Techniques for automated analysis of saliency in human full-body movement

Serafino Gabriele
Lo Luca

Advisors:
Volpe Gualtiero
Camurri Antonio

2024

Abstract

In contemporary research, the topic of salience has become a focal point, with a growing number of studies dedicated to its exploration. Various fields, including Medical, Financial, Military, Psychology, Kinesiology, have approached the problem in diverse ways.
Our thesis aims to develop effective heuristics and algorithmic approaches for detecting salience in the context of movement. Specifically, we focused on the domain of dance movements, defining a concept of salience within this context.
In our purpose the salience is meant as a change in the behaviour of the dancer: static movement with respect to dynamic movement, repetitive movement with respect to a chaotic one, using an arm and then stopping it for using the other one, ans so on...
The study involves a thorough analysis of Motion Capture (MOCAP) data, where we navigate dataset constraints and boundaries to select significant features that aptly represent our chosen domain.
To address the salience detection problem, we employ several approaches: among them, a Supervised Machine Learning (ML) Algorithm.
For the ML approach we defined a specific prototype of Sliding Windows (SW) able to convey information through several frames of the videos, with the goal of giving to the specific model the most significant piece of information in order to have the best results.
Then we tested the algorithm at different levels of generalization.
The goal of this work is to develop a system able to recognize the salience in human full-body context in real time.

Contents

List of Figures

List of Tables

Acronyms

3D Three dimensional. 10

CPD Change Point Detection. 14

EST Event Segmentation Theory. 11

ML Machine Learning. 2, 15

MOCAP Motion Capture. 2, 9, 10, 14

QTM Qualisys Track Manager. 10

SW Sliding Windows. 2

TSV Tab-Separated Values. 10, 12

1 Introduction

The thesis focuses on defining movement saliency and developing a model for its automatic detection and analysis. The work uses the datasets from the archives of Casa Paganini - Infomus, an international Research Center, and consists of videos of dancers performing on-site captured using a motion capture techniques.

1.1 The problem of automatic salience

1.2 Motivation and goals

1.3 Thesis structure

2 Methodologies

2.1 CPD algorithms

2.2 Related Work

The material we started from was provided by researchers at
Casa Paganini - InfoMus [5]. By exploiting the existing results, our goal was to find algorithms and heuristics for dealing with the problem of salience.
The given scenario is the following: three dancers from Casa Paganini - InfoMus [5], Cora Gasparotti (Figure 1), Marianne Gubri (Figure 2) and Muriel Romero (Figure 3), performed dancing movements by emphasizing some specific tasks, in order to highlight a combination of graceful and fluid movements, as well as angular and hurried gestures.

Those performances has been recorded with two professional cameras (frontal and lateral view, 1280 $\times$ 720, 50fps) at Casa Paganini, and the result of this procedure was a set of *takes* (each *take* has a duration between 30 seconds to 180 seconds).
So, for each dancer, we have between 5 to 15 *takes*.

Figure 1:
Cora Gasparotti

Figure 2:
Marianne Gubri

Figure 3:
Muriel Romero

Raw data has been collected by using a Qualisys [6] Motion Capture (MOCAP) system:

- Qualisys Cameras: these are specialized for capturing and tracking movements in three-dimensional space.
  These cameras operate based on the principle of detecting and analyzing markers placed on objects or individuals within their field of view. The cameras emit light, which is then reflected by small, passive markers, such as reflective spheres or retro-reflective devices. This process allows the cameras to discern and meticulously track the position and movement of each marker within their field of view.
  In our context, a Qualisys system endowed of 16-cameras was used (fs = 100Hz). (Figure 4)

Figure 4: Camera Qualisys

- Markers: are integral components of motion capture systems, they are identifiable points placed on objects or the human body. These markers can take various forms, such as reflective spheres, strips, or discs, each designed for specific tracking purposes.
  In our cases we used 64 spherical markers positioned on the whole body. (Figure 5)

Figure 5: Markers Qualisys

The MOCAP data was collected in a Tab-Separated Values (TSV) file by using Qualisys Track Manager (QTM) [6] which is a software used for files handling MOCAP data, hence for each frame the position of each marker in the 3D space (x,y,z) is available.

2.2.1 State of the art

One of the most important work related to our context was the one done by E. Ceccaldi in her PhD Thesis [2], in that discussion, the problem of salience detection, has been defined with some psychological approach related to the body behaviour, for the definition of the salience, *takes* have been segmented (in the time).
In this work, segmentation is based on boundaries between events, as described by a cognitive theory on how this process unfolds in the human mind, namely the Event Segmentation theory [8]. According to this theory, a boundary is perceived whenever a meaningful (i.e. salient) change is perceived in the ongoing situation. Following the theory, in [2] changes were operationalized as follows:
Thus, salience is defined to a higher level as follows:

- C1. Time: which is the timing and the rhythm of the movement, for instance if the dancer accelerates her actions.
- C2. Space: if the attention and the direction of the movement starts to pointing something different.
- C4. Character Location: if the character moves on the stage, or, in our cases, if the dancer moves from a point to another.
- C6. Causes: Which are the causes and appraisal, if a new state of affairs leads to a subsequent event.

Hence the ground truth was taken by considering these 4 different aspects by psychologists which are trained on EST [8] principles by using a software for segmentation: ELAN [1] (Figure 6).
This software allows to make the segmentation of videos (or an audio track) by defining some classes of event segment and a hierarchy among them.
The *features* which have been considered in that case are the following:

- For the *Time* it was used the General Quantity of Movement and the Chest Quantity of Movement.
- For the Space the author employed the Directness of Head Movement.
- The Character Location has been detected by looking the Density of Chest Trajectory.

- The causes, instead, has been used only has an additional ground truth.

All these features has been extracted by using an application (patch) developed for EyesWeb XMI (Figure 7), a software platform able to build over the TSV, which contains our raw data, some useful characteristics such as: kinetic Energy, Point density, Directness, and so on...
For instance, kinetic energy (of all the markers) was taken as a measure of the overall amount of movement.
The Chest Quantity of Movement was evaluated by considering the kinetic energy of the marker related to the chest. The Directness of Head is a measurement of how the dancer's head moves in curvilinear trajectory, the higher is the value, the more the movement following a straight line.
For finishing, the Density of chest trajectory is an indicator of whether movement is localized in a small region in the space rather than spanning the whole space, i.e., higher density indicates that the actor has moved in a smaller region.

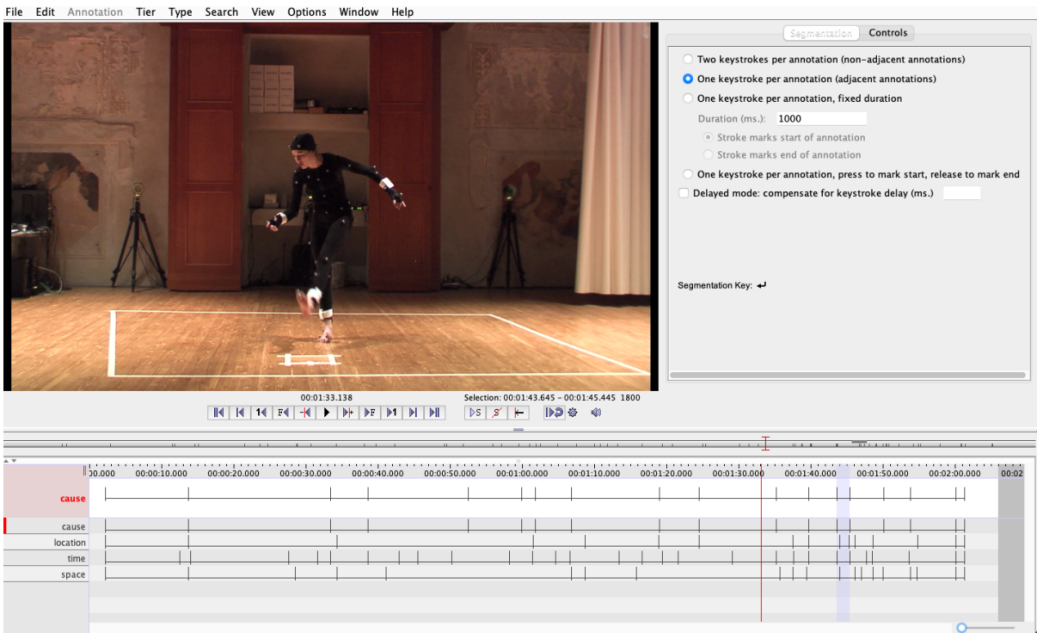

Figure 6: Annotation of the different salience with ELAN.

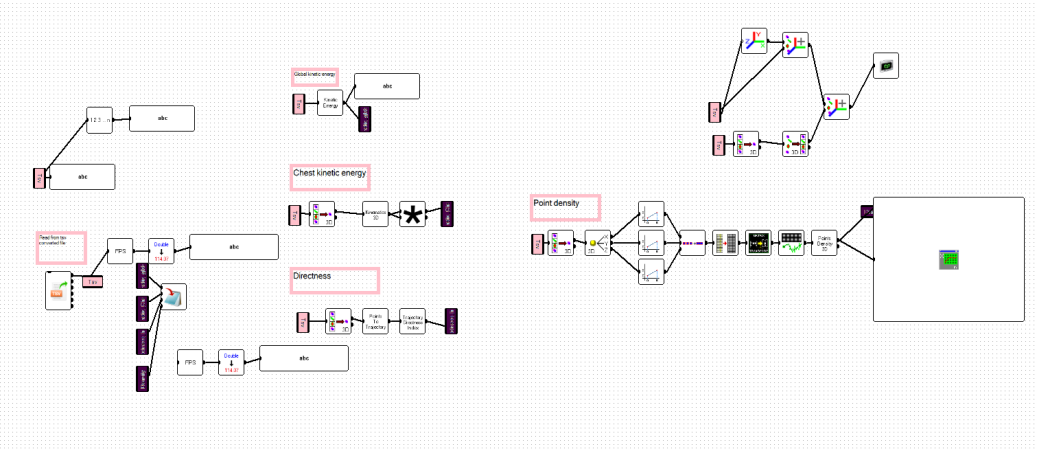

Figure 7: One of the patches implemented on EyesWeb for extracting the features.

2.3 Structure of workflow

In this chapter it will be explained which is the workflow that we did during our study (figure 8).
From the raw data analyzing to the salience classification and experiment testing:

1. Physical Layer: this layer was already achieved since the Motion Capture was given by the researcher at Casa Paganini - InfoMus [5] (chapter 2.2).

2. Salience and Ground Truth: After reviewing the material provided at the previous step, we evaluate which is the better definition of salience in our context, because we had to take in account the different possible movements of the dancers.
   Then, we wrote down the Ground Truth based on that definition of salience (chapter 3).

3. *Takes* selection: With respect to the point 2 we had to choose wisely the *takes* which better represent the context, because not all the videos were adapt for representing our domain (chapter 5.1).

4. Data Cleaning: We cleaned up the data with a Linear interpolation (chapter 5.1.1).

5. Low Level Features: Then, we have defined the *features* at a lower level, global kinetic energy, Repetitiveness etc.
   Hence these features have been collected thanks to EyesWeb (chapter 4).

6. Model Selection: We had explored two different approach for addressing the problem, a machine learning approach, and a Change Point Detection (CPD) technique (chapter 5 or 6).

7. Sliding Windows Template: For the specific model we built a template of sliding window able to convey the right piece of information basing on the used model (chapter 5.1.2 and 6.1).

8. Mid Level Features: For the Machine Learning (ML) approach, mid-level features have been selected, such as mean, variance, entropy etc. (Chapter 5.1.3)

9. Samples Definition and Downsampling: then for the ML model a specific definition of samples has been set (chapter 5.2).

10. Data Normalization: Data has been normalized (chapter 5.1.4)

11. Algorithm Selection: For the specific approach, a specific algorithm was used, for Machine Learning (ML), a Random Forest was employed for doing the classification (chapter 5.3).

12. Parameters Definition: Basing on the approach used parameters have been chosen, for ML a cross-validation has been done (chapter 5.3).

13. Model Evaluation: For finishing, different level of generalization has been explored (chapter 5.4).

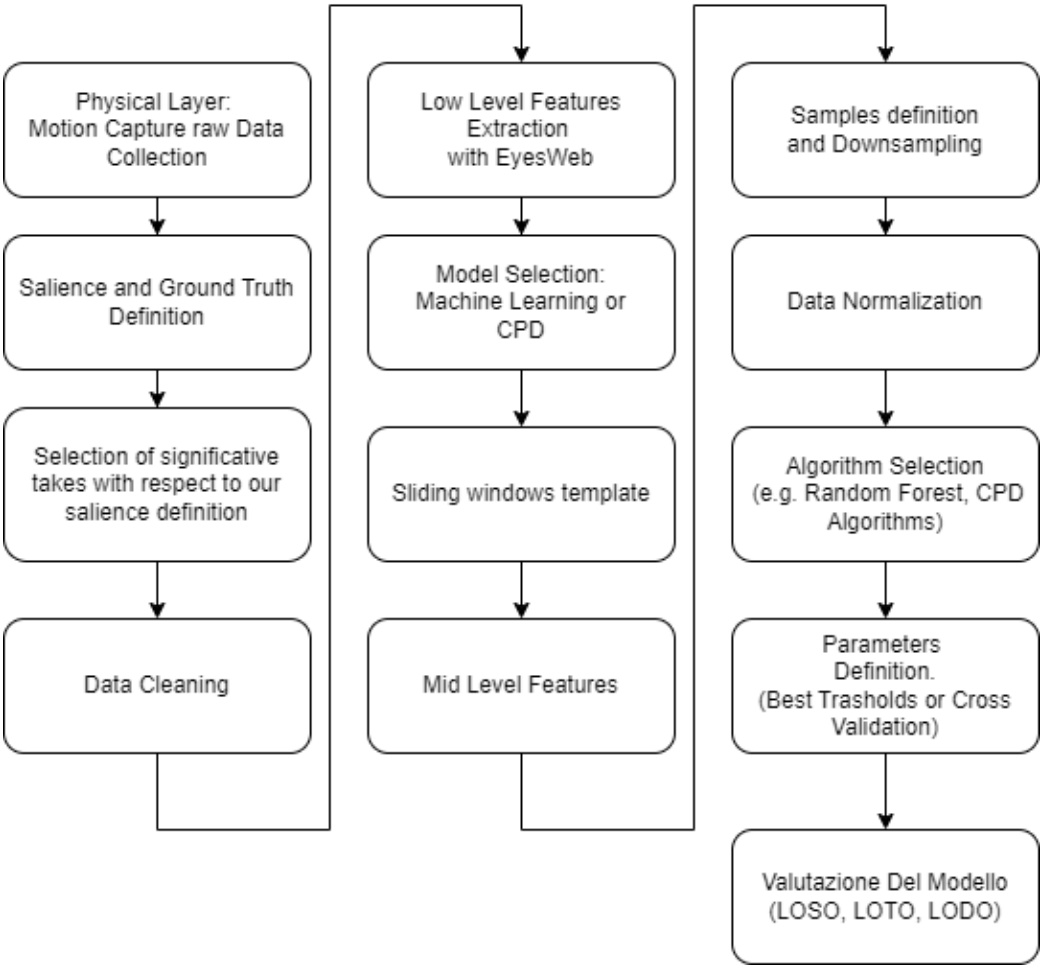

Figure 8: Workflow chart

3 Salience definition

When we started to approach this problem, one of the first challenge that we had to deal with, was to define what is a specific event that change in some way the behaviour of the dancer.
We address this event as a ***Salience***.
The salience in the context of movement can be viewed in different ways, from an holistic one, to a granular one, by considering a movement within a large temporal window, or by viewing it with a short temporal span.
The problem can be addressed from different point of view, for instance, you can make a comparison between a movement which is more *impulsive* or more *sudden* [3], or you can look at the movement in terms of *Lightness* or *Fragility* [4].
The first issue that we have noticed is that, if we thought the salience as something which is too much related to the causes, and the psychological aspect of the human behaviour, any algorithm over that kind of approach gave us bad results, because addressing an intention or an emotion is a tricky challenge.
So, we have focused on looking more the movement of the dancer from the point of view of the Kinesiology, so if there is a movement behaviour which is similar with the observation of the whole set of frames which precedes that specific frame, then for us is not a salience, if instead the frame subjected to evaluation is the first frame before a whole set of frame which has different characteristics in therms of movement with respect to the set of frames before the specific evaluated frame, then for us it is a salience.
After that, by looking the set of salience resulting on the step before, and since the quantity of frames labeled "salience" is significantly lower with respect to the frame which are labeled "non-salience", we became aware that for an algorithm of machine learning this scenario could be a problem, because unbalanced dataset are often hard to address [7]. For trying to reduce the oddness between the two classes we had defined the concept of event as something which is more frequent, by using a more granular segmentation, hence visible changes in the movement such as an arm which starts to move with respect to the other one, or a repetitive movement and then a chaotic one, and so on...

3.1 Ground Truth

Then we thought to rebuild the Ground Truth, so with ELAN [1] we took annotations of all the events building upon the considerations made before (Figure 9). Since each event is detected between two sets of consecutive frames, events can be addressed as follows:

- S_1 - This event occurs when the dancer moves from a specific position in the stage to another position in the stage.
- S_2, S_3 - This event detect a change in the movement from the point of view of the repetitiveness, hence, if the dancer moves from a repetitive and orderly action, to a chaotic one (S_2), or vice versa (S_3).
- S_4, S_5 - It occurs when the dancer moves from a crouched position to an extended one(S_4), or vice versa(S_5).
- S_6, S_7 - Event occurred when the dancer moves from a static position of the head to a dynamic one (S_6), or vice versa (S_7).
- $S_8 - S_{23}$ - The event which occurs when the dancer change the dynamic of her distal points: for instance if she moves her left wrist in the first set of frames, and then moves her right ankle, or if she moves her left ankle, and then moves her right wrist, or if she moves both the wrists and then stops to move one of the other distal parts.
  Since there are 4 different distal parts, left wrist, right wrist, left ankle, right ankle, and all of them can be moving or not, we can think them as a set of boolean variables (1 if moving, 0 if not), we can think the vector v_0=[0,0,0,0] as the vector which represents the situation in which the dancers does not move anything, and the vector v_n=[1,1,1,1] in which the dancer moves all the distal parts (n=15).
  Thus we have 2^4 different states (st_s) and hence 2^4 different events since each one represents the transition from st_i to st_j ($i, j \in$[0,16), i≠j).

So, in total we have 23 different kind of salience.
The salience from S_1 to S_5 consider a more holistic change between the two set of frames, instead the set of salience which goes from S_6 to S_{23} is referred

to a local movement.
During the phase of taking the Ground Truth, we did not make meaningful segments, we just consider the events occurrence, then for avoiding perceptual fusion problem (two different frame are perceived as one if the time between them is lower then 10 ms), we took the start of the segment (which is our salience) by looking frame per frame in the area of the perceived salience. For finishing, all the starting frames of the events have been exported in a text file where each value in that file represent the timestamp in seconds (with centesimal precision) of an occurrence of a frame representing a salience. For having the number of the frame instead of the timestamp we have computed the following formula: n_f=floor(n_s*50), since we have 50 frames per second (floor is the rounding down).

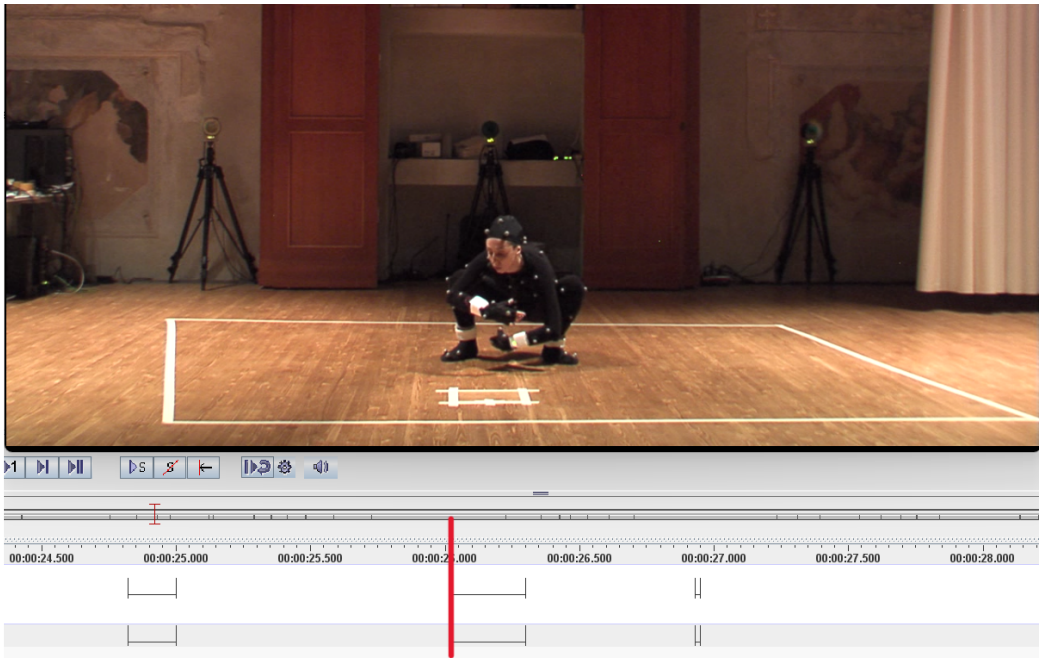

Figure 9:
The red line is over
the frame were
the salience is annotated (S_4).

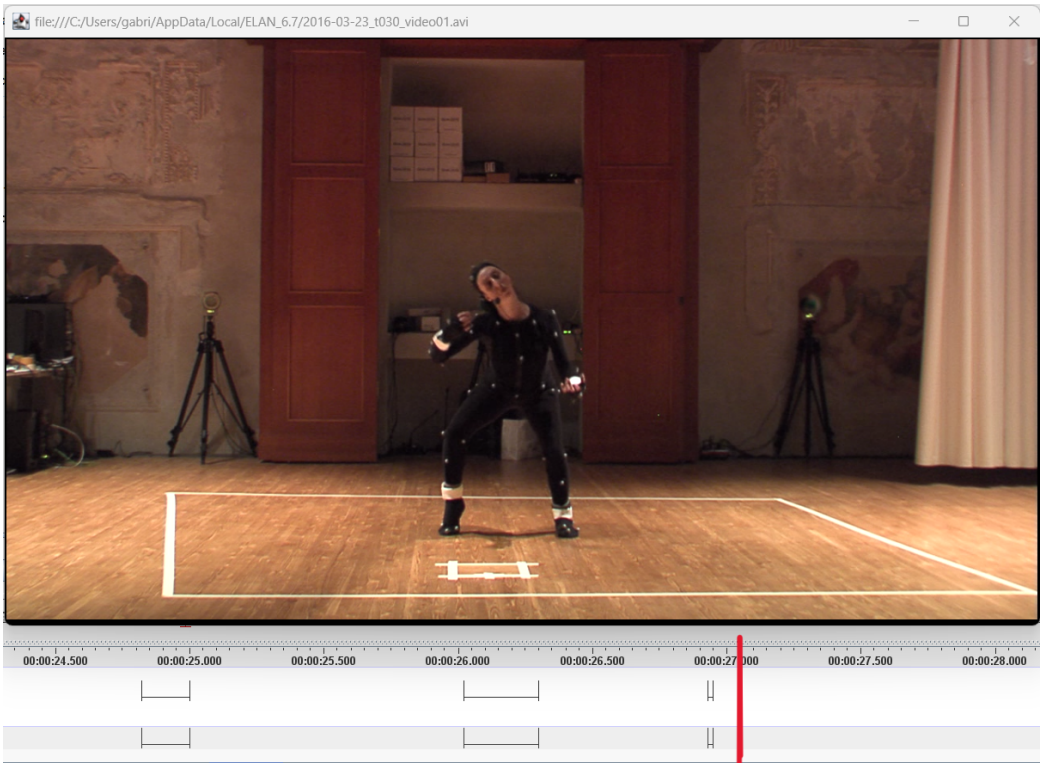

Figure 10:
After ≈ 50 frames (1 second)
with respect to the
salience in Figure 9
the dancer is extended.

4 Low level features

5 Supervised ML approach

5.1 Data

5.1.1 Data Cleaning

5.1.2 Sliding Windows

5.1.3 Mid level features

5.1.4 Normalization

5.2 Samples Creation

5.3 Model Selection

5.4 Results

5.4.1 LOSO

5.4.2 LOTO

5.4.3 LODO

6 Statistic approach

6.1 Sliding windows

6.2 Mid Level Features

6.3 model

6.4 Results

7 Conclusions

7.1 Future Researches

References

[1] The Language Archive. URL: https://archive.mpi.nl/tla/elan.

[2] Ceccaldi E. "CEST: a Cognitive Event based Semi-automatic Technique for behavior segmentation". PhD thesis. Università di Genova, 2022.

[3] R. Niewiadomski et al. "Automated Detection of Impulsive Movements in HCI". In: *Conference CHItaly* (2015).

[4] R. Niewiadomski et al. "Does embodied training improve the recognition of mid-level expressive movement qualities sonification?" In: *Multimodal User Interfaces* (2018).

[5] Casa Paganini. URL: http://www.casapaganini.org/.

[6] Qualisys. URL: https://www.qualisys.com/.

[7] L. Wang et al. "Review of Classification Methods on Unbalanced Data Sets". In: *IEEE Access* (2021).

[8] J.M. Zacks and K.M. Swallow. "Event segmentation". In: *Current directions in psychological science* 16.2 (2007), pp. 80–84.