VIVO: Video Analysis for Corpus-based Audio–Visual Synthesis

Matéo Fayet, Diemo Schwarz, Vincent Tiffon

VIVO: VIDEO ANALYSIS FOR CORPUS-BASED AUDIO–VISUAL SYNTHESIS

Matéo Fayet
STMS, Ircam–SU–CNRS,
PRISM, AMU–CNRS
fayet@ircam.fr

Diemo Schwarz
STMS Ircam–SU–CNRS
Paris, France
schwarz@ircam.fr

Vincent Tiffon
Aix Marseille Univ, CNRS
PRISM, Marseille, France
tiffon@prism.cnrs.fr

Résumé

La synthèse concaténative par corpus audio–visuelle étend le principe de synthèse concaténative sonore au domaine visuel, où en addition du corpus sonore (i.e une collection de segments de son enregistrés accompagnés d'une description perceptive de leurs caractéristiques), l'artiste utilise un corpus d'images statiques avec leurs caractéristiques visuelles perceptives (couleur, texture, détail, luminosité, entropie, mouvement), dans le but de créer une performance audio–visuelle musicale en navigant en temps réel dans ces espaces de descripteurs, i.e. à travers une collection de grains sonores dans un espace de descripteurs audio perceptifs, et à travers un espace de descripteurs visuels, i.e. en sélectionnant des images dans un corpus visuel pour le rendu, et en conséquence naviguer en parallèle au travers des deux corpus de manière interactive par contrôle gestuel tactile. Nous étendons ici ce principe à l'analyse de vidéos pour constituer le corpus visuel, avec l'ajout de quelques descripteurs spécifiques. La question arts–sciences qui est ici explorée dans le cadre d'une création artistique est quels descripteurs visuels sont adaptés à une interaction multi-modale et comment les intégrer depose l'analyse de données vidéo en temps-réel dans un système de synthèse sonore concaténative par corpus dans le but de créer une expérience audio–visuelle multi-modale incarnée.

Abstract

Audio–visual corpus-based synthesis extends the principle of concatenative sound synthesis to the visual domain, where, in addition to the sound corpus (i.e. a collection of segments of recorded sound with a perceptual description of their sound character), the artist uses a corpus of images with visual perceptual description (colour, texture, detail, brightness, entropy, movement), in order to create an audio–visual musical performance by navigating in real-time through these descriptor spaces, i.e. through the collection of sound grains in a space of perceptual audio descriptors, and at the same time through the visual descriptor space, i.e. selecting image frames from the visual corpus for rendering, and thus navigate in parallel through both corpora interactively with gestural control via touch sensing. We extend here this principle to the analysis of videos constituting the visual corpus, by adding video-specific descriptors. The artistic-scientific question explored here based on the realisation of a concrete performance piece is which visual descriptors are suitable for multi-modal interaction and how to integrate them via real-time video data analysis into a corpus-based concatenative synthesis sound system with the aim of creating an embodied multi-modal audio–visual experience.

1. INTRODUCTION

This article presents the integration of video corpora to corpus-based audio–visual synthesis. This method offers new ways of creating gesture-controlled audio–visual live performances, based on cross-modal perception and mobilising inter-modal analogies. With the previously existing CoCAVS system [27], one can create audio–visual musical systems by navigating through a descriptor space selecting sound grains classified depending on their sound description, and through another descriptor space in which the elements are images classifed depending on their visual descriptions. Interaction with both spaces are induced by a touch-pad. An overview of this system is depicted in figure 1. It allows artists to analyse and play their own corpora and perform cross-modal audio–visual interaction depending on their wish or need. The new ViVo [1] project described here aims to extend this system to video streams by 1) allowing users to load video-frame based data 2) develop visual descriptors adapted to video 3) since video takes place over time, allowing users to analyse a video stream in real-time. The system was developed in parallel to an artistic residency based on the use of these tools during which each experiment was a concrete realization of the technological development [6]. This article will describe the artistic and technological context of an audio–visual piece and corpus-based concatenative synthesis in which the project is implemented. We will briefly describe the algorithms, implementation, use cases and the artistic application of ViVo.

2. MOTIVATION AND RELATED WORK

2.1. Concatenative Synthesis

Corpus-based concatenative synthesis (CBCS) was explored in a music creation context since 1999 [25], and

[1] `http://github.com/ircam-ismm/vivo`

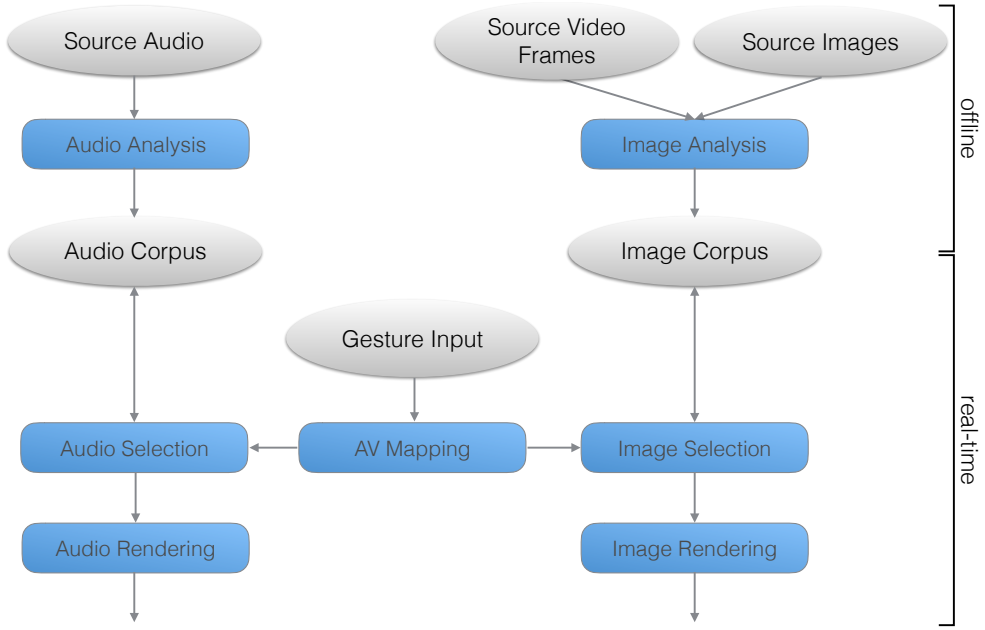

Figure 1. Schema of the ViVo Image Browser.

then applied in real time with CataRT [2] [28, 29] since 2005. This type of digital audio synthesis plays grains from a corpus of sounds segmented and analysed according to their proximity to a target position in a descriptor space [28]. It's usage has now been extended to musical performance, sound design and installation contexts, while also enabling interactive exploration of the corpus by users [26].

Furthermore, Nick Collins' work explores the possibility of simultaneously using audio and visual descriptors for data-driven synthesis in a technical and artistic context. In this project, five audio descriptors and five video descriptors are used and described [2].

More recently, Schwarz proposes with the CoCAVS project to link a corpus of images and a corpus of sounds with the help of two-dimensional mapping, a selection algorithm that choose the closest elements imposed by one of the two corpora, and audio and graphics engines for rendering sound and visual elements. Here, the term mapping refers to an action that acts on two modes at once. In other words, the selection of samples of the two corpora is controlled by a single gesture. This research was carried out during an art–science residency at the IMéRA Institute for Advanced Studies in July 2022, and an electroacoustic composition was presented in the Andromède planetarium. [3] This research was accompanied by the development of seven visual descriptors that characterize the perceptual qualities of each image through numerical values [27].

One of the essential elements of this research concerns audiovisual mapping and the assignment of descriptors in both dimensions. The author points out that these links are more a matter of mobilizing cross-modal analogies rather than being based on human multi-modal perception. For example, the brightness of a sound represented by its spectral centroid can be correlated to the brightness of an image based on luminance analysis. Nevertheless, audio and video descriptors can be correlated in any configuration to allow artistic freedom with no regard to inter-modality intentions [27].

A second element concerns the link of selection between the corpora in two different modalities. To address this, two triggering modes can be used. One is to take the user-controlled target position (usually in 2D) in each corpus independently and trigger the respective nearest elements (*pre-selection pairing*), the other consists in triggering the element of a corpus B after having selected one in the other corpus A via a user-defined cross-modal mapping from A to B (*post-selection pairing*). This mapping can make use of the full dimensionalities of both corpora [27].

This tool shows that the development of multi-modal performances and creations is facilitated by the design of perceptually valid and comprehensible descriptors by humans in the different dimensions. Schwarz also suggests that future work needs to be carried out to develop image descriptors that take into account an estimation of texture and object salience, as well as meta-data [27].

2.2. Multi-modality

Cross-modal perception has been theorized with several different approaches from synaesthesia [15] to the ecological theory of perception [8, 10, 9, 11]. More recently, some research in neurosciences and cognitive psychology show the complexity of cross-modal events in human per-

[2] `http://ircam-ismm.github.io/max-msp/catart.html`

[3] *Performance for Audio–Visual Concatenative Synthesis & Violin* `http://youtu.be/EFAN9fOofd0`

ception [12, 1, 3, 23, 13], and the age-old search for sensory translation in the arts [32, 31]. Nevertheless, ViVo aims to offer a high degree of freedom regarding cross-modal associations. By providing different analysis modules for both audio and visual modalities, people will be able to create their own analogies allowing to test the limits of sensory expectations.

The aim of this project is to extend corpus-based audio–visual synthesis to video content. This involves the development of visual analysis algorithm (in addition to those developed during the CoCAVS project) and an adjustment of the already existing tools in the MuBu/CataRT framework.

2.3. Artistic Context

Even though a plethora of multi-modal artistic work has been done in the experimental cinema creative field, we decided to focus on two main musical trends which are visual music and *vidéomusique*. Not only do these historical references enable us to understand the aesthetics of these artistic movements, they also allow us to effectively plan and forecast the development of a tool adapted to creative needs.

2.3.1. Visual Music

In 2005, Brian Evans suggested a preparatory work on the foundations of visual music theory through a geometric and colourimetric analysis of images [5]. Furthermore, the author extracts from certain visual figures creative behaviours similar to musical compositions. Concepts such as tension and resolution and their respective degrees are discussed. The author also mentions subjectivity regarding visual perception. He specifies for instance that "the perception of colour is inexact, culturally influenced, and personal" [5]. However, certain digital processes (in the HSV spectrum, for example) can be used to analyse colourimetry through generalisable aspects such as brightness and saturation. The author cites Norman McLaren (with compositions like – *horizontal and vertical – lines* [18] [17] and *synchromy* [16]), among others, as one of the key players in visual music, whose works help us to understand the stakes involved in the style of music and film.

2.3.2. Vidéomusique

More recently, Jean-Pierre Moreau's work studies the history and reception of *vidéomusique* (a different aesthetic trend from visual music) and proposes a method for analyzing videomusical works [19] [20]. Among other things, his work has enabled us to build up a musicological lexicon around this artistic practice. Moreau references Michel Chion, a member of the GRM, and Jean Piché, whom we should mention for his pioneering videomusical works. In compositions such as *Sièves* [21] we notice that there is no permanent congruence between the different media. In this regard, one can observe an interesting opposition to McLaren's works which demonstrate near-permanent synchronicity. These different approaches offer us logical prospects for the development of a multi-modal creative tool.

3. VIVO OVERVIEW

The analysis of video parameters for music and sound purposes requires the development and adaptation of tools to extract usable numeric values (whether in real-time or offline). In addition to the seven already existing image descriptors [27], which we will briefly revisit in 3.1, we have chosen to develop four new image and video analysis modules. The analysis modules are based on Jitter within the Max environment [4] for compatibility reasons with the existing systems and the ease of prototyping offered by the software. Indeed, the frameworks used for the audio synthesis (MuBu [5] [24] and CataRT [2]) and video rendering are Max-based, and the integration of new analysis modules is facilitated when they are created in the Max environment. However, every algorithm (because they are mainly based on matrix processing) could easily be implemented in other environments (such as *Unreal Engine* and *TouchDesigner* with Python modules).

3.1. Image Descriptors

ViVo's image descriptors characterise perceptive qualities of single image frames with numerical values. They are listed in table 1 and are explained in the following.

3.1.1. Overall Image Qualities

These three descriptors characterise the global colour qualities of an image with numerical values. They are calculated as the mean and standard deviation of each channel of the HSL colour representation over all pixels, expressing the average colour (hue), saturation, and luminosity of an image, and the degree of their disparity (variance).

3.1.2. Dominant Colour

As the average hue does not necessarily correspond to any colour existing in the image, the *dominant colour* descriptor C is calculated as the mode of the hue histogram weighted by luminance, i.e. the hue which occurs in the most pixels in the image, where bright pixels count more than dark pixels [27]:

$$C = \arg\max_i c_i \tag{1}$$

where c is the histogram of the pixel hue values H_n weighted by the luminance L_n from the 8 bit HSL colour space:

$$c_i = \sum_{n|H_n=i} L_n \tag{2}$$

[4] http://cycling74.com/max

[5] http://forum.ircam.fr/projects/detail/mubu

3.1.3. Keypoint-based Descriptors

The last three image descriptors are derived from a *keypoint* or *salient feature* computer vision analysis of the image by *cv.jit*, detecting "interesting" image details by edges and their intersections. The number of these keypoints found expresses the overall complexity of the image, and their average position and variance the spatial centre and extent of the region with the most salient details in the image [27]. An example of the keypoint analysis can be seen in figure 2.

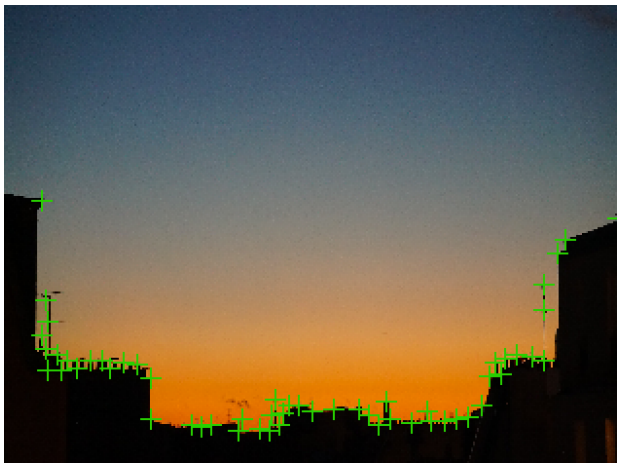

Figure 2. Example of keypoint analysis as performed by the OpenCV computer vision package. Each green cross is one salient feature or keypoint found.

In addition to the seven image descriptors described above, which were developed during the CoCAVS project [27], ViVo contributes four new analysis abstractions which characterise higher-level perceptive attributes of images and videos.

3.1.4. Warmness

The warmness analysis module is based on an algorithmic approach developed by Dimopoulos and Winkler [4]. As described in their article, "effects of cold and warm colours, as an aspect of human colour perception, have been mostly studied by psychologists. They split visible colours into two groups, one of warm and one of cold colours respectively based on their impact on people." In order to algorithmically determine the warmth of an image, one must compute colour warmness of each of the N quantised colours [6] of an HSV (Hue, Saturation, Value) matrix. Then "image warmness is [...] defined as the weighted average colour warmness values of all colour bins". For this we, first, assign a value of $+1$ to warm and -1 to cold colours, based on the hue value H_n of colour bin n in degrees

$$T_n(H_n) = \begin{cases} -1 & \text{if} \quad 75 \quad < H_n < \quad 285 \\ +1 & \text{if} \quad 0 \quad \leq H_n \leq \quad 75 \\ & \text{or} \quad 285 \quad \leq H_n \leq \quad 360 \end{cases} \tag{3}$$

and then calculate the per-pixel warmness θ_n as

$$\theta_n = T_n(H_n) w_n(S_n, Vn), \tag{4}$$

where the weight w_n is based on the saturation S_n and value (brightness) V_n from the HSV colour space:

$$w_n(S_n, Vn) = S_n\, V_n, \quad \text{with} \quad S_n, V_n \in [0, 1] \tag{5}$$

and finally we calculate the global image warmness Θ as

$$\Theta = \sum_{n=1}^{N} f_n \theta_n, \tag{6}$$

with $f_n = c_n / I$ the relative frequency of colour bin n based on c_n, the number of pixels with colour n and I the total number of pixels in the image. These processing steps have been implemented in Jitter and `jit.gen` under a Max abstraction.

3.1.5. Sharpness

The principle of this module is to determine the level of sharpness (as opposed to blur) of an image. When an image is sharp, it shows the edges of the objects that make it up. The blurrier the image, the less perceptible the edges. An edge is "a collection of the pixels whose gray value has a step or roof change, and it also refers to the part where the brightness of the image local area changes significantly" [7]. There are various edge detection algorithms, such as Sobel, Prewitt and Roberts. All three of these algorithms are available on Jitter, giving you a wide choice to suit your needs. For this abstraction, we chose to use the Sobel detection algorithm which offers thicker and brighter representation of edges than its two counterparts [7]. The `jit.sobel` object marks pixels pertaining to edges independently in each plane of an RGB matrix. Therefore, we calculate the mean over all pixels in each plane separately and extract the maximum mean value among the three RGB planes.

3.1.6. Detail

An interesting analysis suggested by Nick Collins [2] during his research on audiovisual concatenative synthesis concerns the analysis of image texture using a Fourier transform algorithm. This provides a spatial frequency-based 2-dimensional spectral representation of an image. The aim of this module is to determine the overall level of detail of an image, and the level of detail for a given frequency range. To do this, we need to create sub-bands of the image spectrum that correspond to frequency ranges. Subsequently, users will be able to modify the size and position of each band, and thus analyse a desired frequency

[6] While the original article [4] uses colour quantisation adapting to the actually occurring colours of an image, we chose to pre-quantise the colour space to 32 steps of hue, and 8 steps of saturation and value, resulting in $N = 2048$ colour bins.

Image Descriptor Name	Explanation
HueAvg, HueVar	Mean and standard deviation of pixel hue (colour value)
SaturationAvg, SaturationVar	Mean and standard deviation of pixel saturation
LuminanceAvg, LuminanceVar	Mean and standard deviation of pixel brightness
Color	Dominant colour
Complexity	Number of keypoints
XAvg, XVar	Mean and standard deviation of keypoint x positions
YAvg, YVar	Mean and standard deviation of keypoint y positions
Warmness	Mean warmness of overall image
Sharpness	Mean sharpness of overall image
Detail (low, mid, high, total)	Mean detail per frequency band or total
Movement (video)	Overall inter-frame movement and detailed (2 axis) movement

Table 1. List of CoCAVS and ViVo image and video descriptors.

range as well as the mean value of every band. This abstraction is developed using the `jit.fft` object. The detail and sharpness descriptors can give quite complementary information about the image content, since even high-detail images can still be more or less sharp, depending on focus and image quality. This is illustrated in the examples in figure 3.

3.2. Video Descriptors

The video-specific descriptors are calculated based on a stream of video frames.

3.2.1. Movement

The optical flow algorithm estimates the overall level of movement between one video frame and the next, as well as the direction (horizontal/vertical) and intensity of movement. According to Horn and Schunk [14], "optical flow is the distribution of apparent velocities of movement of brightness patterns in an image." Starting from the Horn-Schunk algorithm implemented in the *cv.jit* Max Package, [7] we determine the absolute values i.e. the square root of each direction matrix—horizontal and vertical—and get the mean value of the two matrices. Therefore, we can extract both vertical and horizontal average movement, as well as an overall average value of the image optical flow.

3.3. Implementation

The ViVo system is implemented in Max, with the MuBu and CataRT extension libraries, and making use of Max's Jitter operators for handling, processing, and rendering image data (on the CPU and GPU).

It was straightforward to adapt CataRT and its underlying MuBu container for time-based media and data to handle image or video frame references and descriptors, instead of audio segments, keeping the same structuring (one MuBu buffer for each external file (image or video), one data frame for each unit). This can later be extended to handle multiple sub-images in one image file. The only difference is that the image data itself is stored as a Jitter movie, and MuBu just stores the video and frame indices as a reference. CataRT's architecture for importing, analysing, selecting, and rendering audio units was easily adapted to handle image data. Especially the scalable kNN unit selection [30] is completely agnostic of the media type, since it finds nearest neighbours in a Euclidean descriptor space and outputs the ID of the closest unit(s), which is a segment ID for audio, and an image or video frame index for visual corpora.

Because visual descriptors abstractions analyse each frame of a video, ViVo allows us to use it in *real-time* or *offline*. Thus, a first simple way of using the ViVo package would be to analyse a video stream and extract the image descriptors in *real-time*. This example of use will be described in the next section. Another way of using the ViVo package is by classifying each frame in a two-dimensional explorer and being able to display each frame as a point. This will allow us to play sound grains and image frames simultaneously as in the CoCAVS project. In fact, the same rendering engine was used. The changes remain in the pre-processing procedures, where we cut videos into their individual image frames. This is what is known as the ViVo Video Browser [8] (figure 4).

4. ARTISTIC EXPERIMENTS

ViVo's development is driven by the desire to create immersive audiovisual works. In this chapter we describe the genesis of $\epsilon\upsilon\tau\epsilon\rho\pi\eta$, [9] [10] an immersive piece around which the development of ViVo has been built until now. [11]

[7] While Horn and Schunk describe the optical flow algorithm based on vectorial matrices, Jean-Marc Pelletier provides in *cv.jit* an algorithm based on pixel displacement estimation.

[8] Tutorial and demo available at `http://youtu.be/R16AeS8phFI`

[9] Euterpe in Greek.

[10] Video available at `http://youtu.be/HAFOElr8lEs`

[11] This work was produced during an artistic residency at La Fabulerie in Marseille.

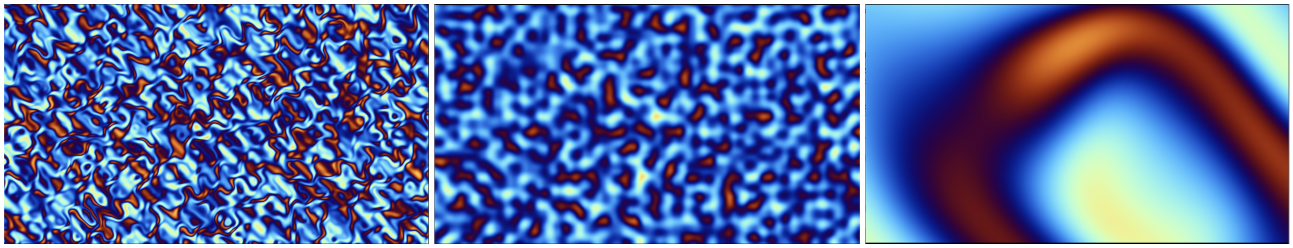

Figure 3. Artificially generated textures to illustrate the image descriptors *Sharpness* vs. *Detail*. From left right: high sharpness/high detail, low sharpness/high detail, low sharpness/low detail.

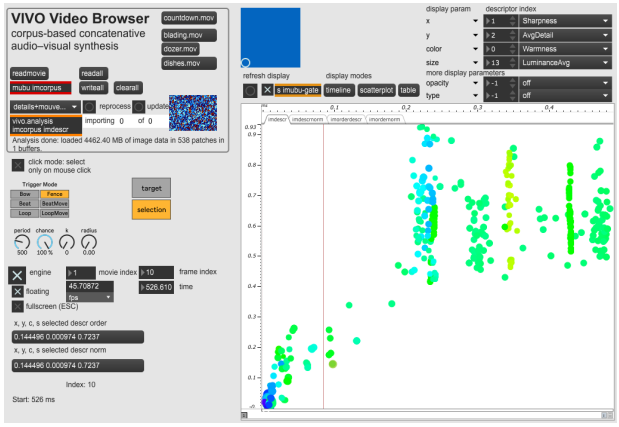

Figure 4. The ViVo Video Browser.

4.1. The Piece $\epsilon\upsilon\tau\epsilon\rho\pi\eta$

4.1.1. Genesis and Aesthetics

ViVo's development is driven by the desire to create a collaborative immersive videomusical work. To achieve these two goals, instrument development and creation have evolved in parallel. The main theme studied here is water. Indeed, growing this piece in Marseille influenced our aesthetics desires not only by the natural environments it is surrounded by but also by being an incubator in experimental arts. In this regard, *Sud* by Jean-Claude Risset [22] remains the main inspiration about sound aesthetics and the global frame of the piece. This twenty minutes creation is performed by the VJ Solal Fayet and musician Matéo Fayet. Both the visual and sound corpora were made of natural material, most of the time recorded and filmed at the same place [6].

4.1.2. The Use of ViVo

One possible use for ViVo is to establish a correspondence (congruent or not) between real-time analysed video parameters and concatenative audio synthesis parameters, as shown in figure 5.

The development of these modules enable automatic control of sound synthesis parameters depending on live displayed video characteristics. Therefore in this piece, the VJ controls the triggering and effects of his own corpus using a MIDI controller and Jitter modules, displaying a 180º video mapping using three projectors. The video stream is locally analysed by ViVo, and the resulting data is sent over UDP using the OSC protocol to the computer in charge of sound rendering. This computer contains a CataRT synthesis engine in which field recorded sounds are loaded, analysed and displayed as grains. This musician is then able to control grain triggering using a Sensel Morph touch interface, [12] and audio parameters and quadrophonic spatialisation using a MIDI controller. The incoming data from ViVo is then directly mapped to some of the main sound synthesis parameters as follows:

- The warmness factor is directly related to the attack and release times of each grain: the warmer the image, the longer the attack and release times.
- The detail factor is mapped to the resampling randomization interval, the `mubu.concat` attribute used to pitch grains: the more complex the image, the larger the randomization interval.
- Blur sharpness analysis is directly linked to grain triggering frequency: the blurrier the image, the lower the trigger frequency. [13]

It is important to note that the proposition of multimodal mapping described above, either congruent or not, were driven by artistic intentions and do not relate to any scientific generality about human multi-modal perception. Moreover, some mapping (e.g. image warmness to sound grain envelope), proved to be quite imperceptible during the major part of the show. Indeed, altering attack and release time when overlapping a lot of long grains can result in unhearable modifications of sound rendering in general. Nonetheless, having put this direct mapping while working and experimenting led and helped define some artistic intents.

This artistic experiment has shown how effective and useful ViVo is in an artistic and real-time context. However, this did not lead to the use of the ViVo Video Browser introduced in section 3.3 which we believe is one of the most interesting and exciting feature of ViVo.

[12] http://morph.sensel.com/

[13] Note here the absence of any use of the movement module, which proved less responsive and effective given the abstract nature of the video feed.

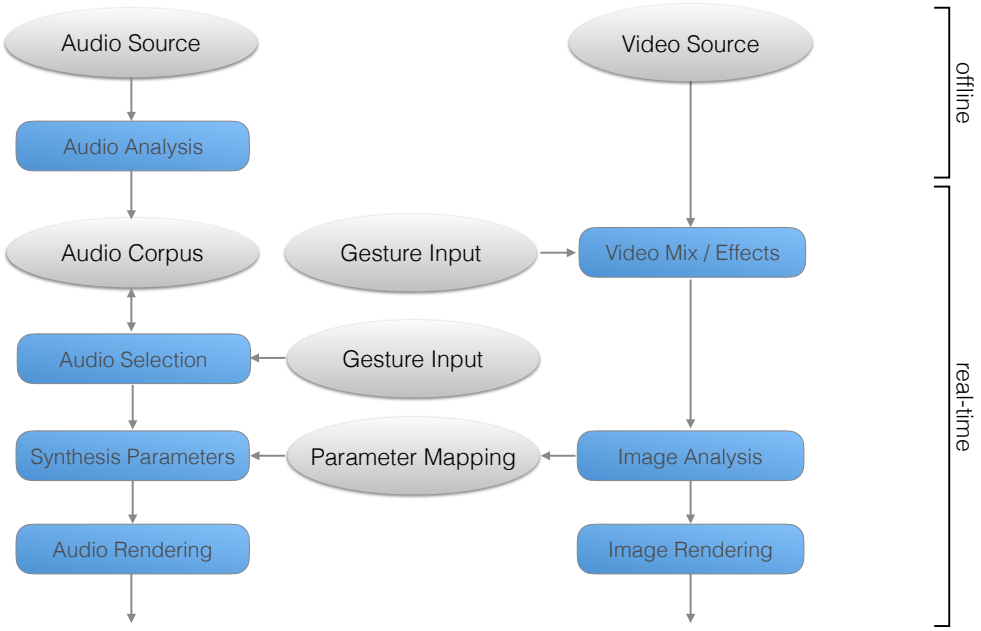

Figure 5. Audiovisual interaction schema, where the ViVo image and video descriptors are analysed in real-time to influence corpus-based synthesis parameters.

5. DISCUSSION AND FUTURE WORK

For future work, there is growing interest in computational content analysis of audiovisual material for interactive repurposing and collage. First, the instantaneous AV descriptors can be used for multi-modal segmentation, i.e. audio descriptors can detect events or transients, or video descriptors can detect scene changes or cuts, which both give rise to meaningful video segments. Further, if the ViVo image and video descriptors are paired with the audio descriptors of video segments, then either audio- or image-descriptor-driven lookup and collage out of a video corpus becomes possible, extending the idea of the video sampler to content-driven synthesis. Finally, one can imagine processing video grains the same as grains are processed in sound granular synthesis. In other words, video grains could be not only a frame but a timed segment from which we could control parameters like length, envelope and speed. These evolutions and experiments should give rise to artistic creations that will help to overcome the limitations and problems associated with the tools developed here.

6. CONCLUSION

ViVo introduces new ways to explore audio–visual corpora within cross-modal interactions. Even though its philosophy and content are based on MuBu/CataRT timbre-based sound exploration (for either congruent or contradictory inter-modal associations), its versatility offers a range of uses to be explored. We strongly believe in its usefulness in the artistic and scientific fields as a basis for sensitive multi-modal exploration. While some recent trends tend to use machine-learning, generative and from end-to-end synthesised materials, this project aims to give users a control over human perception through sensible descriptors. This high-level interpretation given by low-level control offers an alternative and fully customisable way of expressing narration.

7. ACKNOWLEDGMENTS

The basis of this work has been funded by the Arts, Sciences, Societies residency program 2021–2022 of the IMéRA Institute for Advanced Studies, Aix–Marseille Université, and supported by Laboratoire Perception, Représentations, Image, Son, Musique (PRISM), CNRS, Aix-Marseille Université.

8. REFERENCES

[1] Adeli, M., Rouat, J., and Molotchnikoff, S., "Audiovisual correspondence between musical timbre and visual shapes", *Frontiers in human neuroscience*, 8:352, 2014.

[2] Collins, N., "Audiovisual Concatenative Synthesis", in *International Computer Music Conference*, 2007.

[3] Deroy, O. and Spence, C., "Crossmodal correspondences: Four challenges", *Multisensory research*, 29(1-3):29–48, 2016.

[4] Dimopoulos, M. and Winkler, T., "Image warmness: A new perceptual feature for images and videos", in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 1662–1666, IEEE, 2014.

[5] Evans, B., "Foundations of a visual music", *Computer Music Journal*, 29(4):11–24, 2005, ISSN 01489267, 15315169.

[6] Fayet, M., "Vivo : une approche multimodale de la synthese concatenative par corpus dans le cadre d'une oeuvre audiovisuelle immersive", 2024, doi: 10.48550/ARXIV.2404.10578.

[7] Gao, W., Zhang, X., Yang, L., and Liu, H., "An improved sobel edge detection", in *2010 3rd International Conference on Computer Science and Information Technology*, IEEE, July 2010, doi:10.1109/iccsit.2010.5563693.

[8] Gaver, W., "How do we hear in the world? Explorations in ecological acoustics", *Ecological psychology*, 1993.

[9] Gibson, J. J., *The senses considered as perceptual systems.*, Houghton Mifflin, 1966.

[10] Gibson, J. J., "On the analysis of change in the optic array.", *Scandinavian Journal of Psychology*, 1977.

[11] Gibson, J. J., *The Ecological Approach to Visual Perception: Classic Edition*, Houghton Mifflin, 1979.

[12] Grill, T. and Flexer, A., "Visualization of perceptual qualities in textural sounds", in *Proceedings of the International Computer Music Conference (ICMC)*, 2012.

[13] Guellaï, B., Callin, A., Bevilacqua, F., Schwarz, D., Pitti, A., Boucenna, S., and Gratier, M., "Sensus communis: Some perspectives on the origins of non-synchronous cross-sensory associations", *Frontiers in Psychology*, 10:523, 2019, ISSN 1664-1078, doi: 10.3389/fpsyg.2019.00523.

[14] Horn, B. K. and Schunck, B. G., "Determining optical flow", *Artificial intelligence*, 17(1-3):185–203, 1981.

[15] Jones, R. and Neville, B., "Creating visual music in jitter: Approaches and techniques", *Computer Music Journal*, 29:55–70, 2005.

[16] McLaren, N., "*Synchromy*", 1971. Canada, Québec.

[17] McLaren, N. and Evelyn, L., "*Lines - Vertical*", 1960. Canada, Québec.

[18] McLaren, N. and Evelyn, L., "*Lines - Horizontal*", 1962. Canada, Québec.

[19] Moreau, J.-P., *De la perception à la représentation: vers une épistémologie de l'œuvre interdiscursive Pour une analyse de l'œuvre vidéomusicale*, Ph.D. thesis, Langues, Lettres et Arts (ED 354), 2018.

[20] Moreau, J.-P., *De la perception à la représentation - Analyser l'œuvre vidéomusicale*, CREArTe, EME éditions, Louvain-la-Neuve, 2021.

[21] Piché, J., "*Sieves*", 2004. Canada, Toronto, SoundPlay Festival.

[22] Risset, J.-C., "*Sud*", 1985. Cd Ina/GRM C 1003.

[23] Saitis, C., Weinzierl, S., von Kriegstein, K., Ystad, S., and Cuskley, C., "Timbre semantics through the lens of crossmodal correspondences: a new way of asking old questions", in *International Symposium on Universal Acoustical Communication*, Sendai, Japan, 2018.

[24] Schnell, N., Röbel, A., Schwarz, D., Peeters, G., and Borghesi, R., "MuBu & friends – assembling tools for content based real-time interactive audio processing in Max/MSP", in *Proceedings of the International Computer Music Conference (ICMC)*, Montreal, Canada, August 2009.

[25] Schwarz, D., *Data-Driven Concatenative Sound Synthesis*, Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris, 2004.

[26] Schwarz, D., "Concatenative sound synthesis: The early years", *Journal of New Music Research*, 35(1):3–22, March 2006. Special Issue on Audio Mosaicing.

[27] Schwarz, D., "Touch interaction for corpus-based audio–visual synthesis", in *Proceedings of the Conference for New Interfaces for Musical Expression (NIME)*, Mexico City, Mexico, May 2023.

[28] Schwarz, D., Beller, G., Verbrugghe, B., and Britton, S., "Real-Time Corpus-Based Concatenative Synthesis with CataRT", in *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 279–282, Montreal, Canada, September 2006.

[29] Schwarz, D., Cahen, R., and Britton, S., "Principles and applications of interactive corpus-based concatenative synthesis", in *Journées en Informatique Musicale*, GMEA, Albi, France, March 2008.

[30] Schwarz, D., Schnell, N., and Gulluni, S., "Scalability in content-based navigation of sound databases", in *Proceedings of the International Computer Music Conference (ICMC)*, Montreal, QC, Canada, August 2009.

[31] Spence, C. and Di Stefano, N., "Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and sound", *i-Perception*, 13:204166952210928, 05 2022, doi:10.1177/20416695221092802.

[32] Spence, C. and Di Stefano, N., "Sensory translation between audition and vision", *Psychonomic Bulletin & Review*, 10 2023, doi:10.3758/s13423-023-02343-w.