A Multi-layered Conceptual Framework for Expressive Gesture Applications

Antonio Camurri (1)
Giovanni De Poli (2)
Marc Leman (3)
Gualtiero Volpe (1)

(1) DIST - University of Genova, Italy
(2) CSC-DEI – University of Padova, Italy
(3) IPEM – Ghent University, Belgium

Abstract

The paper aims at (i) understanding expressiveness in gestures using computational modeling and (ii) exploit this understanding in artistic applications, where the enhancement of the expressiveness in interactive music/dance/video systems is a major goal. A multi-layered conceptual framework is presented and examples are given of its use in interactive art performances.

1 Introduction

In this paper, we focus on the problem of affective/emotional communication from the perspective of interactive human-machine in art performances. Art is a field where the recognition of expressiveness in gestures is of central importance. The human sciences, in particular aesthetics, have a long tradition in describing basic concepts of artistic expressiveness yet the concepts remain largely ill-defined and badly understood.

Our goal is (i) to better understand expressiveness in gestures using computational modeling and (ii) exploit this understanding in artistic applications, where the enhancement of the expressiveness in interactive music/dance/video systems is a major goal. Enhancing expressive communication in novel art media is not only useful for musicians, choreographers, actors but for all users who develop multimedia content and applications for interactive applications with different degrees of affective and emotional participation.

The paper first defines the notion of expressiveness in gestures as related to artistic human-machine interaction. A multi-layer conceptual framework is presented and examples are given of the use of the system architecture in art applications. The research described in this paper is developed in the framework of the EU IST Project MEGA (Multisensory Expressive Gesture Applications, www.megaproject.org).

2. Expressiveness in Gestures

Traditionally, a gesture is defined as a body motion that conveys information [1]. Many gestures in artistic contexts are called expressive and are meant, not to denote things in the outer world but, to convey information related to the affective/emotional domain. Humanistic theories of expressiveness often refer to the role of an affective/emotional semantics which would be associated with gestures [2]. It seems likely that expressiveness in gestures is conveyed by a set of temporal/spatial characteristics that operate more or less independent from the denotative meanings (if any) of those gestures. In that sense, gestures can be conceived as the vehicles which carry these expressive characteristics and it is likely that expressiveness as such subsumes certain universal patterns and general rules. Our research is focused on understanding and exploiting these patterns and rules through computer modeling.

3. A Layered Conceptual Framework

Gesture recognition can be modeled in terms of cue shapes, sequences of features, parameterized physical models. Yet the modeling of expressiveness in gestures, requires proper techniques that capture the subtle temporal/spatial characteristics of expressiveness. What, then, are the features that define expressiveness?

We assume that they are related to low-level characteristics of the dynamics of the movement, as well as to medium-level features that relate to semantic spaces or maps (e.g., energy-velocity spaces and valence-arousal emotion spaces), as well as to higher-level conceptualizations and taxonomies (synaesthetic and kinaesthetic metaphors). These aspects are taken into account in a layered conceptual framework.

The layered conceptual framework has a role as a supporting framework for the development and implementation of expressiveness in gestures. The multi-layered approach is used to split up the problem of expressive gesture analysis and mapping into different sub-problems. A distinction is made between different levels, from low-level features, to events and patterns, to semantic spaces, and taxonomies. This allows a bottom-up and a top-down approach as well as a flexible interpretation of the concepts in relationship to emotional, affective and sensitive processing.

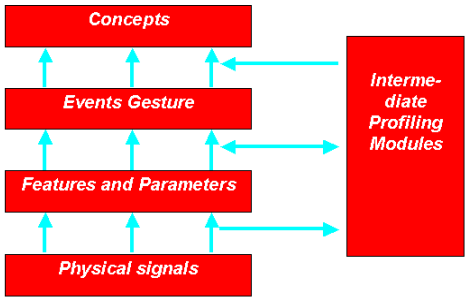


Figure 1: A layered conceptual framework.

The conceptual framework consists of four layers:

Layer 1 – Physical Signals:

This is the information which is captured by the sensors of the computer system. Physical signals may have different formats. They may consist of time variant signals such as sampled audio signals, sampled signals from tactile, infra-red sensors, signals from haptic devices, or events such as MIDI-messages or low-level data frames in video.

Layer 2 – Low-level features and statistical parameters:

Low-level features are extracted and processed (in the statistical sense) in order to carry out a subsequent analysis related to expression. For example, in the audio domain, these low-level features are related to tempo (= number of beats per minute), tempo variability, sound level (measured in dB), sound level variability, spectral shape (which is related to the timbre characteristics of the sound), articulation (features such as legato, staccato), articulation variability, attack velocity (which is related to the onset characteristics which can be fast or slow), pitch, pitch density, degree of accent on structural important notes, periodicity (related to repetition in the energy of the signal), dynamics (intensity), roughness (or sensory dissonance), tonal tension (or the correlation between local pitch patterns and global or contextual pitch patterns), and so on.

Layer 3 – Mid-level features and maps:

In this layer, the purpose is to represent expression in gestures by modeling the low-level features in such a way that they give an account of expressiveness in terms of events, shapes, patterns or as trajectories in spaces or maps. Starting from parameters relevant for detecting expressive content features, particular models can be used.

Most of the research done thus far has been focused on so-called semantic spaces or maps. A semantic map represents categories of semantic features related to emotion and expression on a pre-defined grid. Typically, a gesture is then a trajectory in this space, and each trajectory can be seen as a point in a trajectory-related (super)space.

Energy-velocity spaces have been successfully used for the synthesis of the musical performance. The space is derived from perceptual experiments [3,4] and has thus far used in synthesis of different and varying expressive intentions in a musical performance. The energy-velocity space is correlated with legato-staccato properties versus tempo. Positions in this space are used to define MIDI parameters, as well as audio signal parameters, that control the timing and the dynamics of the notes to be played during a performance. The MIDI parameters typically control tempo and key velocity. The audio-parameters control tempo, legato, loudness, brightness, attack time, vibrato, envelope shape….

Friberg and Bresin [5] start from a semantic map that represents basic emotions. They used the results of Juslin [6] for the synthesis of performances conveying different basic emotions Similar techniques (in reversed direction) can be applied for the analysis of expressiveness in gestures. In [7] an approach based on the metaphor of potential fields has been proposed to analyse movement and map on music objects. [8] gives examples of analysis of

expressive gesture in movement by expressive spaces inspired to Laban's Effort space.

Layer 4 – Concepts and structures:

This layer contains a network that correlates concepts to low-level and mid-level features. The network specifies relationships between these features and concepts. Cross-modality relationships are expressed at this level. The network is typically based on fuzzy logics or probabilistic reasoning systems, such as Bayesian networks. Such a structure may describe the four basic emotions (fear, grief, anger, happiness), or it may map gestures onto the Laban conceptual framework of gesture effort, as in Fig. 2.

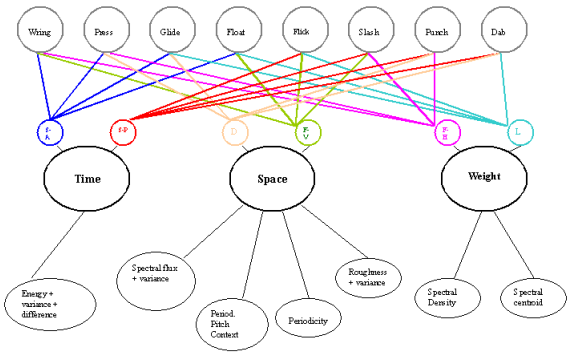


Figure 2: Example of a semantic network which defines concepts and relationships of the Laban effort theory in terms of Layer II features (energy, spectral flux, periodicity, roughness, …)

The difference between the conceptual network and the semantic spaces of the mid-level layer is that in the conceptual network, the concepts function as objects in a discrete space, whereas the semantic spaces of the mid-level layer can be continuous. The main categories of those spaces would be called objects in the conceptual network. The advantage of the conceptual networks is that abstract concepts can be introduced and that relationships between different features and feature maps can be defined.

The schema of Fig. 1 contains an added block devoted to aspects of personalization. The idea is that in certain applications the conceptual architecture can be initialized on personal grounds, such that certain aspects of expressiveness may constrain the processing. The machine may thus be initialized as having an aggressive or soft character using the Intermediate Profiling Modules. The mechanism makes the "archetypical" parameters of the architecture more flexible and personalized, by keeping track of (i) their evolution over time given specific contexts, and (ii) different biasing due to "personality" and focus of interest, etc. This feature has been conceived, but is yet to be worked out in more detail.

5. Examples

In this paragraph a number of examples are given which illustrate the effectiveness of the proposed concepts and approach to the study of expressiveness in art.

5.1. Movement Shapes: Toward a Symbolic Description of Expressiveness in Full-Body Gestures

This example relates to Layer II (low-level features and statistical parameters) and aims at segmenting full-body movements on the basis of expressiveness in movement patterns. The segmentation process is based on stops or pauses. It gives as output a sequence of gestures and each of them can be described by a set of curves describing motion parameters such as contraction and expansion, fluentness, direct and flexible motion [9]. Direct comparisons between the shapes of such curves in the same dance performed with different expressive intentions suggest that different expressive intentions produce difference shapes in such motion curves.

A main focus of the research is on the identification and classification of different shapes and investigation of their relationship with a dancer's expressive intention. Segmentation allows us to describe a dance fragment as a sequence of motion and pause phases each one characterized by measured values of suitable parameters. This can be considered as a first step toward a symbolical representation of the movement: consider, for example, a temporal analysis of the contraction/expansion index. The sequence of pause phases, motion phases, and legato phases is considered. We call a "legato" phase a particular phase of the movement in which a transition between two movement phases is detected without an explicit pause phase in between. From the point of view of the temporal analysis, movement, pause, and legato phases can be characterized by information such as start frame, end frame, duration. Movement phases are distinguished between expansion phases and contraction phases and are characterized by further information such as the maximum value of the index during the movement phase and the offset of such maximum with respect to the start frame of the movement phase.

For example, let us consider the following excerpt from an automated analysis of movement in EyesWeb:

```
expansion(158 , 16 , 9.0 , 146.0 , 522.0 , 170
, -0.235009 ).
legato( 174 , 2 ).
contraction( 176 , 8 , 193.0 , 149.0 , 225.0 ,
178 , 0.417968 ).
pause( 184 , 13 , 9 ).
```

Where each line represents a bell-shaped curve approximating a segment of the gesture:
- Expansion and Contraction have the following main parameters: start frame of the segment, length in frames/samples, value of the first sample, final value, max value in this bell-like shape, offset);
- pause: a segment with no movement. Parameters: Start frame, Length, Zeroes (no. of samples effectively at zero in the Length interval);
- legato: a transition between two segments in which a partial overlap causes no real zero values between the segments. Parameters: Start Frame, Length.

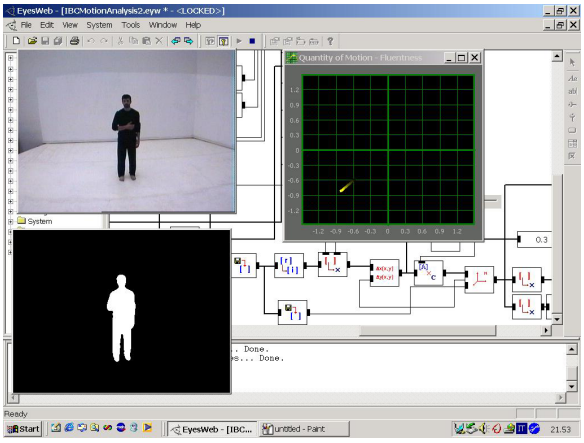
This example shows a kind of representation which might open novel perspectives for gesture analysis: for example, a rule-based system could be applied in order to make inferences depending on sequences of contractions, pauses, and expansions and the measures related to them. Statistical sequential models such as HMM could be applied as well to detect movement high-level features.

5.2. The QuantityOfMotion/Fluentness Expressive Space

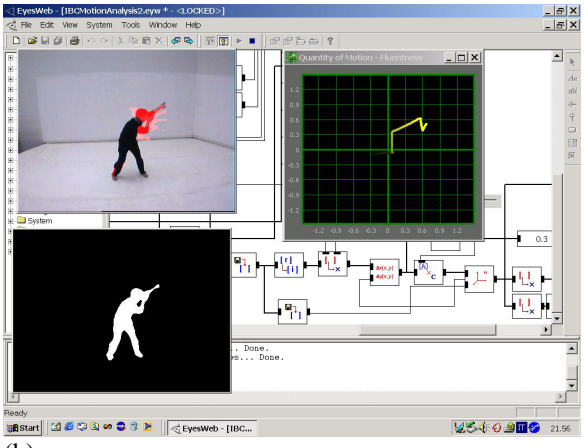
In this section, we present an example of the extraction of expressive cues from a Layer I and II representation and its mapping onto a Layer III representation, in particular a 2D expressive space. The example, implemented as an EyesWeb patch, was demonstrated recently as a part of a public demonstration in the ISTV (EU-IST Project MEGA, www.megaproject.org) at IBC2001 (Amsterdam (NL), September 14th - 18th, 2001).

A first set of EyesWeb software modules is devoted to the extraction of features from the low-level representation. The dancer's silhouette is extracted using background subtraction and a value proportional to the actual quantity of motion is calculated in the following way (roughly): the difference between the silhouette in the current and previous frame is computed, then the last k differences obtained in this way are summed, thus obtaining an image as a sum of the last k differences between couples of frames. The resulting area of such image is an approximation of the quantity of motion. An adaptive threshold, depending on the running average of the quantity of motion in a window of 1 second is then applied in order to distinguish motion and pause phases. Duration of pause and motion phases is measured and the ratio between such durations in a window of 1second is taken as one of the contributors to the measurement of the fluentness of the movement. Actually fluent movements are also continuous with few and short breaks, while more rigid, strong movements are emphasized by more frequent and, eventually, long pauses. Both quantity of motion and fluentness features are situated at Layer II.

A second set of EyesWeb software modules then maps the two extracted features onto an expressive space along two dimensions: quantity of motion and fluentness. Other expressive intentions and cues may be (partially) identified as regions in this space: e.g., rigidness, hardness, heaviness, softness. Such expressive intentions can be mapped on output channels. For example, a fluent movement may influence in real time the interpretation of a musical score by producing a legato performance, by means of suitable deviations of energy/timing cues.



(a)



(b)

Figure 5a and b: EyesWeb patch for the expressive space example.

Figures 5a and 5b show the running EyesWeb patch application for this experiment. In the first image, the

dancer is not moving: the current position in the expressive space (window in the right) is moving toward the bottom left parts of the 2D space (yellow stripe), a position characterized by low quantity of motion and low fluentness (e.g., the amount of pause phases is dominating the amount of motion phases). In figure 5b, a high-energy gesture is displayed. The red shadow around the dancer (upper-left window) is proportional to the quantity of motion and the position in the expressive space (yellow stripe in the right window) is moving toward the top-right region in that window, characterized by high quantity of motion and high fluentness.

5.3. Real-time Synthesis of Expressiveness in Sound

Mapping the dynamics of extracted expressive features onto trajectories in abstract expressive spaces (as described in Section 5.2) allows real-time synthesis of expressive outputs depending on the evolution of such trajectories. For example, a dancer may perform a choreography dancing on a real-time computer generated expressive performance of a musical score. Computer is able to control in real time, like a real pianist, the expressive character of his performance. Analysis of expressive cues in movement (e.g., fluent/rigid, etc) can be performed and the dancer's expressive intentions are mapped in a predefined abstract space. The dancer's expressive intentions influence the automatic music performance in a coherently way, i.e., if the dancer is moving heavily, music also will become heavy.

In this perspective, an automatic expressive music performance can be generated, according to trajectories in an abstract space. From this point of view the mid-level map represents a control space, which determines, at an abstract level, the expressive content and the interaction between the dancer and the musical message. The control space is characterized by a set of adjectives describing different expressive intentions of the performer. In general, this space lets artists to organize their own abstract space by defining expressive points and positioning them in the space. A label is associated to each point describing the meaning of the gesture, as e.g. heavy, light, rigid, fluid.

An EyesWeb module (called ISpace), part of the MEGASE, was designed for computing low-level features (Layer II) starting from a position (x-y coordinates) in the abstract space. The module receives as input a trajectory in the abstract space, obtained from the dancer's gesture analysis. A suitable mapping strategy is employed in order to vary coherently and gradually the expressiveness (i.e., morphing among happy, solemn and dark). Morphing can be realized with a wide range of graduality (from abrupt to very smooth), allowing to adapt the system to different situations.

Analysis-by-synthesis method was applied to estimate which kind of morphing technique ensures the best perceptual result. The computer-generated performances showed appropriate expressive meaning in all the points of control space, computing intermediate points of the space using a quadratic interpolation. It has to be noticed that expressive content of a performance is revealed on a time scale which is longer of that of a single event; therefore, in order to obtain a fruition which is coherent with the artist's intentions, the system allows to slow down the movements of the user, so to avoid unwanted "expressive discontinuities" in correspondence of abrupt movements. To this end, suitable smoothing strategies have been developed for movement data coming form the dancer.

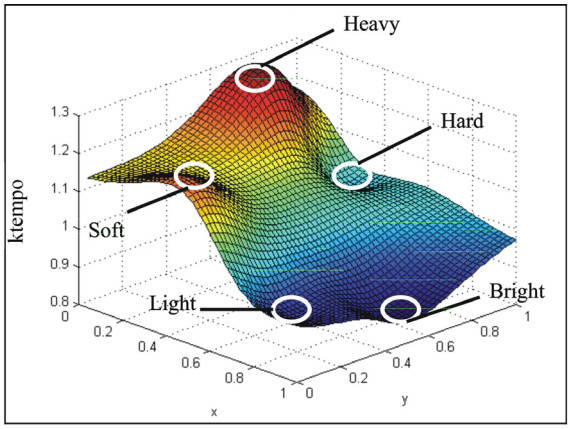


Figure 6: Mapping surface among points of the control space (x-y plane) and the parameter.

On the basis of music performance analysis [4] some low-level features have turned out to be particularly important for expressive music rendering, for instance Tempo, Legato, Intensity, Phrasing, etc. Their mean value and range variation (k and m) are the second layer features computed by this module. Another module (called ExpressiveSeq) makes use of these parameters in order to calculate the Layer I features, i.e. the deviations which have to be applied to the score for the rendering of the desired expressive intention and the corresponding MIDI messages. Figure 6 shows an example of the mapping surface computed by the module ISpace relating points of a control space with the parameter Ktempo, i.e., the k for the Tempo low-level feature. On the basis of movements on the x-y plane, the variations of the parameter Ktempo to be applied to the performance are thus computed. A value greater than 1 stands for rallentando (gradually slackening in tempo), while a value lower than 1 stands for

accelerando (gradually accelerating in tempo). The model computes intermediate points of the space using a quadratic interpolation.

7. Conclusion

This paper reflects the main findings of a first year project that has a focus on the expressiveness of gestures. The *MEGA System Environment (MEGASE)* is the main scientific/technological result and concrete product of the EU-IST MEGA project. It is build onto the EyesWeb open software platform (www.infomus.dist.unige.it/eywindex.html) and it is composed by hardware components and software libraries integrated or connected to EyesWeb.

Our explorations have led to the development of a layered conceptual framework for the description of expressiveness in terms of low-level, mid-level and high-level representations. At this stage, modules for low-level features of expressiveness have been implemented related to multiple sensory modalities. The extracted features have been mapped to higher level concepts such as basic emotions and energy-velocity maps.

8 Acknowledgments

This work has been partially funded by EU Project IST 20410 MEGA (Multisensory Expressive Gesture Applications) .

References

[1] Wanderley, M. and M. Battier, 2000. *Trends in Gestural Control of Music.* (Edition électronique.) Paris: IRCAM.

[2] Coker, W., 1972. Music and Meaning – A *Theoretical Introduction to Musical Aesthetics*. New York: The Free Press.

[3] Canazza, S., De Poli, G., Rinaldin, S. and A. Vidolin, 1997. Sonological Analysis of Clarinet Expressivity. In M. Leman (Ed.) *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, Berlin, Heidelberg: Springer-Verlag.

[4] Canazza, S., De Poli, G. and A. Vidolin, 1997. Perceptual Analysis of the Musical Expressive intention in a Clarinet Performance. In M. Leman (Ed.) *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, Berlin, Heidelberg: Springer-Verlag.

[5] Bresin, R. and Friberg, A., 2000. Emotional Coloring of Computer-Controlled Music Performances. *Computer Music Journal*, 24:4, 44-63.

[6] Juslin, P., 1997. Perceived Emotional Expression in Synthesized Performances of a Short Melody: Capturing the Listener's Judgment Policy. *Musicae Scientiae*, 1:2, 225-256.

[7] Camurri, A., Frixione, M., and C. Innocenti, 1994. A Cognitive Model and a Knowledge Representation Architecture for Music and Multimedia. *Interface - Journal of New Music Research*, 23:4, 317-347.

[8] Camurri A., Hashimoto S., Ricchetti M., Trocca R., Suzuki K., and G. Volpe, 2000. EyesWeb – Toward Gesture and Affect Recognition in Dance/Music Interactive Systems, *Computer Music Journal*, 24, 57-69.

[9] Camurri, A., Mazzarino, B., Trocca, R. and G. Volpe, 2001. Real-Time Analysis of Expressive Cues in Human Movement, *Proc. Intl. Conf CAST01*, Bonn: GMD.