Multimodal Analysis of Expressive Gesture in Music and Dance Performances

Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti,
Renee Timmers, and Gualtiero Volpe

InfoMus Lab, DIST - University of Genova
Viale Causa 13, I-16145, Genova, Italy
{toni,bunny,rtimmers,rmat,volpe}@infomus.dist.unige.it
http://infomus.dist.unige.it

Abstract. This paper presents ongoing research on the modelling of expressive gesture in multimodal interaction and on the development of multimodal interactive systems explicitly taking into account the role of non-verbal expressive gesture in the communication process. In this perspective, a particular focus is on dance and music as first-class conveyors of expressive and emotional content. Research outputs include (i) computational models of expressive gesture, (ii) validation by means of continuous ratings on spectators exposed to real artistic stimuli, and (iii) novel hardware and software components for the EyesWeb open platform (www.eyesweb.org), such as the recently developed Expressive Gesture Processing Library. The paper starts with a definition of expressive gesture. A unifying framework for the analysis of expressive gesture is then proposed. Finally, two experiments on expressive gesture in dance and music are discussed. This research work has been supported by the EU IST project MEGA (Multisensory Expressive Gesture Applications, www.megaproject.org) and the EU MOSART TMR Network.

1 Introduction

Expressive gesture is a key concept in our research (Camurri et al., 2001). This paper tries to face it and introduces two experiments aiming at understanding the non-verbal mechanisms of expressive/emotional communication.

Several definitions of gesture exist in the literature. The most common use of the term is with respect to natural gesture, which is defined as a support to verbal communication. For Cassel and colleagues (1990) “A natural gesture means the types of gestures spontaneously generated by a person telling a story, speaking in public, or holding a conversation”. McNeill (1992) in his well-known taxonomy divides the natural gestures generated during a discourse in four different categories: iconic, metaphoric, deictic, and beats. In a wider perspective Kurtenbach and Hulteen (1990) define gesture as “a movement of the body that contains information”. A survey and a discussion of existing definition of gesture can be found in (Cadoz and Wanderley, 2000).

In artistic contexts and in particular in the field of performing arts, gesture is often not intended to denote things or to support speech as in the traditional framework of natural gesture, but the information it contains and conveys is related to the affec-

A. Camurri and G. Volpe (Eds.): GW 2003, LNAI 2915, pp. 20–39, 2004.

tive/emotional domain. From this point of view, gesture can be considered "expressive" depending on the kind of information it conveys: expressive gesture carries what Cowie et al. (2001) call "implicit messages", and what Hashimoto (1997) calls KANSEI. That is, expressive gesture is the responsible of the communication of information that we call *expressive content*. Expressive content is different and in most cases independent from, even if often superimposed to, possible denotative meaning. Expressive content concerns aspects related to feelings, moods, affect, intensity of emotional experience.

For example, the same action can be performed in several ways, by stressing different qualities of movement: it is possible to recognize a person from the way he/she walks, but it is also possible to get information about the emotional state of a person by looking at his/her gait, e.g., if he/she is angry, sad, happy. In the case of gait analysis, we can therefore distinguish among several objectives and layers of analysis: a first one aiming at describing the physical features of the movement, for example in order to classify it (quite a lot of research work can be found in the computer vision literature about gait analysis, see for example Liu et al., 2002); a second one aiming at extracting the expressive content gait coveys, e.g., in terms of information about the emotional state that the walker communicates through his/her way of walking. From this point of view, walking can be considered as an expressive gesture: even if no denotative meaning is associated with it, it still communicates information about the emotional state of the walker, i.e., it conveys a specific expressive content. In fact, in this perspective the walking action fully satisfies the conditions stated in the definition of gesture by Kurtenbach and Hulteen (1990): walking is "a movement of the body that contains information". Some studies can be found aiming at analysing the expressive intentions conveyed through everyday actions: for example, Pollick (2001) investigated the expressive content of actions like knocking or drinking

If on the one hand expressive gestures partially include natural gestures, that is, natural gestures can also be expressive gestures, we face on the other hand a more general concept of expressive gesture that includes not only natural gestures but also musical, human movement, visual (e.g., computer animated) gestures. Our concept of expressive gesture is therefore somewhat broader than the concept of gesture as defined by Kurtenbach and Hulteen, since it considers also cases in which, with the aid of technology, communication of expressive content takes place even without an explicit movement of the body, or, at least, the movement of the body is only indirectly involved in the communication process (e.g. the allusion at movement in musical signals). This can happen, for example, also when using visual media. The expressive content is conveyed through a continuum of possible ways ranging from realistic to abstract images and effects: cinematography, cartoons, virtual environments with computer animated characters and avatars, expressive control of lighting and colour in a theatre context (e.g., related to actor's physical gestures). Consider, for example, a theatre performance: the director, the choreographer, or the composer can ask actors, dancers, musicians, to communicate content through specific expressive gestures (e.g., dance and/or music interpretation). At the same time, technology enables the director to extend artistic language: he can map motion or music features onto particular configurations of lights, on movement of virtual characters, on automatically generated computer music and live electronics. In this way, he can create an "extended" expressive gesture that - while still having the purpose of communicating an expressive content - it is only partially related to explicit body movement: in a way, such "extended expressive gesture" is the result of a juxtaposition of several dance, music, and visual gestures, but it is not just the sum of them, since it also includes the artistic point of view of the artist who created it, and it is perceived as a whole multimodal stimulus by human spectators.

Our research on expressive gesture is finalized to the development of interactive multimedia systems based on novel interaction paradigms enabling a deeper experience and participation of the user by explicitly observing and processing his/her (multimodal) expressive gesture. Since artistic performance uses non-verbal communication mechanisms to convey expressive content, we focused on performing arts, and in particular on dance and music, as the main test-beds where computational models of expressive gesture and algorithms for expressive gesture processing can be developed, studied, and tested.

In particular, our attention has been focused on two aspects:

- Expressive gesture as a way to convey a particular emotion to the audience;
- Expressive gesture as a way to induce intense emotional experience in the audience (see e.g. Scherer, 2003).

Each of them has been recently subject of experiments at our Lab aiming at understanding which features in an expressive gesture are responsible for the communication of the expressive content, and how the dynamics of these features correlates with a specific expressive content.

In this paper, we concretely illustrate our approach by presenting two experiments focused on these two aspects.

The first one aims at (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions (in term of basic emotions) to the audience during a dance performance and (ii) testing the developed models and algorithms by comparing their performances with spectators' ratings of the same dance fragments.

The second one investigates the mechanisms responsible for the audience's engagement in a musical performance. The aim of this experiment is again twofold: (i) individuating which auditory and visual cues are mostly involved in conveying the performer's expressive intentions and (ii) testing the developed model by comparing its performance with spectators' ratings of the same musical performances.

For the analysis of expressive gesture in these experiments we developed a unifying conceptual framework, described it in the next section.

2 A Unifying Layered Conceptual Framework

The experiments presented in this paper address expressive gesture in music and dance performance.

While gesture in dance performance mainly concerns the visual/physical modality (even if the auditory components can be relevant as well), gesture in music performance uses both the auditory and the visual channels to convey expressive information, and, thus, it is multimodal in its essence. Gesture in music performance is not only the expressive and functional physical gesture of a performer, but it also includes expressive gesture present in the produced sound. When we define gesture in terms of structural units that have internal consistency and are distinguished in time and quality from neighbouring units, it is possible to analyse gesture in both (acoustic and visual) modalities. Multimodality is therefore a key issue. In order to deal with multimodal input a unifying conceptual framework has been adopted, derived from (Camurri, De Poli, Leman, 2001). It is based on a layered approach ranging from low-level physical measures (e.g., position, speed, acceleration of body parts for dance gestures; sampled audio signals or MIDI messages for music gesture) toward descriptors of overall gesture features (e.g., motion fluency, directness, impulsiveness for dance gestures; analysis of melodic, harmonic, agogic qualities of a music phrase for music gesture).

This layered approach is sketched in Figure 1. Each layer is depicted with its inputs, its outputs, and the kind of processing it is responsible for. In the following sections, each layer will be discussed with respect to its role in the two experiments.

Our conceptual framework, here presented for analysis, can also be applied for synthesis of expressive gesture: for example for the generation and control of the movement of avatars, virtual characters, or robots in Mixed Reality scenarios, as well as for the synthesis and interpretation of music. Examples of synthesis of expressive movement and expressive audio content are well documented in literature: see e.g. the EMOTE system (Chi et al., 2000) for generation of movement of avatars and virtual characters based on high level motion qualities, and the systems for synthesis of expressive music performances developed at KTH (Friberg et al, 2000) and by the DEI-CSC group at the University of Padova (Canazza et al., 2000).

Finally, it should be noticed that in the perspective of developing novel interactive multimedia systems for artistic applications, such a framework should be considered in the context of a broader scenario (Mixed reality, Distributed Collaborative Virtual Environments) in which virtual subjects (e.g., virtual characters) who behave both as observers and as agents perform the four layers of processing in the analysis of observed expressive gestures and in the synthesis of expressive gestures to communicate (directly or remotely) with other real and virtual subjects.

3 Analysis of Expressive Gesture in Dance Performance

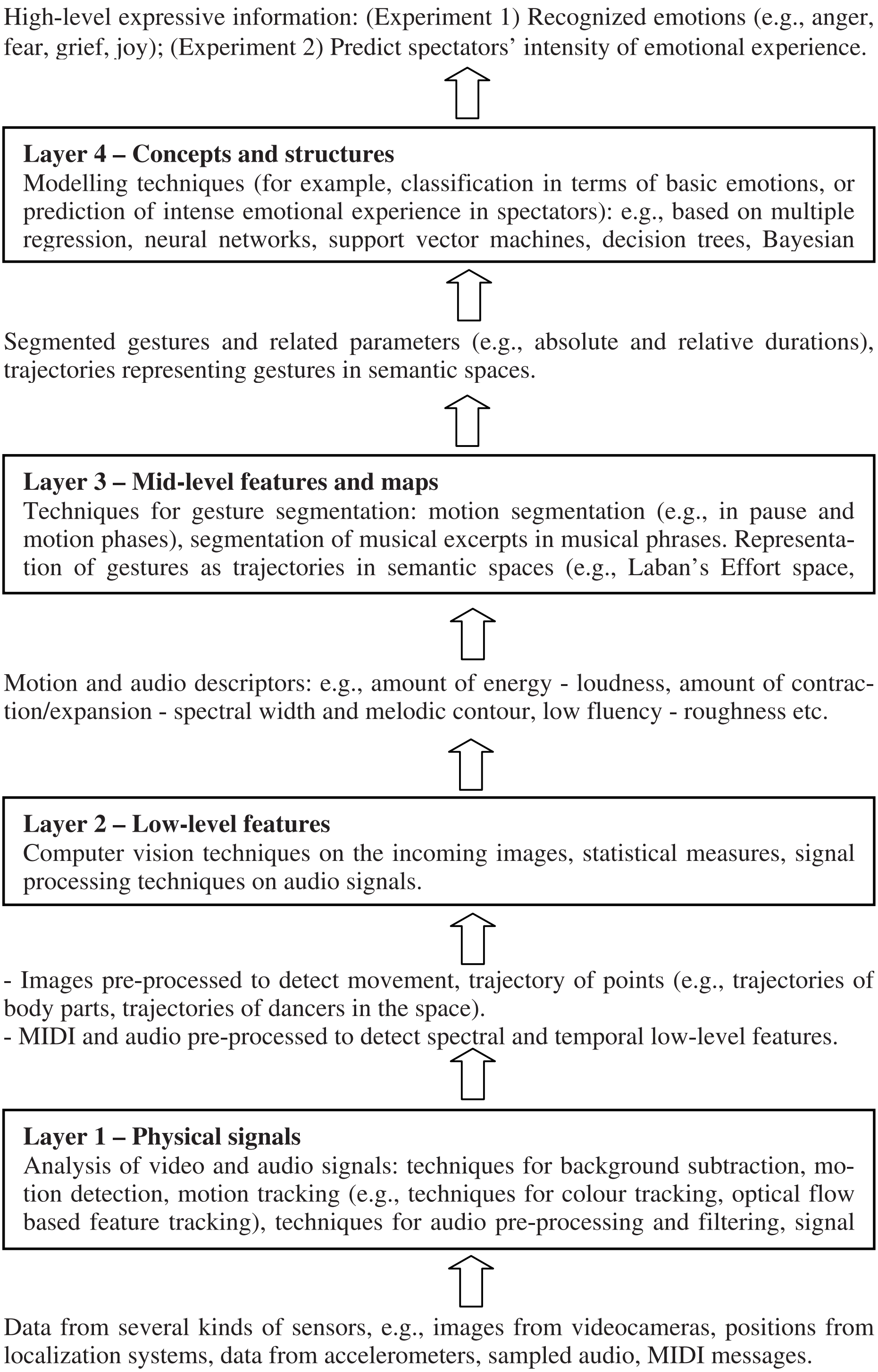
As an example of analysis of expressive gesture in dance performance, we discuss an experiment carried out in collaboration with the Department of Psychology of the University of Uppsala (Sweden) in the EU-IST MEGA project.

The aim of the experiment was twofold: (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions to the audience during a dance performance and (ii) testing the developed models and algorithms by comparing their performances with spectators' ratings of the same dance fragments.

In the case of this experiment, expressive gesture was analysed with respect to its ability to convey emotions to the audience. The study focused on the communication through dance gesture and recognition by spectators of the four basic emotions: anger, fear, grief, and joy.

The research hypotheses are grounded on the role of the Laban's dimensions in dance gesture, as described in Laban's Theory of Effort (Laban, 1947, 1963):

- The time dimension in terms of overall duration of time and tempo changes also elaborated as the underlying structure of rhythm and flow of the movement;
- The space dimension in its aspects related to Laban's "personal space" e.g., to what extent limbs are contracted or expanded in relation to the body centre;



High-level expressive information: (Experiment 1) Recognized emotions (e.g., anger, fear, grief, joy); (Experiment 2) Predict spectators' intensity of emotional experience.

Layer 4 – Concepts and structures
Modelling techniques (for example, classification in terms of basic emotions, or prediction of intense emotional experience in spectators): e.g., based on multiple regression, neural networks, support vector machines, decision trees, Bayesian

Segmented gestures and related parameters (e.g., absolute and relative durations), trajectories representing gestures in semantic spaces.

Layer 3 – Mid-level features and maps
Techniques for gesture segmentation: motion segmentation (e.g., in pause and motion phases), segmentation of musical excerpts in musical phrases. Representation of gestures as trajectories in semantic spaces (e.g., Laban's Effort space,

Motion and audio descriptors: e.g., amount of energy - loudness, amount of contraction/expansion - spectral width and melodic contour, low fluency - roughness etc.

Layer 2 – Low-level features
Computer vision techniques on the incoming images, statistical measures, signal processing techniques on audio signals.

- Images pre-processed to detect movement, trajectory of points (e.g., trajectories of body parts, trajectories of dancers in the space).
- MIDI and audio pre-processed to detect spectral and temporal low-level features.

Layer 1 – Physical signals
Analysis of video and audio signals: techniques for background subtraction, motion detection, motion tracking (e.g., techniques for colour tracking, optical flow based feature tracking), techniques for audio pre-processing and filtering, signal

Data from several kinds of sensors, e.g., images from videocameras, positions from localization systems, data from accelerometers, sampled audio, MIDI messages.

Fig. 1. The layered conceptual framework.

- The flow dimension in terms of analysis of shapes of speed and energy curves, and frequency/rhythm of motion and pause phases.
- The weight dimension in terms of amount of tension and dynamics in movement, e.g., vertical component of acceleration.

These cues were predicted to be associated in different combinations to each emotion category. Details can be found in (Camurri, Lagerlof, Volpe, 2003).

3.1 Description of the Experiment

An experienced choreographer was asked to design a choreography such that it excluded any propositional gesture or posture and it avoided stereotyped emotions.

In Uppsala, five dancers performed this same dance with four different emotional expressions: anger, fear, grief and joy. Each dancer performed all four emotions. The dance performances were video-recorded by two digital videocameras (DV recording format) standing fixed in the same frontal view of the dance (a spectator view). One camera obtained recordings to be used as stimuli for spectators' ratings. The second video camera was placed in the same position but with specific recording conditions and hardware settings to simplify and optimise automated recognition of movement cues (e.g., high speed shutter). Dancers' clothes were similar (dark), contrasting with the white background, in an empty performance space without any scenery. Digitised fading eliminated facial information and the dancers appeared as dark and distant figures against a white background.

The psychologists in Uppsala then proceeded in collecting spectators' ratings: the dances were judged with regard to perceived emotion by 32 observers, divided in two groups. In one group ratings were collected by 'forced choice' (chose one emotion category and rate its intensity) for each performance, while the other group was instructed to use a multiple choice schemata, i.e., to rate the intensity of each emotion on all four emotion scales for each performance.

At the same time, at the InfoMus Lab we proceeded in extracting motion cues from the video recordings and in developing models for automatic classification of dance gestures in term of the conveyed basic emotion.

3.2 Automated Extraction of Motion Cues

Extraction of motion cues followed the conceptual framework described in Section 2.

3.2.1 Layer 1

In the case of analysis of dance performance from video, layer 1 is responsible for the processing of the incoming video frames in order to detect and obtain information about the motion that is actually occurring. It receives as input images from one or more videocameras and, if available, information from other sensors (e.g., accelerometers). Two types of output are generated: processed images and trajectories of body parts. Layer 1 accomplishes its task by means of consolidated computer vision techniques usually employed for real-time analysis and recognition of human motion and activity: see for example the temporal templates technique for representation and recognition of human movement described in Bobick and J. Davis (2001). It should be noticed that in contrast to Bobick and J. Davis research, we do not aim at detecting or recognizing a specific kind of motion or activity. The techniques we use include feature tracking based on the Lucas-Kanade algorithm (Lucas and Kanade, 1981), skin color tracking to extract positions and trajectories of hands and head, an algorithm to divide a body silhouette in sub-regions, and Silhouette Motion Images (SMIs). A SMI is an image carrying information about variations of the silhouette shape and position in the last few frames. SMIs are inspired to motion-energy images (MEI) and motion-history images (MHI) (Bradsky and J. Davis, 2002, Bobick and J. Davis, 2001). They differ from MEIs in the fact that the silhouette in the last (more recent) frame is removed from the output image: in such a way only motion is considered while the current posture is skipped. Thus, SMIs can be considered as carrying information about the "amount of motion" occurred in the last frames. Information about time is implicit in SMI and is not explicitly recorded. We also use an extension of SMIs, which takes into account the internal motion in silhouettes. In such a way we are able to distinguish between global movements of the whole body in the General Space and internal movements of body limbs inside the Kinesphere.

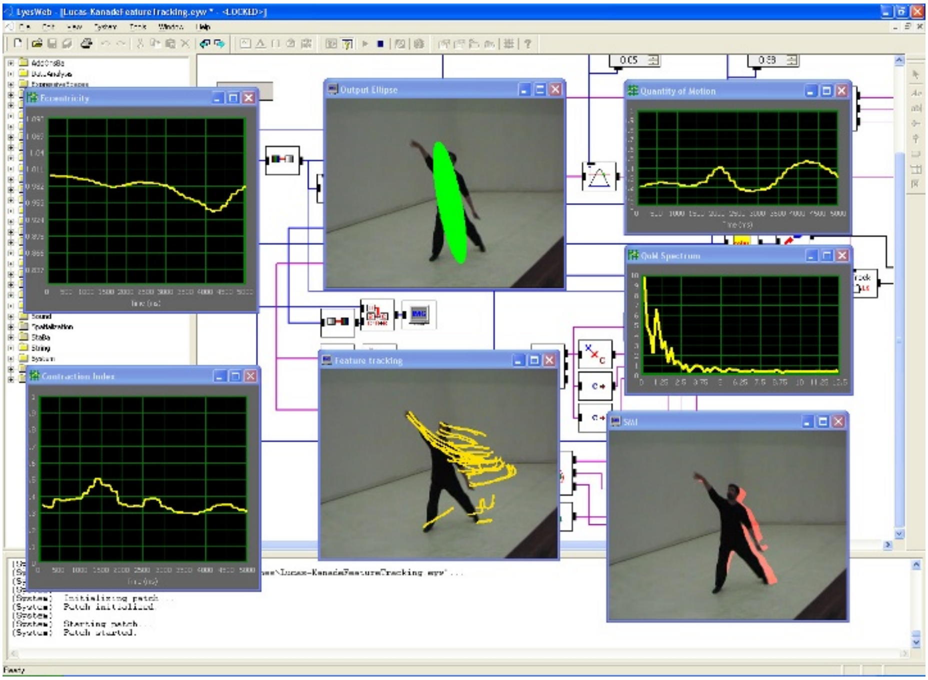
3.2.2 Layer 2

Layer 2 is responsible of the extraction for a set of motion cues from the data coming from low-level motion tracking. Its inputs are the processed images and the trajectories of points (motion trajectories) coming from Layer 1. Its output is a collection of motion cues describing movement and its qualities. According to the research hypotheses described above, the cues extracted for this experiment include:

- Cues related to the amount of movement (energy) and in particular what we call Quantity of Motion (QoM). QoM is computed as the area (i.e., number of pixels) of a SMI (Camurri, Lagerlof, Volpe, 2003). It can be considered as an overall measure of the amount of detected motion, involving velocity and force.
- Cues related to body contraction/expansion and in particular the Contraction Index (CI). CI is a measure, ranging from 0 to 1, of how the dancer's body uses the space surrounding it. The algorithm to compute the CI (Camurri, Lagerlof, Volpe, 2003) combines two different techniques: the individuation of an ellipse approximating the body silhouette and computations based on the bounding region. The former is based on an analogy between the image moments and mechanical moments (Kilian, 2001): the eccentricity of the approximating ellipse is related to body contraction/expansion. The latter compares the area covered by the minimum rectangle surrounding the dancer with the area currently covered by the silhouette.
- Cues derived from psychological studies (e.g., Boone and Cunningham, 1998) such as amount of upward movement, dynamics of the Contraction Index (i.e., how much CI was over a given threshold along a time unit);
- Cues related to the use of space: length and overall direction of motion trajectories;
- Kinematical cues (e.g., velocity and acceleration) calculated on motion trajectories.

For those cues depending on motion trajectories a Lucas-Kanade feature tracker has been employed in Layer 1. A redundant set of 40 points randomly distributed on the whole body has been tracked. Points have been reassigned each time dancers stopped their motion (i.e., a pause was detected) so that a small and not significant amount of points is lost during tracking. Overall motion cues have been calculated by averaging the values obtained for each trajectory.

Figure 2 shows an example of extraction of motion cues using the EyesWeb open platform and specifically the EyesWeb Expressive Gesture Processing Library.



```
(System)  ...es\Lucas-KanadeFeatureTracking.eyw' ...
(System)  Initializing patch ...
(System)  Patch initialized.
(System)
(System)  Starting patch ...
(System)  Patch started.
```

Fig. 2. An EyesWeb application extracting motion cues (QoM, CI, Kinematical cues).

3.2.3 Layer 3

Layer 3 is in charge of segmenting motion in order to individuate motion and non-motion (pause) phases. QoM has been used to perform such segmentation. QoM is related to the overall amount of motion and its evolution in time can be seen as a sequence of bell-shaped curves (*motion bells*). In order to segment motion, a list of these motion bells has been extracted and their features (e.g., peak value and duration) computed. An empirical threshold has been defined for these experiments: the dancer is considered to be moving if its current QoM is above the 2.5% of the average value of the QoM computed along each whole dance fragment.

Segmentation allows extracting further higher-level cues, e.g., cues related to the time duration of motion and pause phases. A concrete example is the Directness Index (DI), calculated as the ratio between the length of the straight trajectory connecting the first and the last point of a motion trajectory and the sum of the lengths of each segment constituting the trajectory. Moreover, segmentation can be considered as a first step toward the analysis of the rhythmic aspects of the dance. Analysis of the sequence of pause and motion phases and their relative time durations can lead to a first evaluation of dance tempo and its evolution in time, i.e., tempo changes, articulation (the analogous to music legato/staccato). Parameters from pause phases can also be extracted to individuate real still standing positions from active pauses involving low-motion (hesitation or subtle swaying or tremble e.g., due to instable equilibrium or fatigue).

Furthermore, motion fluency and impulsiveness can be evaluated. They are related to Laban's Flow and Time axes. Fluency can be estimated starting from an analysis of the temporal sequence of motion bells. A dance fragment performed with frequent stops and restarts (i.e., characterized by a high number of short pause and motion phases) will result less fluent than the same movement performed in a continuous, "harmonic" way (i.e., with a few long motion phases). The hesitating, bounded performance will be characterized by a higher percentage of acceleration and deceleration in the time unit (due to the frequent stops and restarts), a parameter that has been demonstrated of relevant importance in motion flow evaluation (see, for example, Zhao 2001, where a neural network is used to evaluate Laban's flow dimension).

A first measure of impulsiveness can be obtained from the shape of a motion bell. In fact, since QoM is directly related to the amount of detected movement, a short motion bell having a high pick value will be the result of an impulsive movement (i.e., a movement in which speed rapidly moves from a value near or equal to zero, to a peak and back to zero). On the other hand, a sustained, continuous movement will show a motion bell characterized by a relatively long time period in which the QoM values have little fluctuations around the average value (i.e., the speed is more or less constant during the movement).

Fluency and impulsiveness are also related to the spectral content of the QoM: a movement having significant energy at high frequencies is a candidate to be characterized by low fluency.

3.2.4 Layer 4

In this experiment, Layer 4 collects inputs from Layers 2 and 3 (18 variables have been calculated on each detected motion phase) and tries to classify a motion phase in term of the four basic emotions anger, fear, grief and joy.

As a first step, statistical techniques have been used for a preliminary analysis: descriptive statistics and a one-way ANOVA have been computed for each motion cue. Results of such preliminary analysis can be found in (Mazzarino, 2002; Camurri, Lagerlof, Volpe, 2003; Volpe, 2003).

Decision tree models have then been built for classification. Five training sets (85% of the available data) and five test sets (15% of the available data) have been extracted from the data set. The samples for the test sets were uniformly distributed along the four classes and the five dancers. Five decision trees have been built on the five training sets and evaluated on the five test sets. The Gini's index of heterogeneity has been used for building the decision trees. Decision trees have been selected for this study since they produce rules that can be used to interpret the results. Comparison with other classification techniques (e.g., Neural Networks, Support Vector Machines) remains for possible future work.

The above-described techniques in the four layers have been implemented in our EyesWeb open software platform (Camurri et al. 2000). Free download of technical documentation and full software environment are available at www.eyesweb.org. The Expressive Gesture Processing Library (Camurri et al., 2003) includes these and other processing modules.

3.3 Results

Results from spectators' ratings are described in (Camurri, Lagerlof, Volpe, 2003). The results obtained on the five decision trees can be summarized as follows (results for the best model are reported in Tables 1 and 2 showing the confusion matrices for the training set and for the test set respectively).

Table 1. Confusion matrix for the training set for the best decision tree.

Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	71.9	28.1	46	10	2	6
Fear	60	61.7	38.3	15	37	1	7
Grief	86	47.7	52.3	10	19	41	16
Joy	74	64.9	35.1	13	8	5	48

Table 2. Confusion matrix for the test set for the best decision trees.

Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	41.7	58.3	5	3	0	4
Fear	13	30.8	69.2	6	4	2	1
Grief	12	41.7	58.3	2	0	5	5
Joy	13	46.1	53.8	4	0	3	6

Two models (3 and 5) fit the data set quite well; the rates of correct classification on the training set for these two models averaged over the four classes are 78.5% and 61.6%, respectively). Three models (1, 2, and 4) have difficulties in classifying fear. The rates of correct classification on the training set for these three models averaged over the four classes are 41.9%, 38.7%, and 36.0%, respectively). Models 2 and 4 also have problems with joy, which means that they distinguish correctly only between anger and grief.

A similar situation can be observed in the evaluation carried out on the test set: only models 3 and 5 are able to classify all four emotions correctly. Model 1 cannot classify fear, while models 2 and 4 cannot classify fear and joy.

The rates of correct classification on the test set for the five models averaged on the four classes are respectively: 40%, 36%, 36%, 26%, and 40%. Thus the average rate of correct classification on the five models is 35.6%. Except for model 4, they are all above chance level (25%). Model 5 can be considered as the best model, since it has a rate of correct classification of 40% and is able to classify all four emotions.

These rates of correct classification that at a first glance seem to be quite low (40% the best model) should however be considered with respect to the rates of correct classification from spectators who have been asked to classify the same dances. In fact, spectators' ratings collected by psychologists in Uppsala show a rate of correct classification (averaged over the 20 dances) of 56%.

The rate of correct automatic classification (35.6%) is thus in between chance level (25%) and the rate of correct classification for human observers (56%).

Furthermore, if the rate of correct classification for human observers is considered as reference, and percentages are recalculated taking it as 100% (i.e., relative instead of absolute rates are computed), the average rate of correct automatic classification with respect to spectators is 63.6%, and the best model (i.e., model 5) obtain a rate of correct classification of 71.4%.

By observing the confusion matrix of the best model (both for the test set and for the training set) it can be noticed that fear is often classified as anger. This particularly holds for the test set, where fear is the basic emotion receiving the lowest rate of correct classification since 6 of the 13 motion phases extracted from fear performances are classified as anger. Something similar can be observed in spectators' ratings (Camurri, Lagerlöf, Volpe, 2003).

A deeper comparison with spectator's ratings shows that while anger is generally well classified both by spectators and by the automatic system (60% for automatic recognition vs. 60.6% for spectators), quite bad results are obtained for fear (below chance level for the automatic classification). The biggest overall difference between spectators and automatic classification was observed for joy (70.4% for spectators vs. 27.7%, just above chance level, for automatic classification). In the case of grief instead, automatic classification performs better than human observers (48.3% for automatic classification vs. 39.8% for spectators): this happens in five cases and mainly for grief. In seven cases, the rate of correct classification for the automatic system is below chance level (and this always happens for fear). In one case, automatic classification did not succeed in finding the correct emotion (Fear – Dancer 4), but spectators obtained 67% of correct classification. In another case, spectators' ratings are below chance level (Grief – Dancer 5), but automatic classification could obtain a rate of correct classification up to 50%.

Dancer 1 obtained the lowest rates of correct classification both from spectators and from the models. Dancer 5 obtains similar rates from both. Dancer 2 is the best classified by spectators and also obtains a quite high rate (with respect to the other dancers) in automatic classification.

4 Analysis of Expressive Gesture in Music Performance

The second experiment investigates the mechanisms responsible for the audience's engagement in a musical performance. The aim of this experiment is again twofold: (i) individuating which auditory and visual cues are mostly involved in conveying the performer's expressive intentions, and (ii) testing the developed model by comparing its performance to spectators' ratings of the same musical performances.

In this experiment, expressive gesture was analysed with respect to its ability to convey the intensity of emotion to the audience. The study focused on the communication through visual and auditory performance gesture of emotional intensity and the effect of it on spectators' emotional engagement.

The research hypotheses combine hypotheses from Laban's Theory of Effort (Laban, 1947, 1963) with hypotheses stemming from performance research (see Clarke and Davidson, 1998; Palmer, 1997; Timmers, 2002), from research on the intensity of emotion and tension in music and dance (see Krumhansl, 1996; 1997; Sloboda and Lehmann, 2001; Scherer, 2003):

1. Emotional intensity is reflected in the degree of openness (release) or contraction (tension) of the back of the performer;
2. Emotional intensity is communicated by the main expressive means for a pianist: tempo and dynamics;
3. Intensity increases and decreases with energy level (speed of movements, loudness, tempo);
4. Intensity is related to the performer's phrasing: it increases towards the end of the phrase and decreases at the phrase boundary with the introduction of new material.

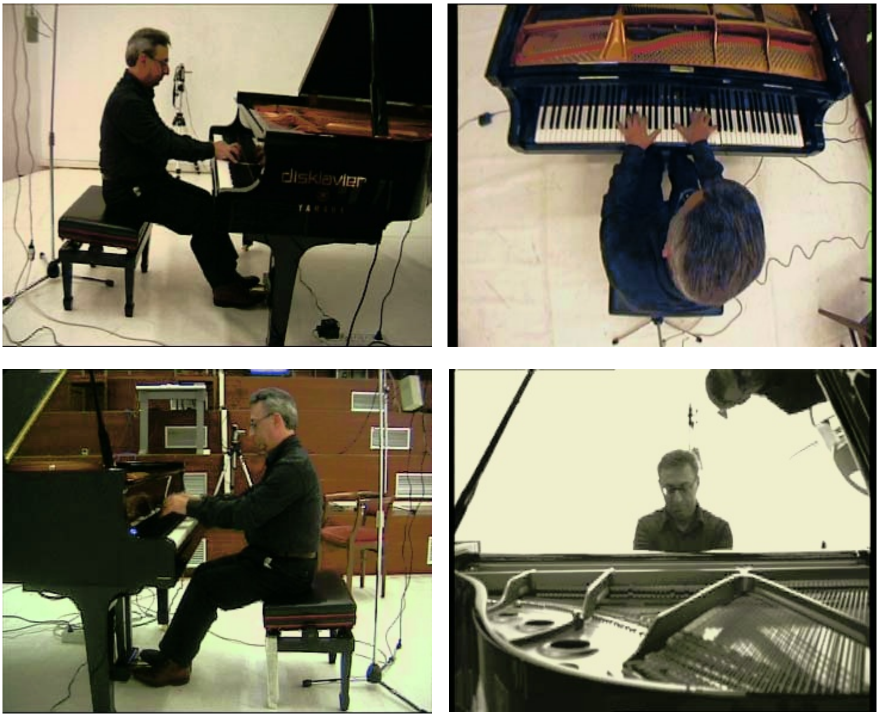


Fig. 3. Video recordings of the piano performances (left, top, right, and front views).

4.1 Description of the Experiment

A professional concert pianist (Massimiliano Damerini) performed Etude Op.8 No.11 by Alexandr Scriabin on a Yamaha Disklavier at a concert that was organized for the experiment's purpose. He performed the piece first without public in a normal manner and an exaggerated manner, and then with public in a normal, concert manner. Exaggerated means in this case with an increased emphasis in expressivity consistent with the style of performance of early 20th Century pianist style.

The Scriabin Etude is a slow and lyrical piece (Andante cantabile) in a late Romantic style that has a considerable amount of modulations. According to the pianist, the piece can be played with several degrees of freedom. Theoretically, the piece has a simple A B A with coda structure in which the A sections are repeated (A A' B A''

A''' C), but the pianist interpreted the line of the music differently: The first main target of the music is a release of tension halfway the B section. Everything preceding this target point is a preparation for this tension release. The A section is anyway preparatory; it leads towards the start of the B section, which is the real beginning of the piece. After this release of tension, the music builds up towards the dramatic return of the theme of the A section. This prepares for the second possible point of tension release halfway the coda at a general pause. The release is however not continued and the piece ends most sad.

The pianist performed on a grand coda piano (Yamaha Disklavier), which made it possible to register MIDI information of the performance. In addition, we did audio and video recordings from four sides (see Figure 3). The video recordings from the left were presented to the participants of the experiment.

Twelve students participated in the experiment; among them were four musicians. The participants saw the performances on a computer screen and heard them over high-quality (Genelec) loudspeakers twice. At the first hearing, they indicated the phrase boundaries in the music by pressing the button of the joystick. At the second hearing, they indicated to what extent they were emotionally engaged with the music by moving a MIDI-slider up and down. The order of the repeated performances was randomised over participants. The whole procedure was explained to them by a written instruction and a practice trial.

4.2 Analyses and Results

4.2.1 Layers 1 and 2

The key-velocity and onset-times of notes were extracted from the MIDI files (layer 1). From these data, the average key-velocity for each quarter note was calculated as well as inter-onset-intervals (IOI's) between successive quarter notes (layer 2). The quarter-note IOI is an accurate measure of local duration, while key-velocity corresponds well to local loudness.

The resulting profiles of quarter note key-velocity and quarter note IOI were highly similar for the three performances. The global pattern of increase and decrease in dynamics is indicated by arrows at the bottom of Figure 4. Local duration shows a similar pattern in the opposite direction. In addition, it shows the characteristic peaks of phrase-final lengthenings.

For the analysis of the movement of the pianist, we concentrated on the movement of the head, which shows both backward-forward movement (y-direction) and left-right movement (x-direction). The position of the head was measured, using the Lucas and Kanade feature-tracking algorithm (Lucas & Kanade, 1981) that assigns and tracks a specified number (in our case 40) of randomly assigned moving points within a region (layer 1). Velocity and acceleration has been calculated for each trajectory using the symmetric backward technique for the numeric derivative (layer 2). Average values of position and velocity among the forty trajectories were calculated for both the x and y component. In addition, the velocity values were integrated for the x and y movement to get a general measure of amount of movement over time. Redundancy in the number of points (i.e., forty points instead, for example, of just the barycentre of the blob) allowed us to get more robust and reliable values for velocity. A low-pass filter was applied to smooth the obtained data. Measures were summarized per quarter note in order to be comparable to the other measures.

The position of the head is plotted in Figure 5 for the two dimensions: left-right (upper panel) and backward-forward (bottom panel). The movement of the head was especially pronounced and especially consistent over performances in the left-right direction (correlation between p1 and p2 and between p2 and p3 was 0.79; it was 0.83 between p1 and p3). The periodic movement is relatively fast in the middle parts if the piece (B and A'') and slower in the outer parts. This suggests an intensification towards the middle followed by a relaxation towards the end.

As for the spectator ratings, firstly, the number of people who indicated a phrase-boundary was calculated for each quarter note in the performance by summing the number of boundary indications per quarter note over participants. This sum per quarter note was expressed as a multiple of chance-level, where chance-level corresponded to an even distribution of the total of segment-indications over quarter notes. This segmentation measure will be abbreviated as SM.

Secondly, the indication of emotional engagement was measured at a sampling rate of 10 Hz using a MIDI-slider that had a range from 0 to 127. The average level of the MIDI-slider (emotion measure, abbreviation EM) per quarter note was calculated for each participant separately. An EyesWeb patch application was developed to collect, process, and store participants data in real-time.

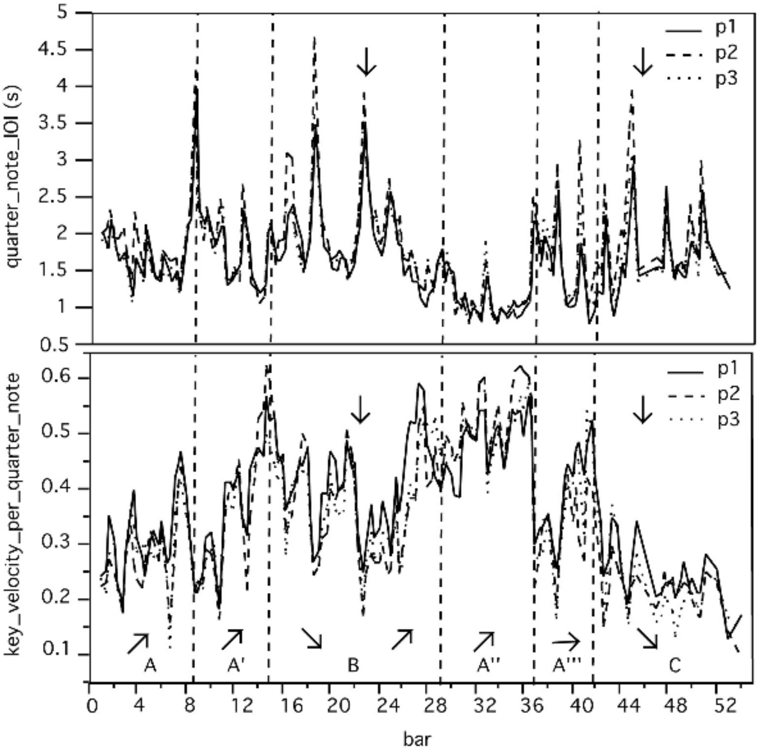


Fig. 4. The duration per quarter note and the key-velocity per quarter note as it varies throughout the Skriabin Etude. Separate plots for the three performances of the piece. Vertical lines indicate section boundaries. Arrows are explained in the text.

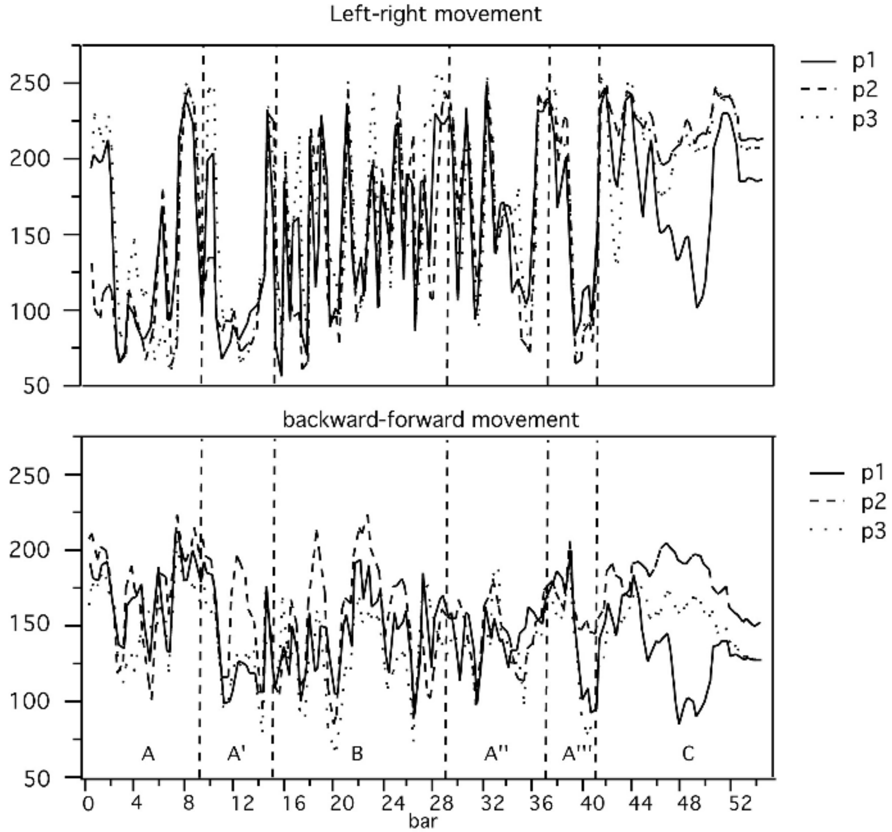


Fig. 5. The position of the head plotted per quarter note. Upper panel shows left-right position (x) and bottom panel the backward-forward position (y). Separate plots for the three performances of the piece. Vertical lines indicate section boundaries.

4.2.2 Layer 3

Correlations between performance measures were calculated to check the coherence between measures. Key-velocity and IOI were negatively correlated (r = -.51 on average). Velocity of head movement was positively correlated with key-velocity (r = .45 on average) and negatively with IOI (r = -.25 on average). The low correlation between values was partly due to the asynchrony between the periodicity of the measures. If peak-values (maximum for key and movement velocity and minimum for IOI) per two bars were correlated, agreement between movement and sound measures became higher. Especially the two velocity measures turned out to be highly correlated (r = .79 on average for key and movement velocity).

All performance measures showed a periodic increase and decrease. To check the relation between these periodicities and the musical structure, the location of mimima in key-velocity, and maxima in IOI, x-position and y-position were compared to the location of phrase-boundaries. Generally, the Skriabin Etude has a local structure of two-bar phrases. The forward and the left position of the performer were taken as start/end point for periodicity. IOI was most systematically related to the two-bar phrasing of the Skriabin-piece, followed by key-velocity. 55% of the phrase-boundaries were joined by a slowing down in tempo. The other phrase-boundaries were directly followed by a slowing down in tempo (a delay of 1 quarter note). For the key-velocity, 42% of the phrase-boundaries coincided with a minimum in key-velocity, 15% was anticipated by a minimum and 28% followed by a minimum. The period-boundaries of the movement of the pianist hardly synchronized with the score-phrasing. The location of these boundaries varied greatly with respect to the two-bar score-phrasing.

4.2.3 Layer 4

In this study, we had four hypotheses concerning the communication of intensity of emotion in musical performances.

Hypothesis 1 predicts that intensity of emotion is positively correlated with backward-forward movement (y). This hypothesis is easily tested and contradicted: the correlation between listeners' indication of intensity of emotion and backward-forward position is negative (r is -.23, -.50, -.29 for p1, p2 and p3, respectively). It is also contradictory with respect to the other performance data: y-position is negatively correlated with velocity and positively with IOI, this means that the performer moves forward in soft and slow and therefore less intense passages and backwards in louder and faster and therefore more intense passages (see hypothesis 3).

Hypothesis 2 predicts that tempo and dynamics cooperate to communicate intensity of emotion. This is made problematic by the fairly low correlation between IOI and key-velocity and by its different relation towards the score-phrasing. Instead the performance data suggests a differentiation in function between the two expressive means, whereby tempo strongly communicates phrasing.

Hypothesis 3 predicts high movement to correspond with intense dynamics and fast tempi. As we have seen in the previous section, dynamics and movement-velocity agree more strongly than movement-velocity and tempo. Especially the variation in velocity-peaks corresponds.

Hypothesis 4 relates phrasing to intensity of emotion. A clear phrase ending is predicted to coincide with a release in emotional intensity.

A series of multiple regression analyses were done. In the first analysis, quarter note IOI, key velocity, and movement velocity were used to predict EM. In the second analysis, the same variables were used to predict SM. In the third analysis, peak values per hyper-bar were used to predict average emotion measure per hyper-bar. All analyses were done directly and with a time-delay of one, two and three quarter notes of the performance data with respect to the listeners' data. The best R^2 obtained will be reported..

The subjective segmentation measure was rather well predicted by the model given the Rs^2 of .34, .33, .30 for p1, p2 and p3, respectively. From this model, IOI was the only significant variable. In other words, duration was a fairly good predictor of the variation in number of participants indicating a section-boundary. More participants indicated a phrase-boundary for longer durations.

EM was well predicted by the quarter note model, but even better by the second model that took the peak-values per hyper-bar to predict the average EM per hyper-bar. The quarter-note regression analysis had an R^2 of .45, .68, .50 for p1, p2, and p3 respectively, while the hyper-bar peak-value regression had an R^2 of .53, .82, and .56. Velocity was always the most significant variable and was the only significant variable for the hyper-bar peak-value regression. For the quarter-note regression movement velocity also reached significance for p2 and p3, and IOI for p2. All R^2 were relatively high for p2, which suggests that the exaggerated expression of this performance increased communication.

As a comparison, the analyses were repeated with x-position and y-position as independent movement variables instead of the more general movement velocity variable. The results did not improve or change from this alteration, instead x and y-position did not contribute significantly to any of the regressions.

These results confirm a differentiation between expressive means: tempo primarily communicated segmentation, while dynamics communicated emotion. Velocity of the movement was correlated with dynamics and may therefore also have reflected emotional intensity, but the sounding parameters were the main communicative factors.

The results are suggestive counter-evidence for hypothesis 4. The lack of tempo to explain variations in emotional intensity contradict that phrase-final lengthenings caused a release in emotional intensity. There was however another way in which phrasing and the dynamics of tension and release did relate, which was at a higher and more global level. Phrase final lengthenings occurred at a high rate and a local scale. At this scale the relation was weak. Major phrase boundaries that were indicated by a drop in tempo and dynamics were however followed by a clear release in intensity. Moreover the global variation of dynamics to which the variation in emotional intensity was so strongly related was the performer's way of communication of the overall-form: the first part is an introduction and builds up to the B section, which he considered as the real beginning of the piece. This beginning is again a preparation for the first target of the piece: the release of tension at the middle of the B section (see downward pointing arrows in Figures). Hereafter tension builds up towards the dramatic return of the theme, which leads via a repeat of the theme in contrasting dynamics to the second important target of the theme: the second possible release of tension at the general pause. After the general pause, the release is not given and all hope is lost. The piece ends most sad. The pianist most skilfully expressed this interpretation in the patterning of dynamics (see arrows in the key-velocity panel of Figure 4). The resulting phrasing is over the entire piece with subdivisions at measures 22 and 36. The return of the theme is the culminating point of the piece where after tension can release. According to the pianist, this tension cannot however be fully resolved.

4.4 Summary and Conclusions from the Experiment

This study had two aims (i) individuating which auditory and visual cues are mostly involved in conveying the performer's expressive intentions and (ii) testing the developed model by comparing their performances with spectators' rating of the same musical performances. The auditory and visual cues most involved in conveying the performer's expressive intentions were hypothesised to be key-velocity, IOI, movement velocity, and the openness or contraction of the performer's posture. In addition, a relation between phrasing and emotional tension-release was expected.

The analyses of performance data suggested an opposite relation between emotional intensity and the performer's posture. The pianist leaned forward for softer passages and backward for intensive passages. In addition it suggested a differentiation in expressive means with tempo on one side and key-velocity and movement velocity on the other side.

When relating the performers data to the listeners' data, this differentiation in expressive means was confirmed. Tempo communicates phrase boundaries, while dynamics is highly predictive for the intensity of felt emotion. Hardly any evidence was found for movement cues to influence listeners' ratings. The sound seemed the primary focus of the participants and vision only subsidiary. The local phrase-boundaries indicated by tempo did not lead to release of emotional intensity. The modulation of dynamics over a larger time-span communicates the overall-form of the piece and, at that level, intensity did increase and decrease within phrases.

5 Discussion

The modelling of expressive gesture is receiving growing importance from both research and industry communities, even if we can consider it at its infancy. The main contributes of our research are to the definition of a unified multimodal conceptual framework for expressive gesture processing, to the results obtained from the two types of experiment described in the paper. Further, a collection of software modules for cue extraction and processing has been developed to support such research. The conceptual framework proved to be useful and effective in two different scenarios, well represented by the two experiments described in the paper.

In the first experiment, we focused on the communication of basic emotions from a dancer to the audience, while in the second experiment we focused on the mechanisms that possibly cause emotional engagement in the audience.

The dance experiment can be considered as a first step and a starting point toward understanding the mechanisms of expressive gesture communication in dance. A collection of cues that have some influence in such a communication process has been individuated, measured, and studied. A first attempt of automatic classification of motion phases has been carried out and some results have been obtained (e.g., an average rate of correct classification not particularly high, but well above chance level). Some directions for future research also emerged. For example, other classification techniques could be employed and their performances compared with what we obtained with decision trees. Some aspects in dance performance that were only marginally considered should be taken into account. In particular, aspects related to rhythm should be further investigated. Expressive cues like impulsiveness and fluency should be further worked out. Moreover, perceptual experiments would be needed to empirically validate the extracted expressive cues.

The music experiment can be considered as a first step towards the understanding of the relation between movement and sound parameters of a performance, their expressive forms and functions, and their communicative function for spectators. A next step, should involve a larger variety of performances and a larger collection of calculated cues. Cues should be fitted to the responses of individual spectators in order to get a deeper as well as broader understanding of these complex phenomena.

Acknowledgements

We thank Ingrid Lagerlöf and Marie Djerf for the joint work carried out in the experiment on expressive gesture conveying emotions in dance performances, Roberto Dillon and other colleagues at the DIST - InfoMus Lab, the EyesWeb development team (Paolo Coletta, Massimiliano Peri, and Andrea Ricci), the students Andrea Pertino and Riccardo Casazza for their contribute in the set up of the music performance experiment and in the extraction and analysis of data, and last but not least Massimiliano Damerini for discussion and his artistic contributes in providing the material of the studies.

References

1. Bobick, A.F., Davis J. (2001), “The Recognition of Human Movement Using Temporal Templates”, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3): 257-267
2. Bradsky G., Davis J. (2002), “Motion segmentation and pose recognition with motion history gradients”, Machine Vision and Applications 13:174-184.
3. Cadoz C., and Wanderley, M. (2000), “Gesture – Music”, in Wanderley, M. and Battier M. eds. (2000), “Trends in Gestural Control of Music.” (Edition électronique) Paris: IRCAM.
4. Camurri A., Lagerlof I., and Volpe G. (2003), “Emotions and cue extraction from dance movements”, International Journal of Human Computer Studies, Vol.59, No.1-2. pp.213-225, Elsevier.
5. Camurri, A., De Poli G., Leman M. (2001), “MEGASE - A Multisensory Expressive Gesture Applications System Environment for Artistic Performances”, Proc. Intl Conf CAST01, GMD, St Augustin-Bonn, pp.59 – 62.
6. Camurri A., Hashimoto S., Ricchetti M., Trocca R., Suzuki K., Volpe G. (2000) “EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems.” Computer Music Journal, 24:1, pp. 57-69, MIT Press, Spring 2000.
7. Canazza S. De Poli G., Drioli C., Rodà A., Vidolin A. (2000), “Audio morphing different expressive intentions for Multimedia Systems”, IEEE Multimedia, July-September, Vol. 7, N° 3, pp. 79-83.
8. Chi D., Costa M., Zhao L., and Badler N. (2000), “The EMOTE model for Effort and Shape”, ACM SIGGRAPH '00, New Orleans, LA, pp. 173-182.
9. Clarke, E. F. Davidson, J. W. (1998), “The body in music as mediator between knowledge and action”, in W. Thomas (Ed.).Composition, Performance, Reception: Studies in the Creative Process in Music, Oxford University Press, 74-92
10. Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W. and Taylor J. (2001), “Emotion Recognition in Human-Computer Interaction”, IEEE Signal Processing Magazine, no. 1.
11. Friberg A., Colombo V., Frydén L., and Sundberg J. (2000), “Generating Musical Performances with Director Musices”, Computer Music Journal, 24(3), 23-29.
12. Hashimoto S., (1997), “KANSEI as the Third Target of Information Processing and Related Topics in Japan”, in Camurri A. (Ed.) “Proceedings of the International Workshop on KANSEI: The technology of emotion”, AIMI (Italian Computer Music Association) and DIST-University of Genova, pp101-104.
13. Kilian J. (2001), “Simple Image Analysis By Moments”, OpenCV library documentation.
14. Krumhansl, C. L. (1996), “A perceptual analysis of Mozart’s piano sonata K.282: Segmentation, tension and musical ideas”, Music Perception, 13 (3), 401-432.
15. Krumhansl, C. L. (1997), “Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine’s choreography of Mozart’s *Divertimento* No. 15”, Musicae Scientiae, 1, 63-85.
16. Kurtenbach and Hulteen (1990), “Gesture in Human-Computer Interaction”, in Benda Laurel (Ed.) The Art of Human-Computer Interface Design.
17. Laban R., Lawrence F.C. (1947), “Effort”, Macdonald&Evans Ltd. London.
18. Laban R. (1963), “Modern Educational Dance” Macdonald & Evans Ltd. London.

19. Lagerlof, I. and Djerf, M. (2001), “On cue utilization for emotion expression in dance movements”, Manuscript in preparation, Department of Psychology, University of Uppsala.
20. Liu Y., Collins R.T., and Tsin Y. (2002), “Gait Sequence Analysis using Frieze Patterns”, European Conference on Computer Vision.
21. Lucas B., Kanade T. (1981), “An iterative image registration technique with an application to stereo vision”, in Proceedings of the International Joint Conference on Artificial Intelligence.
22. McNeill D. (1992), “Hand and Mind: What Gestures Reveal About Thought”, University Of Chicago Press,
23. Palmer, C. (1997). “Music Performance”, Annual Review of Psychology, 48, 115-138.
24. Pollick F.E., Paterson H., Bruderlin A., Sanford A.J., (2001), “Perceiving affect from arm movement”, Cognition, 82, B51-B61.
25. Scherer, K.R. (2003) “Why music does not produce basic emotions: pleading for a new approach to measuring the emotional effects of music”, in Proc. Stockholm Music Acoustics Conference SMAC-03, pp.25-28, KTH, Stockholm, Sweden.
26. Sloboda J.A., Lehmann A.C. (2001) “Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude”, Music Perception, Vol.19, No.1, pp.87-120, University of California Press.
27. Timmers, R. (2002). “Freedom and constraints in timing and ornamentation: investigations of music performance”. Maastricht: Shaker Publishing.
28. Wanderley, M. and Battier M. eds. (2000), “Trends in Gestural Control of Music.” (Edition électronique) Paris: IRCAM.
29. Zhao, L. (2001), “Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures”, Ph.D Dissertation, University of Pennsylvania.