation within video sequences

The realtime rendering of the video is done with the *AniMiro* software application, developed by André Bernhardt at *reactiveshop GmbH*, in Karlsruhe, Germany. *AniMiro* runs on customized Linux systems using OpenGL for rendering and supports multi-camera and multi-screen projections. Its ability to capture and interpret data from sensors and its palette of realtime effects allow it to generate the NET deformations for BorderLands directly from the content of the video frames as well as from Therman data received through OSC messages. Each frame of the video sequences is encoded in jpeg so as to avoid key frames in order to be able to play them freely forwards or backwards.

A system to navigate within the MTS could be implemented using a hidden Markov Model (HMM) whose transition probabilities are given by distance matrixes (see Fig. 4). However, this cannot be applied directly to the artistic project for technical reasons. Indeed, this project uses hand-made morphing to smooth the transitions between videos. These morphing involve several frames around the branching point itself, which would interfere with an instantaneous branching between two frames depending on the distance matrix. Further work need to be done to achieve the design of a navigation system specific to the artistic project. In the meanwhile, the conditions used to change the context for the sound interactions and the level of visitor's activity can be fed, using OSC messages, to the software developed by humatic [7]. That software, developed under the supervision of Christian Graupner, allows indeed to perform a navigation process by defining, independently for each video sequence, conditions to jump to other video sequences at precise time codes.

3. PERSPECTIVES

Annotations using features-based segmentation will be examined in an upcoming numediart project. The mapping of sensors to sound parameters through a physical model was successful and would merit to be investigated further. Attempts should be made to formalize such mapping for the wireless Numediart MARG (Magnetic, Angular Rate, and Gravity) sensors.

4. ACKNOWLEDGMENTS

We would like to thank Christian Graupner for his collaboration through his “BorderLands” project. We would also like to thank the ZKM [10], Karlsruhe, Germany, where Christian Graupner received the support from the Institute for Visual Media, and Todor Todoroff the support from the Institute for Music and Acoustics.

numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631).

5. REFERENCES

- [1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. “SURF: Speeded-Up Robust Features”. In: *9th European Conference on Computer Vision*. Graz, Austria 2006. P.: 25.
- [2] *Borderlands Project - Humatic*. Borderlands Project –Humatic. URL: <http://worx.humatic.net/p/BL2010/>. P.: 25.
- [3] box2d.org. URL: <http://www.box2d.org/>. P.: 29.
- [4] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. Cambridge, MA: O’Reilly, 2008. P.: 26.
- [5] S. Heymann et al. “SIFT implementation and optimization for general-purpose GPU”. In: *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. Citeseer, 2007. P.: 25.
- [6] Ming K. Hu. “Visual Pattern Recognition by Moment Invariants”. In: *IRE Transactions on Information Theory* IT-8 (1962). Pp. 179–187. P.: 25.
- [7] humatic. URL: <http://www.humatic.de>. P.: 30.
- [8] David G. Lowe. “Object Recognition from Local Scale-Invariant Features”. In: *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*. Washington, DC, USA: IEEE Computer Society, 1999. ISBN: 0769501648. URL: <http://portal.acm.org/citation.cfm?id=850924.851523>. P.: 25.
- [9] box2d for Max and PureData. URL: <http://charles.bascou.free.fr/box2d/>. P.: 29.
- [10] Zentrum für Kunst und Medientechnologie. URL: <http://www.zkm.de/>. P.: 31.
- [11] Michèle Noiret, Todor Todoroff, and Paolo Atzori. “Twelve Seasons”. 2001. URL: http://www.michele-noiret.be/index.php?page=f_proj_t. P.: 27.
- [12] Michèle Noiret, Todor Todoroff, and Fred Vaillant. “Mes Jours et mes Nuits”. 2002. URL: <http://www.michele-noiret.be/index.php?page=55>. P.: 27.
- [13] Joseph A. Paradiso and Neil Gershenfeld. “Musical Applications of Electric Field Sensing”. In: *Computer Music Journal* 21.2 (1997). Pp. 69–89. P.: 28.
- [14] Xavier Siebert et al. “MediaCycle: Browsing and Performing with Sound and Image libraries”. In: *QPSR of the numediart research program*. Ed. by Thierry Dutoit and Benoît Macq. Vol. 2. 1. numediart Research Program on Digital Art Technologies. 2009. Pp. 19–22. URL: http://www.numediart.org/docs/numediart_2009_s05_p3_report.pdf. P.: 26.
- [15] Andrey Smirnov. URL: <http://asmir.info/about.htm>. P.: 28.
- [16] ThereminCenter. URL: <http://theremin.ru/>. P.: 28.
- [17] Todor Todoroff et al. “Extension du Corps Sonore - Dancing Viola”. In: *Proc. NIME '09*. Pittsburgh, Pennsylvania 2009. Pp. 141–146. P.: 30.
- [18] Dancing Viola. URL: <http://www.numediart.org/projects/04-2-dancing-viola/>. P.: 30.
- [19] C. Wu. *SiftGPU: A GPU implementation of scale invariant feature transform (SIFT)*. 2007. P.: 25.
- [20] D. Zhang and G. Lu. “Shape-based image retrieval using generic Fourier descriptor”. In: *Signal Processing: Image Communication* 17.10 (2002). Pp. 825–848. ISSN: 0923-5965. P.: 26.
- [21] H Zhang et al. “Image and Vision Computing; Construction of a complete set of orthogonal Fourier–Mellin moment invariants for pattern recognition applications”. In: (2009). ISSN: 02628856. DOI: 10.1016/j.imavis.2009.04.004. P.: 26.

AUDIOGARDEN: TOWARDS A USABLE TOOL FOR COMPOSITE AUDIO CREATION

Christian Frisson¹, Cécile Picard², Damien Tardieu³

¹ Laboratoire de Télé-détection et Télécommunications (TELE), Université catholique de Louvain, Louvain-la-Neuve, Belgium

² Freelance researcher / pl-area

³ Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS), Université de Mons, Belgique

ABSTRACT

This project presents a new approach to sound composition for soundtrack composers and sound designers. We propose a tool for usable sound manipulation and composition that targets sound variety and expressive rendering of the composition. We first automatically segment audio recordings into atomic grains which are displayed on our navigation tool according to their timbre. To perform the synthesis, the user selects one recording as model for rhythmic pattern and timbre evolution, and a set of audio grains. Our synthesis system processes then the chosen sound material to create new sound events based on onset detection of the recording model and similarity measurements between the model and the selected grains. A large variety of sound events such as those encountered in virtual environments or other training simulations.

KEYWORDS

MediaCycle, Multimedia Databases, Content-based Navigation, Interfaces, Sound Design, AudioGarden

1. INTRODUCTION

Soundtrack composers and sound designers aim at creating auditory experiences [2]. In order to produce soundtracks for movies or video games, Foley artists mainly rely on prerecorded sound material, or record it themselves. While the use of prerecordings is easy to implement, the number of samples in a database is often limited due to memory constraints. Another possibility to generate such sounds is sound synthesis. A large variety of synthesis methods exist, but each of them is usually more suited for a reduced range of sounds. A very common technique for texture synthesis is the data driven concatenative synthesis, also referred to as mosaicing [8]. Concatenative synthesis approaches aim at generating a meaningful macroscopic waveform structure from a large number of shorter waveforms. They typically use databases of sound snippets, or grains, to create a given target phrase. Unlike granular synthesis where no analysis is performed on the audio units and where the unit size is defined arbitrarily [7], concatenative synthesis selects the audio units according to a set of audio descriptors. Physical modeling can be introduced to further refine granular synthesis [3, 1]. A very important issue for applications of granular synthesis to sound design is the control of the synthesis process. Vocem, introduced by Lopez et al. [5], is one of the first graphical interfaces for real-time granular synthesis, with high-quality audio output and very short latencies. Parameters allow the user to easily control the creation and the distribution of the grains. With MoSevius, Lazier et al. [4] first attempt to apply unit selection to real-time performance-oriented synthesis with direct and intuitive controls based on descriptor values such as energy, spectral flux or spectral centroid, as well as voicing and instrument name. For a more musical context, Misra et al. [6] focus on a single framework

that starts with recordings and proposes a flexible environment for sonic sculpting in general. Another class of control methods relies on a wise visualisation of the grains database in order to adequately select them. In Catart, Schwarz proposes to display the grains in a two-dimensional space according to descriptor values or output of dimension reduction techniques such as multidimensional scaling analysis or principal component analysis [8]. Following these ideas, we propose an approach that combines hypermedia navigation and a synthesis process into an adequate multimodal user interface for sound composition and design.

Our specific contributions are:

- a method for automatic analysis of audio recordings, extraction and classification of meaningful audio grains as new database.
- a technique for automatic synthesis of coherent soundtracks based on the arrangement of audio grains in time.
- a usable interface for database manipulation and sound composition.

2. CREATING SOUNDS

2.1. Synthesis Method Overview

The synthesis method is based on content driven concatenative synthesis. The main idea is to segment a target sound, that will be used has a time structure model and timbre evolution model, and to replace each segment by a grain contained in a database. This method involve four steps: segment the target, extract feature from the target segments and the grains, choose the grains that will replace each target segment and finally concatenate the chosen grains to create the final sound.

2.2. Sound Segmentation

2.2.1. Method

Sounds are segmented by finding the local minima of the spectral flux. This simple method allows to find onsets quite reliably. One problem is that onsets do not always coincides with energy minima resulting in a segmentation after the actual beginning of the sound and to small clicks in the synthesis. An alternative method would be to segment by finding the local minima of the energy of the sounds. This method results in less clicks in the synthesis but the rhythm feeling can be lost in the synthesis because energy minima do not correspond to perceived onsets. So the best solution would probably be to segment using energy minima and keep onset information (obtained using spectral flux) as a feature to be used during the synthesis process. The synthesis method thus needs to be adapted.

2.2.2. Segmentation In Mediacycle

Segmentation utilities have been added to Mediacycle (see Fig. 1). First, each media can now have children represented by a vector of pointers to media. Second, the support for segmentation plugins has been added, allowing various kind of segmentation depending on the type of media and on the application. The role of the plugins is to make all the computation allowing to find the segment boundaries and then to add these segments to the children vector. We developed one such plugin that implement the segmentation method previously described.

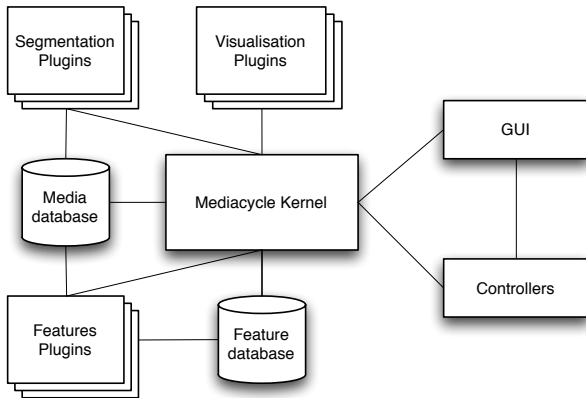


Figure 1: Mediacycle Architecture

2.3. Audio Features

To describe the grains and the target segments, the following features are computed:

- MFCC : to describe the spectral envelope of the sounds,
- SFM : to describe the noisiness of the sounds,
- Duration,
- 10-point temporal envelope interpolation.

2.4. Choosing Grains

The choice of the grain that will replace a segment is a very important issue in the synthesis process. It raises the problem of similarity measurement, and further the problem of similarity measurement in context. That is, if one chooses a grain to replace the first segment this will impact the subsequent similarity measurements. An other example of the kind of problem that can arise is given in figure 2. The grains and the target segments are displayed in a 2 dimensional feature space. In this example, if we choose the grain by similarity, the same one will always be chosen whatever is the segment. So we have to provide either transformation of the feature space or mapping function that allows consistent choice of the grains.

We propose three different methods:

- subtract the mean of both feature sets,
- subtract the mean of both feature sets and normalize the standard deviation,
- do nothing.

Those are very simple and further research still has to be done, but the second one, for instance, give very good results when the target and the grains are drum sounds from different drum kits. By normalizing the mean and standard deviation, the sounds of similar items of the different drum kits (snare drum, kick drum ...) end in the same region of the space, so the grain choice is very consistent (example).

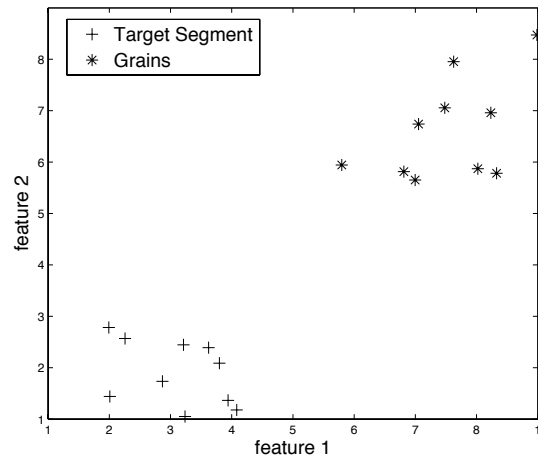


Figure 2: Example of similarity problem

2.5. Adding Grains

After selecting the grains, we need to add them in order to create the final synthesized sound. First of all we apply a tapered window on the grain in order to guarantee the smoothness of the composition. Then we propose three different addition methods:

- Simple: Grains are positioned at the same position as the original segment,
- Squeezed: Grains are concatenated, so rhythmic properties of the target is not preserved,
- Padded: for each segment, the closest grain is added, then if this first grain is shorter than the segment, the second closest is concatenated.

In the first and third cases, grains iteratively added to the current sound, starting from an empty sound, to allow superposition between grains.

3. USER INTERFACE DESIGN

3.1. Prototyping with mockups and storyboards

As there are very few computerized systems or analog practices that propose a workflow similar to the method we described here, we had to design a user interface fed by our own creativity. To achieve a certain level of mutual understanding of what we believe to be a suitable design, we produced throughout several brainstormings many mockups of the visual user interface and a storyboard of the expected scenario of usage, as illustrated in Figure 3, using paper [9] or whiteboards (shooting backups with cameras).

Drawing mockups prevented us from diving directly into the implementation of software prototypes, particularly a two-browser solution (one for selecting rhythmic patterns from sound events, the second for timbres from audio grains) that would have been harder and slower to implement and less straightforward in terms of interaction than the solution we opted for, a single browser revealing temporal and timbral features.

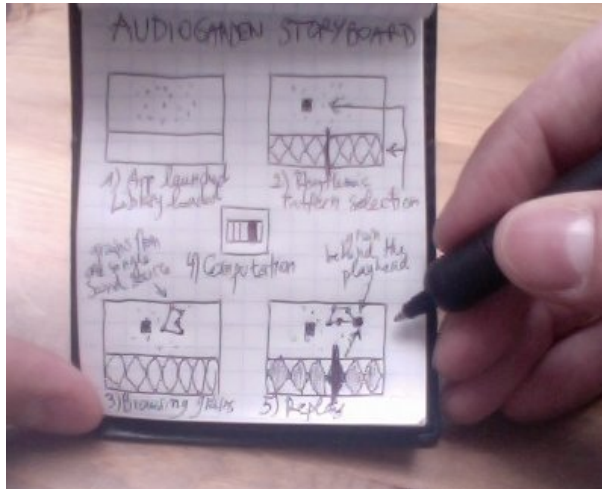


Figure 3: Storyboard of an expected scenario of usage of the desired workflow quickly drawn on a small notebook.

3.2. Proposed scenario of interaction

The scenario consists in:

1. browsing, listening to and selecting:
 - (a) one sound event for its rhythmic pattern,
 - (b) several audio grains for their timbral character,
2. easily constructing a new sound event that “updates” the chosen sound event with different timbral features;
3. listening to the new sound event, optionally saving it (thus making it appear on the browser);
4. renewing the aforementioned cycle (steps 1. and 2.), by either choosing another sound event or different grains, or starting again with no audio content set.

3.3. Proposed Visualisations

3.3.1. Disc

The first proposed visualisation is shown on Fig. 4. The position of the points is computed as follows:

- The radius is proportional to the inverse of the logarithm of the duration of the sounds. Thus long sounds, that can be used as targets are positioned in the center of the display and short sounds that can be used as grains are on the periphery of the circle.
- The angle depends on the timbre features (for now only MFCC). It is proportional to the coordinates of the sound on the first principal component of the MFCC.

This visualisation is useful to explore the grain database and to experiment with the synthesis method.

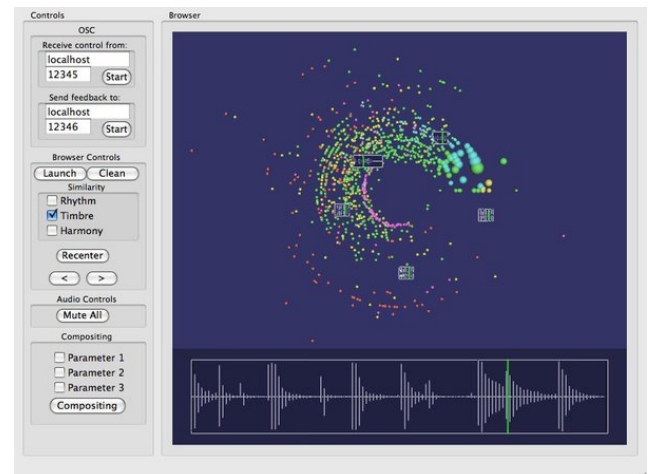


Figure 4: Screenshot of an early prototype of the user interface, featuring a two-pane view: audio database browser by similarity on top, waveform of the sound being “composited” on the bottom.

3.3.2. Flower

The second visualisation is shown on Fig. 5. Each long sound (a sound that has been segmented) is represented by a circle. The long sound itself is in the center of the circle, while the segments of this sound form a circle around it. The grains in the circle are in chronological order. The circles are placed in the 2D space depending on the average timbre, that is, the coordinates of the center of the circles are proportional to the two first principal components of the MFCC and SFM features.

In this visualisation it is possible to select the grains one by one, or if one clicks on the center of a circle while hitting a special key, the entire circle is selected at once. This visualisation is very useful to mix sounds, i.e. using one sound as a target and all the segments of one or several other sounds as grains.

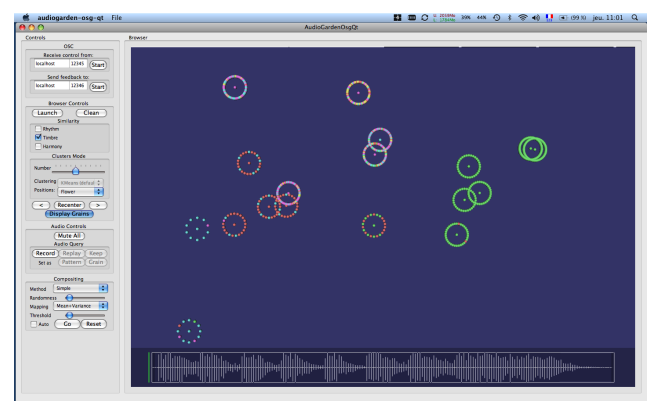


Figure 5: Screenshot of an early prototype of the user interface, featuring a two-pane view: audio database browser by similarity on top, waveform of the sound being “composited” on the bottom.

4. SUMMARY OF MEDIACYCLE IMPROVEMENTS

- Waveform display at the bottom of the window to show one sound of particular interest, such as the target or the synthesis in audiogarden,
- Possibility to record a sound from the computer input, play it back and add it to the library,
- Internal support for media segments,
- Support for segmentation plugins.

5. CONCLUSION AND PERSPECTIVES

In this project, we designed and developed a tool for sound creation. First of all, it has been the occasion to test the flexibility of the Mediacycle framework. In the course of the development, Mediacycle has proved to be a very good tool to quickly design new audio application and test various kind of sound analysis and data display on screen. By allowing a very fast prototyping, it allows to test various configurations and then select the best one in a very short period of time. In addition with paper mockups, such frameworks can be very useful tools for research and development. Some improvements have also been done, such has the support for segments that will be useful for many other applications. But the main achievement of the project is the AudioGarden software. Standing on content driven concatenative synthesis, this software proposes a new way to create sounds. Informal tests showed that a large variety of sounds can be created, but still more formal tests needs to be performed. Some aspect needs improvements. The synthesis method could be improved in many ways. First the segmentation and grain addition could be changed such has described in section 2.2 to allow both a segmentation on low energy part and a synthesis that preserve the rhythm of the target. New mappings between the target and the grains could also be explored. We proposed three simple ones, that give very good results for some kind of sounds, but sometimes unexpected results on other sounds. This may involve sound similarity perception researches and experiment. Finally different visualisation could be proposed for different usages, this would need to give the tool to different kind of users and look at the way they use it and listen to their suggestions.

6. ACKNOWLEDGMENTS

numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631).

C. Picard obtained a Short-Term Scientific Mission (STSM) funding from the COST Action Sonic Interaction Design (SID)¹.

We want to thank the One Laptop Per Child Project (OLPC) for providing the Free Sound Samples Library under a Creative Commons license².

7. REFERENCES

- [1] P. R. Cook. "Toward Physically-Informed Parametric Synthesis of Sound Effects." In: *In Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-99)*. 1999. Pp. 1–5. P.: 33.

¹COST SID: <http://www.cost-sid.org>

²OLPC Sound Sample Library: http://wiki.laptop.org/go/Sound_samples

- [2] Christoph Cox and Daniel Warner, eds. *Audio Culture: Readings in Modern Music*. Continuum International Publishing Group, 2004. ISBN: 9780826416148. P.: 33.
- [3] D. Keller and B. Truax. "Ecologically-based Granular Synthesis". In: *Proceedings of the International Computer Music Conference (ICMC)*. Ann Arbor, USA 1998. P.: 33.
- [4] Ari Lazier and Perry Cook. "MOSIEVIUS: Feature driven interactive audio mosaicing". In: *Proceedings of the International Conference on Digital Audio Effects*. London, UK 2003. P.: 33.
- [5] Daniel Lopez, Francesc Marti, and Eduard Resina. "Vocem: An Application for Real-Time Granular Synthesis". In: *Proceedings of the Digital Audio Effects (DAFx)*. 1998. P.: 33.
- [6] A. Misra, P. R. Cook, and G. Wang. "Musical Tapestries: Re-composing Natural Sounds". In: *Proceedings of International Computer Music Conference (ICMC '06)*. New Orleans, USA: International Computer Music Association, 2006. P.: 33.
- [7] Curtis Roads. "Introduction to Granular Synthesis". In: *Computer Music J.* 12.2 (1988). P.: 33.
- [8] Diemo Schwarz. "Concatenative Sound Synthesis: The Early Years". Ed. by Adam T. Lindsay. In: *Journal of New Music Research* 35.1 (Mar. 2006). Pp. 3–22. P.: 33.
- [9] Carolyn Snyder. *Paper Prototyping: The Fast and Easy Way to Define and Refine User Interfaces*. Morgan Kaufmann, 2003. ISBN: 1558608702. P.: 34.