Prototyping a New Audio-Visual Instrument Based on Extraction of High-Level Features on Full-Body Motion

Project Proposal for the eNTERFACE 2015 International Workshop

Principal Investigators

Joëlle Tilmanne[1], Nicolas d'Alessandro[1]

Team Candidates

S. Laraba[1], A. Moinet[1], A. Puleo[1], T. Ravet[1], L. Reboursière[1], M. Tits[1], F. Zajéga[1], A. Heloir[2] , R. Fiebrink[3]

Affiliations

[1]numediart Institute, University of Mons *(Belgium)*,
[2]LAMIH, University of Valenciennes *(France)*
[3]Goldsmiths, University of London *(UK)*

Abstract

Skeletal data acquisition generates a huge amount of high-dimensionality data. In many fields where mocap techniques are now used, practitioners would greatly benefit from high-level representations of these motion sequences. However, meaningful motion data dimensionality reduction is not a trivial task and the selection of the best set of features will largely depend on the considered use case, hence enhancing the need for a fast customization and prototyping tool. Numediart has worked this year on the first version of such a motion analysis framework: *MotionMachine*. The goal of this project is to build, based on *MotionMachine*, a new audio-visual instrument that uses full-body motion to drive sound and visuals. We will develop this new instrument as a proof of concept: the elaborate choice of higher-level feature extraction techniques will greatly improve human/computer interaction and lead to more expressive experiences.

Objectives

The main objective of this eNTERFACE project is to investigate the use of a fast prototyping motion analysis framework for the extraction of meaningful high-level features in motion capture sequences. A new audio-visual instrument using full-body motion to drive sound and visuals will be designed using high-level features extracted, thanks to the MotionMachine framework. In this new instrument setup, we will use sonification and visualization as a test bed for validating the relevance of the new features that we implemented. This vision has been shaped by many previous eNTERFACE projects, involving motion capture, real-time gesture recognition and innovative approaches towards statistical generation and mapping [1,2,3]. The design of this instrument will also be a test case for the real-time communication abilities of MotionMachine with external visualization or sound synthesis modules, as well as existing gesture recognition toolkits such as Wekinator [8]. For this workshop, we decided to focus on two main aspects:

Objective 1: Feature Extraction

Although motion capture technologies exist for quite a while, most works have focused on using the high-dimensionality data coming from data captures as such. However, past works have led us to believe that meaningful dimensionality reduction was one of the main topics to be addressed for most of the mocap research fields: analysis, synthesis, gesture recognition, etc. Proposals for very different motion features can be found in the literature [9, 10] and our purpose in this work is to gather some of them under a common easy-to-use framework. Moreover, the extraction of these features will be implemented as real-time modules which could both be used offline for prototyping and online for performance. The implementation of these features will be designed to be easily adapted for different data topologies (e.g. skeletons with a different number of nodes or joint names).

In addition to these built-in MotionMachine features, the communication of MotionMachine with existing motion recognition toolkits for feature extraction will be tested and implemented during the workshop. Continuous features extracted in MotionMachine will be selected as input to e.g. the Wekinator gesture recognition/mapping module [8] and the output of the recognition/mapping will be sent from Wekinator back to MotionMachine to be used as any other feature and be available as an output to drive the visualization or sound synthesis modules.

Objective 2: New Instrument Design

In order to evaluate the relevance of above-mentioned feature extraction strategies, we want to develop a new audio-visual instrument as a test-bed. The extracted features will be mapped onto two main modules: visualization and sound synthesis. The visualization research will take place in two different environments: web and 3D software like Blender, respectively extending XML3D [11] and Tanukis [12] projects. The sound synthesis work will extend the point cloud pitch-synchronous approach used in HandSketch [13], developed for many years in previous eNTERFACE projects and now on its way to be release as a commercial framework. HandSketch uses a descriptor-based approach to synthesis, enabling descriptor-to-descriptor high-level mapping strategies. The main objective is to use the feature extraction tools within a fast prototyping loop, driven by the need to create an expressive instrument.

Background

Over the ten last years, an important amount of motion capture techniques have emerged. However most of these techniques, such as inertial suits[1] or optical markers tracking[2], did remain expensive and cumbersome. More recently, the democratization of *depth cameras* – like the Microsoft Kinect – has considerably changed the scope of markerless motion capture research. Indeed the massive dissemination of these sensors gave many new research groups the chance to jump in this field and then provide various resources (databases, software, results) to the scientific community [5]. This technological breakthrough has brought motion capture into new application domains, like health [6] and the arts [7].

Skeletal data acquisition generates a huge amount of high-dimensionality data. Such raw data is neither very readable, nor very reusable. The design of interesting mapping strategies for gesture-controlled interaction is often hampered by these considerations and very basic gestures are usually selected for such interaction because of the lack of existing or readily-available higher-level motion features. In many fields where mocap techniques are now used, practitioners would greatly benefit from high-level representations of these motion sequences. Such features could be understood by sight or highly related to the studied phenomenon (expertise, style, etc.), so that they could be mapped with sonification, visualisation or used in machine learning. In literature, we can find e.g. top-down description of full-body motion [14], geometrical descriptors of the 3D skeleton [9] or application of dimension-reduction techniques to extract higher-level aspects of gestures. Our perspective on this situation is that there is no generic solution to motion high-level feature extraction. The optimal solution will be highly dependent on the use case. There is therefore a need for a very usable prototyping platform, where a set of feature extraction modules can be quickly plugged in and their efficiency evaluated in real-time, both from offline and online skeletal data. At numediart, we have developed such a tool for real-time prototyping of motion data processing, called *MotionMachine*. It is used as an independent feature extraction layer, taking OSC 3D data from various sensors and streaming OSC features in real-time. It is a C++ library and it currently has bindings with openFrameworks for the visualisation module.

During eNTERFACE 2013 and 2014, our team has worked on the problem of gesture recognition based on skeletal data. These early works have demonstrated the need for high level motion feature extraction in order to reduce the dimensionality of the motion data and hence reduce the complexity and augment the performances of motion recognition models, especially when the training data is scarce.

Several tools exist to manipulate and visualize motion capture data with the goal of creating motion-enabled applications. The MOCAP toolbox [10] is among the very popular tools, giving access to many high-level processing functions, but it is only available for Matlab and therefore not very suitable for real-time, iterative and interactive testing and performance. When it comes to performance-driven tools, we find software like the RAMToolkit [15] with the opposite issue: the tool is not generic and rather very specific to a single use case (Kinect-captured dancers). In the field of mapping software for new interface design, the complexity of motion capture data is still not properly faced. For instance, tools like Max or LibMapper [16] lack of serious MOCAP toolkits with high-level feature extraction mechanisms. Most of the time, such abstraction of raw data is achieved and fine-tuned for a specific use case.

---

[1] MetaMotion IGS-190: http://www.metamotion.com/gypsy/gypsy-gyro.htm

[2] NaturalPoint OptiTrack: http://www.naturalpoint.com/optitrack

Technical Description

In this project, we will design a new audio-visual instrument that uses full-body motion to drive sound and visuals. We will develop this new instrument as a proof of concept: the elaborate choice of higher-level feature extraction techniques can greatly improve human/computer interaction and lead to more expressive experiences. Practically, we will convert on-the-fly skeletal data from the Qualisys or the Kinect into a set of higher-level features with *MotionMachine*. Some of these features corresponding to gesture recognition will be extracted through the Wekinator platform, connected to MotionMachine. Then these features will be used to manipulate parametric visualisation and sound synthesis. Our purpose is to use visualisation and sonification as a test-bed for validating the relevance of the explored features. We want these features to lead to more expressive interaction.

This part of the proposal gives more details on the different technologies that are envisioned and gives the main research and development axes that will be followed in order to build the new system. We also give greater insights about the devices, environments and prototyping strategies that will be aligned in this project. Finally we also describe the project management that will be deployed. In this section, most of the following text refers to module names that are depicted in Figure 1. We also highlight the workpackages of the project and introduce a naming convention (WP_N) that will be reused in the section where the schedule is described.

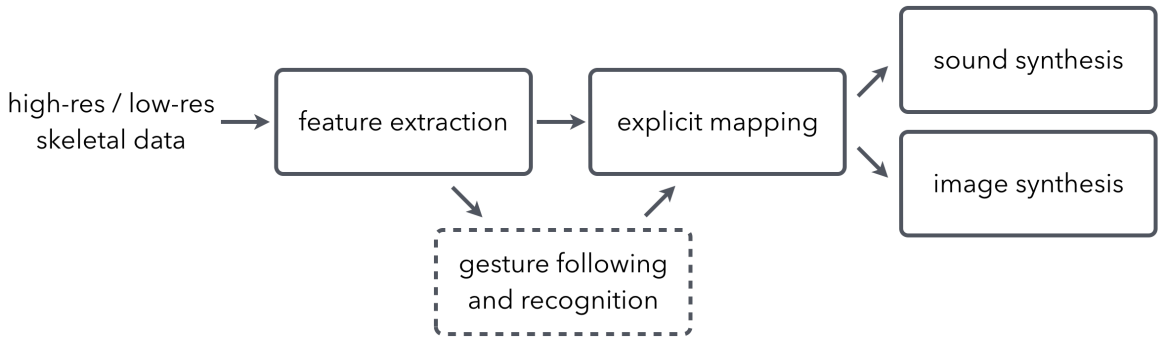

Figure 1 – Processing pipeline for the instrument prototyping use case.

Research and Development Axes

The new digital instrument developed in this project will be composed of five main components, identifying the groups of researchers working on these problems: the extraction of relevant high level features based on skeletal data (either from Qualysis or Kinect), the visualization task, the sound synthesis module, the development of mapping strategies and finally the assessment of the whole system:

1. Feature Extraction (WP_1)

As described in the Background section, the high dimensionality of motion capture data makes it hard to make use of these data as such. In this work package, some high level features will be implemented within the MotionMachine framework in order to give readily available features to the naïve users who want to design new gesture controlled interfaces. Thanks to the coupling of motion extracting features with basic visualization of said features, the feature selection for mapping and visualization will be easier and more natural. Some features such as the 39

descriptive features proposed by Müller [b] have already been implemented and a list of most relevant features will be collected by the projects participants interested in motion analysis and discussed with the artists and motion recognition experts.

Furthermore, the implementation of these features will be made robust to different skeleton hierarchies (e.g. Kinect or optical mocap data) for the system to be able to generalize easily between different mocap systems.

Coupling of the MotionMachine feature extraction with external modules such as Wekinator for gesture recognition and mapping will also be investigated, since our goal is not to re-implement existing tools but rather to offer a framework in which motion can easily be analyzed and meaningful features extracted, selected and understood thanks to the visualization.

2. Mapping (WP_2)

The output of WP_1 will be a series of motion descriptors based on feature extraction and gesture recognition. The mapping task is about defining and testing user interaction scenarios, i.e. relations between the calculated motion descriptors and the audiovisual space with which we aim to interact. Mapping creates one-to-one, one-to-many or many-to-one connections between inputs and outputs. In a previous eNTERFACE workshop [2], we explored the descriptor time scales as a mapping strategy. Indeed we have combined short-term mapping (decisions taken on the input data value frame-by-frame) and long-term mapping (observation of input data value in the context of its past behavior: slope, min/max, peaks, pattern, etc.) mechanisms. As MotionMachine can act as a realm-time mocap data ring buffer, we think that this is particularly appropriate to opt for such a design approach.

3. Visualization (WP_3)

In this WP, two different use cases of visualization will be tackled:

- Mocap in the browser: On the one hand, modern browsers that implement the full HTML5 specifications offer a large choice of UI elements capable of displaying complex datasets and foster rich interaction. On the other hand, the new XML3D standard [11] offers an HTML5 namespace letting any traditional web developer edit sophisticated 3D scenes inside the DOM. By combining these two technologies, it is possible to get the best of both worlds: merging together high-end interactive graphics and high-level UI libraries.

- Tanukis: We will extend an advanced 3D avatar animation framework, called Tanukis (developed in Blender) [12], exploring the use of the extracted high-level features for changing various properties of the character animation and rendering. Indeed the direct use of mocap for character animation is an opaque approach and it is very difficult to change such data in a creative way. We think that using high-level motion features along with the raw mocap data can greatly enrich the animation possibilities. For example, using gesture dynamics could help to drive a given exaggeration in the animation.

4. Sound Synthesis (WP_4)

In this workpackage, we will extent the HandSketch sound engine [13]. This engine analyses a database of monophonic sounds (speech, singing, woodwind, etc.) with a PSOLA-like algorithm, extracts several features on each sound grain and uses these

features to locate each grain inside a 2D or 3D space. This creates a “point cloud” representation of the sound database on which we can then map a given controller.

In this project, we want to greatly extend the feature extraction part of the engine, by adding descriptors and adapt the algorithm to more types of sounds. We also want to explore the idea of finding and defining sound trajectories within the space, based on the long-term evolution of the original sound. The HandSketch sound engine has mainly been used with highly melody-based multitouch interfaces. Hence we will need to integrate several adaptations, so as to consider its usage with full-body motion features. Indeed there is a need for the sound engine to take more musical decision towards the input motion, considering a full-body skeleton is less keen to directly drive melodic contents. The integration of long-term feature evolution, both on motion and sound, is an interesting starting point to consider.

5. Fast Prototyping Framework (WP_5)

The key idea in order to validate the overall musical instrument design approach is to be able to put such scientific techniques (motion capture, machine learning, visualization and sound synthesis) into a prototyping framework that is fast, interactive and expressive. We consider that it is very meaningful to design and evaluate the overall framework to be as interdisciplinary as possible. Indeed this work stands between different kinds of expertise, and ultimately the need for artists to put their hands on key technologies so as to create a new instrument. Defining a new appropriate design language is a transversal objective of this project.

The MotionMachine toolkit is currently a C++ library, interfacing openFrameworks for visualization. We want to keep this core tool and work on the communication with various external software that are acknowledged for a give task (machine learning, web visualization, avatar animation, sound synthesis), while keeping a clear and consistent API to design the instrument.

Prototyping Cycle

As in any HCI application development, the team workflow is made of iterations between various phases, including research & development (as described above) but also updating the musical scenario on which we are working (WP_6) and validating our instrument and results in front of a panel of external observers (WP_7). The main aspect is the evaluation of the overall UX (full-body musical instrument) and the system outputs (visuals and sounds) by external listeners/observers. We wish to demonstrate our ongoing prototype to as many eNTERFACE researchers as possible and establish a first informal benchmarking of successful strategies.

Facilities and Equipment

The team will essentially work with available devices brought by the participating labs. Obviously we will bring our own laptops. Moreover, we will have at our disposal several acquisition systems, Kinects, LeapMotions, and maybe the Qualysis optical motion capture system installed in the numediart room. We would also require some space for setting up our motion capture test space (either with Kinect or access to the numediart room for using the Qualysis system).

Project Management

Joëlle Tilmanne and Nicolas d’Alessandro will supervise the whole project. They should stay on the site of the workshop for the whole period. Based on the

subscribed participants, sub-teams will be gathered around the specific workpackages of the project. The methodology that is promoted in this project aims at staying flexible and adapt to our successive prototyping cycles. We will work with guidelines inspired by various Agile techniques, such as organizing scrum meetings or collectively defining the development backlog.

Project Schedule

In this part we gather the various workpackages that have been highlighted in the technical and set them down on a one-month schedule, plus some extra tasks:

- **WP_1 – Feature Extraction**: development of new motion feature extraction techniques based on the full-body skeletal data (database and streaming);
- **WP_2 – Mapping**: exploration of mapping strategies for the connection between full-body motion, visualization and sound synthesis;
- **WP_3 – Visualization**: development of new visualization of motion data, both on the web and within the Tanukis avatar animation framework;
- **WP_4 – Sound Synthesis**: extension of the HandSketch sound engine so as to tackle the new musical experiences (working from full-body data);
- **WP_5 – Fast Prototyping Framework**: development of an integrated API for sketching new musical instruments, communicating with various tools;
- **WP_6 – Iteration on the Musical Scenarios**: inline reassessment of the digital musical instrument scenarios that we use in our case study;
- **WP_7 – Assessment of UX and Results**: organization of external observation regarding our overall UX and the visual/sonic results produced by our system;
- **WP_8 – Reporting and Publishing**: general dissemination tasks.

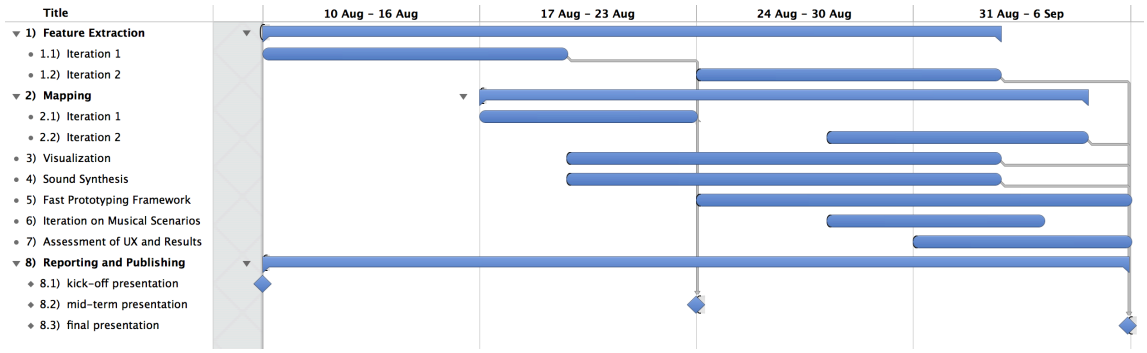

Figure 2 – Scheduling of the project (workpackages and milestones).

Deliverables and Benefits

Here are the main deliverables and benefits that the team will provide:

- software component for the high-level feature extraction algorithms
- software configurations for the recognition of extracted feature patterns
- software component for the extended HandSketch audio synthesis engine
- new interactive setup corresponding to the musical instrument use case
- new software and models for visualizing mocap data: web and Blender
- open sessions during the workshop for welcoming observers
- a scientific report, distributed in the required format

Team Profile

Project leaders: J. Tilmanne (motion capture, motion analysis),
N. d'Alessandro (real-time systems, sound synthesis, HCI).

Team proposed: A. Moinet (software design, sound), T. Ravet (software design, motion capture), S. Laraba (motion capture), A. Puleo (software design), M. Tits (motion analysis), R. Fiebrink (machine learning, sotware design, HCI), A. Heloir (animation, web software), L. Reboursière (computer music), F. Zajéga (visual art).

Collaborators that we are looking for: As described in the project schedule, this workshop will need pretty advanced software developers for most of the time. The first half of the month will be more oriented towards data analysis and statistical modeling, as the second half will require expertise in real-time applications. The second half of the project will also involve more testing of the synthesis results and human-computer interaction properties of our software. Therefore for this second half, we are also looking for HCI or Cognitive Sciences profiles.

References

[1] J. Tilmanne *et al.*, *"A Database for Stylistic Human Gait Modeling and Synthesis,"* Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, pp. 91-94, 2008.

[2] M. Astrinaki *et al.*, *"Is This Guitar Talking or What?"* Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, pp. 47-56, 2012.

[3] N. d'Alessandro *et al.*, *"Towards the Sketching of Performative Control with Data,"* [to appear in] Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, 2013.

[4] S. Mitra and T. Acharya, *"Gesture Recognition: A Survey,"* IEEE Trans. on Systems, Man and Cybernetics, C: Applications and Reviews, vol. 37, n° 3, pp. 311-324, 2007.

[5] Z. Zhang, *"Microsoft Kinect Sensor and Its Effects,"* IEEE Multimedia, vol. 19, n° 2, pp. 4-10, 2012, DOI: 10.1109/MMUL.2012.24.

[6] E. E. Stone and M. Skubic, *"Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment,"* International Conf. on Pervasive Tech. for Healthcare, pp. 71-77, 2011.

[7] Y. Kim, M. Lee, S. Nam and J. Park, *"User Interface of Interactive Media Art in a Stereoscopic Environment,"* Lecture Notes in Computer Science, vol. 8018, pp. 219-227, 2013.

[8] R. Fiebrink, P. R. Cook and D. Trueman, *"Human Model Evaluation in Interactive Supervised Learning,"* Proceedings of the SIGCHI Conference on Human-Computer Interaction (CHI'11), Vancouver, BC, May 7-12, 2011.

[9] M. Müller, *"Information Retrieval for Music and Motion"*, Springer-Verlag, 2007.

[10] B. Burger and P. Toiviainen, *"MoCap Toolbox – A Matlab Toolbox for Computational Analysis of Movement Data,"* Proc. of the 10th Sound and Music Computing Conference, (SMC). Stockholm, Sweden: KTH Royal Institute of Technology, 2013.

[11] F. Klein, T. Spieldenner, K. Sons and P. Slusallek, *"Configurable Instances of 3D Models for Declarative 3D in the Web,"* Proc. of the 19th International Conference on Web 3D Technology (Web3D), pp. 71-79, ACM, 2014.

[12] *Tanukis*, http://frankiezafe.org/index.php?id=239

[13] N. d'Alessandro and T. Dutoit, "HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet" Proc. of International Conference on New Interfaces for Musical Expression, pp. 78–81, 2007.

[14] R. Laban, "A Vision of Dynamic Space," *London: The Falmer Press*, 1984.

[15] *RAMToolkit*, http://interlab.ycam.jp/en/projects/ram/ram_dance_toolkit

[16] *LibMapper*, http://idmil.org/software/libmapper