"MUSICA SULL'ACQUA": A MOTION TRACKING BASED SONIFICATION OF AN AQUARIUM IN REAL TIME

Stefano Baldan, Luca A. Ludovico, and Davide A. Mauro
Laboratorio di Informatica Musicale (LIM)
Dipartimento di Informatica (DI)
Università degli Studi di Milano, Via Comelico 39/41, I-20135 Milan, Italy
http://www.lim.dico.unimi.it
(baldan, ludovico, mauro)@dico.unimi.it

ABSTRACT

This paper presents a temporary multimedia installation set up at the *Civic Aquarium of Milan*. Thanks to four web cameras located in front of the tropical fishpond, fish are tracked and their movements are used to control a number of music-related parameters in real time. In order to process multiple video streams, the open-source programming language Processing has been employed. Then, the sonification is implemented by a Pure Data patch. The communication among the parts of the system has been realized through Open Sound Control (OSC) messages. This paper describes the key concepts, the musical idea, the design phase and the implementation of this installation, discussing also the major critical aspects.

1. INTRODUCTION

This paper aims at describing a temporary multimedia installation based on a set of four web cameras, a personal computer running *ad hoc* software for video and sound processing, and four loudspeakers for sonification purposes.

The installation has been realized in the framework of the initiative called "L'avventura della Scienza" (that can be translated to "Adventures in Science"), organized by the *Università degli Studi di Milano* and held in Milan from March 12 to April 9 in 2012. The main idea was making science tangible and enjoyable in its many facets, both for children and for adults. A number of initiatives has been organized by researchers, ranging from shows to workshops, from exhibitions to meetings. An interesting aspect is given by the choice of non-traditional locations for academic didactic activities, such as cinemas, theaters, etc. For further details, please refer to the official Web site. [1]

In this framework, the Computer Science research area planned a number of labs and other didactic activities at the *Civic Aquarium of Milan*. Such a location suggested us to rethink water - and in particular fish tanks - as an acoustic environment. Usually fish are perceived as silent creatures, so much that the phrase "as dumb as a fish" is a common saying to describe a close-mouthed person. On the contrary, fish make a lot of underwater sounds when moving, and many of them are also able to vocalize during female courtship and territorial defense. A recent research [1] has even mapped the neural circuitry for vocalization in larval fish and demonstrated a highly conserved pattern between fish and more recent branches in the evolutionary tree, including birds, amphibians and mammals.

Our installation, called "Musica sull'Acqua" (in English: *water music*), is an attempt to give the fish a sort of voice which is audible to the human ear, by tracking their movements and using them to drive in real time a computer-generated music performance. This process will be explained in detail in the next sections.

2. IDEATION AND DESIGN

Inside the *Civic Aquarium of Milan* there are many fish tanks which could be interesting for our goals, most of them reconstructing the natural habitat typically found in the north of Italy: lakes, rivers, swamps, and so on. One of the few exceptions is a large tropical fishpond which hosts many beautifully colored species and mesmerizingly waving sea anemones. Just for these reasons, namely for the great variety of species, each one with its own features and peculiarities, this site has been chosen for our installation. Fig. 1 provides a surface view of the fishpond.

As regards the physical layout of the devices, we placed an array of cameras along the whole width of the tank in order to perform fish tracking, as well as an array of speakers to play sounds. As shown in Fig. 2, the tank is divided in three sectors by four equally spaced columns, and this peculiarity has been exploited by placing both a speaker and a web camera near each of them. These couples of devices have been embedded inside a unique protective structure, which externally appears like a colored box (see Fig. 3).

Web cameras, placed on top of each speaker, have been accurately configured and pointed in order to capture a global panoramic view of the tank. An effort has been made to avoid overlapping regions or blind spots between adjacent cameras.

Cameras and speakers have been connected to a desktop PC, which has been programmed to process the four inde-

[1] http://www.avventuradellascienza.unimi.it

Figure 1. The tropical fish tank of the *Civic Aquarium of Milan*. In the lower left corner a sea anemone can be clearly recognized.

Figure 2. An overview of the installation. The big tank appears as segmented in three parts, due to the presence of pillars.

Figure 3. The protective structures, each containing both a loudspeaker (pointing outwards) and a web camera (pointing inwards). This image is a zoom on the central segment of the installation.

pendent video inputs through computer vision algorithms and to generate in real time the musical performance by using the extracted parameters.

Hardware has been protected by tweaked wooden cases, and cables have been run, fixed and hidden along the border of the tank and the columns.

The resulting quadraphonic setup allows sound spatialization for the whole length of the tank.

3. VIDEO CAPTURE AND COMPUTER VISION

In order to implement the installation, the first step consists in the extraction of visual parameters from the fish tank.

One of the key practical problems to solve was which kind of cameras to use for the project. Four possibilities have been considered:

1. USB web cameras;
2. DV camcorders;
3. analog AV cameras;
4. IP web cameras.

The latter ones were fit for our requirements. In fact IP web cameras are generally used for video surveillance, so they have to be both reliable and affordable: a perfect choice for a low-budget project like ours. Moreover, they are stand-alone devices which send a Motion JPEG (M-JPEG) stream through TCP/IP, [2] so they can work under any platform without the need of *ad hoc* drivers. Finally, they communicate through a wired Ethernet connection; as a consequence, they can be placed far from the receiver device, and a multiple-camera setup can be easily built by adopting a simple Ethernet switch.

Computer vision algorithms have been implemented using the *Processing Development Environment (PDE)*. *Processing* is an open source programming language and environment for people who want to create images, animations, and interactions, and it aims at learning, prototyping, and production activities. Software pieces written using *Processing* are called "sketches". This programming language has been adopted because of its portability (it is based on Java), its lively and supporting community and its open source license. [3]

The whole process can be briefly described as follows. The data streams coming from the cameras are first normalized, then a binary mask is created by subtracting the background and highlighting the pixels belonging to the fish. From the resulting binary image, blobs representing the fish are detected and their movement across consecutive frames is tracked. Finally, a set of features is extracted from each blob, e.g. its centroid, area and normalized color components. Now we will discuss each point in detail.

[2] TCP/IP stands for the Internet protocol suite, namely the set of communications protocols used for the Internet. The acronym comes from its most important protocols, i.e. Transmission Control Protocol (TCP) and Internet Protocol (IP).

[3] For further details on *Processing*, please refer to the official Web site at `http://processing.org`.

The first transformation of the incoming images is a conversion into a normalized RGB color space. This can be obtained through the following equations:

$$R = \frac{r}{r+g+b}, G = \frac{g}{r+g+b}, B = \frac{b}{r+g+b} \quad (1)$$

where R, G and B are the normalized red, green and blue components for each pixel respectively, while r, g, b are their counterparts in the original image.

This step solves the problem of light variations in the video footage, caused either by the environmental illumination system or by the automatic adjustments of exposition time and aperture of the cameras, which could not be deactivated for the models used in this project.

Then, a background-subtraction algorithm is used to separate the image areas representing the fish from the ones containing water, rocks and so on. This is still a challenging matter, where different techniques can give diverse results depending by the context they are applied to. Therefore, a review of the state of the art for background subtraction algorithms has been conducted in order to choose the most suitable solution. Please refer to Section 5 for an in-depth discussion on this issue.

Once foreground and background pixels are marked, the two pass connected component labeling algorithm described in [2] is used to detect blobs in the image. The final result can be improved by discarding blobs with a too small or too large area, and recursively merging blobs whose bounding boxes[4] intersect. Blobs are then tracked over time, using distance-based criteria as proposed in [3]. Each blob should now roughly correspond to a single fish and give information about its position, size, color, etc.

Finally, an Open Sound Control (OSC) message is generated for each blob in any frame [4]. By using OSC namespace, messages are labeled depending on which camera they come from and categorized into three different types:

1. *noteon*, if they correspond to a blob which has been detected for the first time in the current frame;
2. *update*, if they correspond to a previously tracked blob which is present also in the current frame;
3. *noteoff* if they correspond to a previously tracked blob which has been detected no more.

The latter messages carry the corresponding blob label as their unique argument, while both *noteon* and *update* messages contain the following information: blob label, x and y centroid coordinates, area, and average values for the normalized red, green and blue components of the blob pixels.

Fig. 4 illustrates blob tracking in *Processing* for the first video streams.

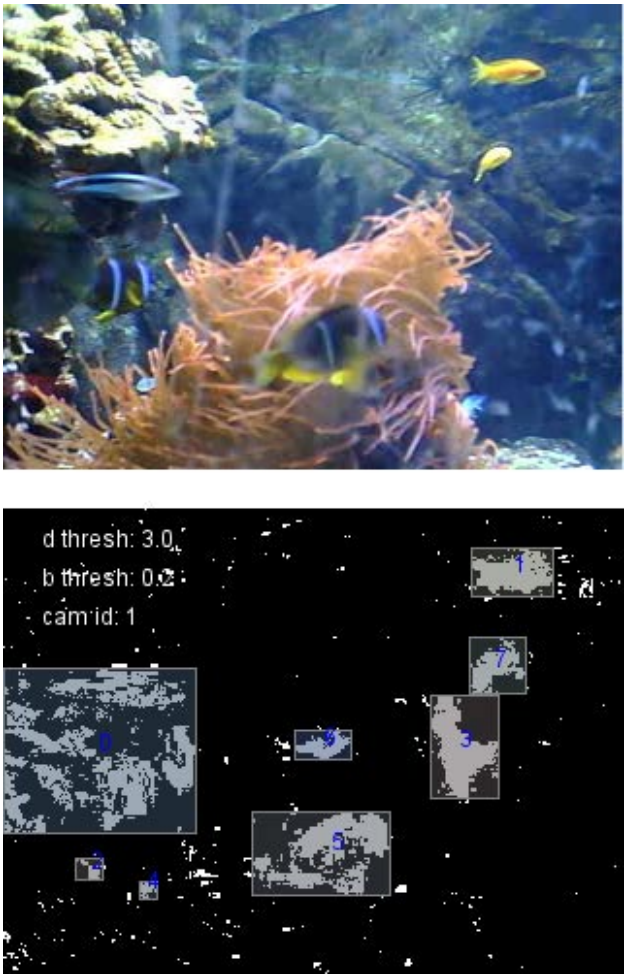

Figure 4. Blob tracking performed by *Processing*. Blob #0 derives from the intersection of two fishes, whereas blob #3 is due to the waving movements of sea-anemone's tentacles.

4. MAPPING VIDEO ONTO MUSIC CHARACTERISTICS

In the sound design, particular efforts have been done to reach two goals: the acoustic installation had to be pleasant and captivating from a musical and aesthetic point of view, being at the same time as self-explanatory as possible. Users should be able to immediately notice the connection between fish movements and the corresponding sounds.

In order to obtain these effects, such a mapping has been designed to simulate some perceptive phenomena typical of a real-world environment and to follow musical rules at the same time.

In our approach, each blob is considered as an independent sound source. It would not be completely correct to consider it a solo instrument, since its timbre can change according to blob parameters, as explained below. The correspondences among blob characteristics and audio features are the following ones:

- The position of the centroid along the horizontal axis has been related to quadraphonic spatialization and panning. When a fish moves from left to right, the corresponding sound follows it from the leftmost to the rightmost loudspeaker. This is also common practice in traditional stereophonic sound placement: sound sources placed to the left of the listener are panned towards the leftmost speaker, while sources placed

[4] For "bounding box" we mean the smallest rectangle containing all the pixels of a connected component.

to the right are panned towards the rightmost loudspeaker;

- The position of the centroid along the vertical axis has been related to note pitches. The underlying reason for this mapping is that the words and concepts used to describe pitch all refer to something vertical: pitch can be *high* or *low*; musical notes are arranged in scales, which can be played in *ascending* or *descending* order; note intervals can be thought as *jumps* between notes at different *heights* in the scale. When a fish moves upwards, it produces ascending sounds on a discretized set of admissible scale grades. The arbitrary choice of allowing only a fixed subset of frequencies is due to aesthetic considerations. The scale models supported by the installation are: diatonic (major and natural minor), whole tone, pentatonic, octatonic, and some other modes typical of jazz. Both the current scale model and its tonic are randomly selected at regular time periods, in order to prevent the sense of repetitiveness and immobility typical of some aleatory music;

- Blob area has been related to loudness, so that either a big fish or a fish passing near the camera produces a loud sound; similarly, a fish which is turning and pointing away from the camera produces a vanishing effect, due to its bounding box resizing. This mapping simulates the real world relation between loudness and object distance: sound intensity is generally low for very far objects, and increases as the sound source gets closer to the ears of the listener. In a similar way, also the perceived size of the object decreases as it gets further away from the eyes of the observer. In general, objects which appear big to the eyes often sound loud to the ears, and vice versa;

- The RGB components influence the timbre. Since timbre is usually referred as the "colour" of a certain tone, the choice was quite straightforward. Each blob generates a mixture of three sampled sounds. These sound components are weighted by the color components of the blob. Red is a color usually associated with fire, energy, blood and violence, therefore the red component has been mapped to the harshest sample; blue is associated with water, relaxation, sadness and cold, so it has been mapped to a mellow tone; green lies somewhere in the middle, thus the corresponding sample is neutral from this point of view. Fish whose color is completely different (e.g. brilliant yellow vs. dark gray) obviously generate different timbres, but once again smooth variations can be obtained even on a single blob, due to lighting or filter effects of the water.

- Finally, to add some more rhythm and dynamics to the performance, blob velocities have been derived by their position change and used to control other samples which, coherently with the purposes of the installation, are somehow related to a water environment: a basin, a toilet and a small aquarium waterfall. Those particular sounds vaguely recall the air friction on fast moving objects and are therefore an appropriate match to the visual parameter. The water-related sounds get louder when the fish accelerates, and softer as it decelerates. Such sounds are low-pitched (almost a rumble) for the big fish and high-pitched (similar to a hiss) for the small ones.

Thanks to the great number of fish and to their different physical characteristics and behaviors, the overall result is a very rich multiplicity of randomly-generated sounds. Nevertheless, the choice of picking notes from scales of our music tradition allows to obtain an euphonic and pleasant soundtrack for the tropical fishpond.

From a technical point of view, the sonification has been implemented as follows. OSC data coming from *Processing* sketches are received by a *Pure Data* (PD) patch. *Pure Data* is a powerful and versatile graphical development environment specifically designed for audio applications. Once again, this software is a multi-platform, open-source initiative. [5]

The resulting patch, partially shown in Fig. 5, implements the polyphonic synthesizer employed in this installation.

5. RELATED WORKS

Even if there are thousands of multimedia installations around the world, the approach adopted here is quite original, and - from the artistic point of view - few works can be compared to our experimentation.

In this context, a similar experience is the one of "Quintetto" by Quiet Ensemble, an installation based on the study of the casual movement of fish used as input for the production of sounds. The layout is composed of 5 vertical aquariums holding a fish in each as a video camera records its movements which are then translated into sounds through a computer software. The key differences are:

- The interpretation of single animals as separated sound sources, with no possibility to interact;

- The presence of some form of external control over the performance, e.g. as regards the lighting effects;

- A different mapping of the acquired characteristics onto a smaller set of sound parameters.

Further details and video recordings can be retrieved at `http://www.quietensemble.com/quintetto.html`.

It is worth to cite the "The Accessible Aquarium Project" `http://sonify.psych.gatech.edu/research/aquarium/` whose goal is to make dynamic exhibits such as those at museums, science centers, zoos and aquaria more engaging and accessible for visitors with vision impairments by providing real-time interpretations of the exhibits using innovative tracking, music, narrations, and adaptive sonification. See [5] for further details.

As mentioned in Section 3, one of the most relevant problems to solve was background subtraction. Luckily, this

[5] For further details on *Pure Data*, please refer to the official Web site at `http://puredata.info`.

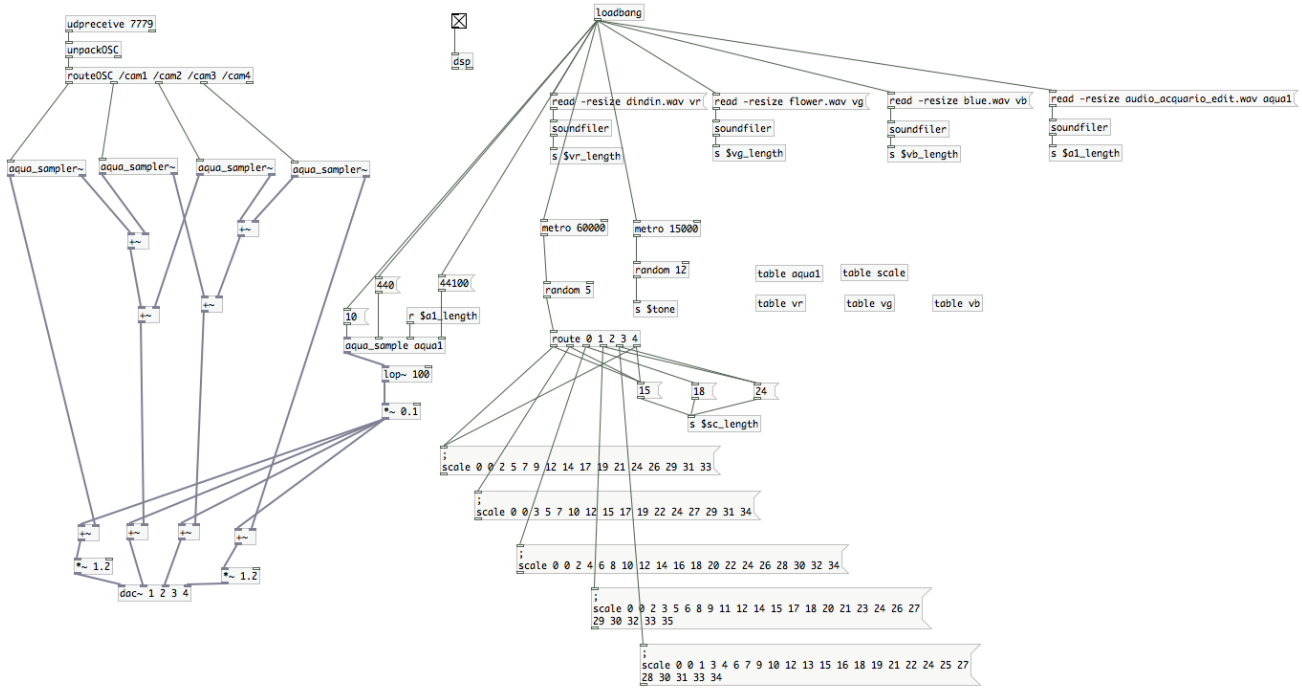

Figure 5. A part of the *Pure Data* patch that realizes the polyphonic synthesizer.

problem has already been discussed in scientific literature. The simplest form of background subtraction is frame differencing [6]: a reference frame representing the pure background is taken, and then subtracted from the successive frames. After the subtraction, pixels whose values lie above a certain threshold are considered as foreground, whereas the rest is marked as background. Unfortunately, this technique could not be used in our installation for two reasons: i) the difficulty in capturing a reference frame for the background inside a very crowded fish tank, and ii) the variability in the background caused by illumination changes, water ripples and so on.

To handle such a situation, a statistical representation of the background is required. The most straightforward approach is taking a window of frames and using the average or the median of pixel values as an estimation of the background [6]. The median is usually preferred over the average, as it is much less disturbed by the presence of outliers. Approximate solutions such as the running average or the running median usually offer a slightly worse accuracy, but they significantly reduce computational and memory requirements.

Other techniques implemented and tested for our application include kernel density estimators [7], texture background modeling through local binary patterns [8], a mixed approach using both texture and color information [9], single Gaussian [10] and mixture of Gaussian background modeling [11] [12]. The latter proved to be the best compromise among accuracy, computational cost and memory requirements for this case study.

6. CONCLUSIONS

This paper has discussed in detail the design and realization phases of the installation "Musica sull'Acqua" held at the *Civic Aquarium of Milan*.

The key idea is capturing the physical characteristics of the fish and their random movements around the tropical tank in order to produce a sonorization which is both pleasant and easily understandable for visitors. The former aspect has been realized through the use of mellifluous timbres and pitched sounds belonging to well-known scales of the Western tradition. The latter aspect has inspired an accurate choice of audio parameters to drive.

Acknowledgments

The authors gratefully wish to acknowledge Nicoletta Ancona from the *Civic Aquarium of Milan* for her cooperation in "L'avventura della scienza" initiative. This work has been partially funded by the *Enhanced Music Interactive Platform for Internet User (EMIPIU)* project.

7. REFERENCES

[1] A. Bass, E. Gilland, and R. Baker, "Evolutionary origins for social vocalization in a vertebrate hindbrain–spinal compartment," *Science*, vol. 321, no. 5887, pp. 417–421, 2008.

[2] L. Shapiro and G. Stockman, *Computer Vision*. Prentice Hall, 2002.

[3] A. Francois, "Real-time multi-resolution blob tracking," DTIC Document, University of Southern California, Los Angeles, USA, Tech. Rep. IRIS-04-422, 2004.

[4] M. Wright, "Open sound control: an enabling technology for musical networking," *Organised Sound*, vol. 10, no. 3, pp. 193–200, 2005.

[5] A. Pendse, M. Pate, and B. Walker, "The accessible aquarium: identifying and evaluating salient creature features for sonification," in *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2008, pp. 297–298.

[6] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings 2004 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4. IEEE Comput. Soc, 2004, pp. 3099–3104.

[7] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *Computer VisionECCV 2000*, pp. 751–767, 2000.

[8] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 657–662, 2006.

[9] J. Yao and J. Odobez, "Multi-layer background subtraction based on color and texture," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[10] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 780–785, 1997.

[11] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1997, pp. 175–181.

[12] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Cat No PR00149*, vol. 2, no. c. IEEE Comput. Soc, 1999, pp. 246–252.