Vis Comput (2016) 32:83–98
DOI 10.1007/s00371-014-1057-8

ORIGINAL ARTICLE

Laban descriptors for gesture recognition and emotional analysis

Arthur Truong · Hugo Boujut · Titus Zaharia

Published online: 3 January 2015

Abstract In this paper, we introduce a new set of 3D gesture descriptors based on the laban movement analysis model. The proposed descriptors are used in a machine learning framework (with SVM and different random forest techniques) for both gesture recognition and emotional analysis purposes. In a first experiment, we test our expressivity model for action recognition purposes on the Microsoft Research Cambridge-12 dataset and obtain very high recognition rates (more than 97 %). In a second experiment, we test our descriptors' ability to qualify the emotional content, upon a database of pre-segmented orchestra conductors' gestures recorded in rehearsals. The results obtained show the relevance of our model which outperforms results reported in similar works on emotion recognition.

Keywords Gesture expressivity model · Laban movement analysis · Motion features · Gesture recognition · Expressivity analysis · Machine learning

1 Introduction

When human beings interact with their environment, they perform what we call "gestures", defined as motions of the body that contain meaningful information [1]. Within this framework, action recognition and expressivity analysis are highly challenging issues that involve both computer vision and machine learning methodologies. The generic objective is to semantically qualify body movements, postures, gestures, or actions with the help of mid- or high-level features built upon low-level visual features.

Gesture analysis recently received a growth of interest, notably since 3D body tracking became facilitated by the emergence of general public, affordable depth cameras (e.g., Kinect). Gesture analysis and interpretation is highly useful for numerous applications: e-health, video games, artistic creation, video surveillance, human-computer interaction, immersive and affective communication. However, the issue of high-level, semantic interpretation of gestures still remains a challenge, which requires the elaboration and development of effective gesture descriptors.

Throughout the past decade, several researchers investigated the possibility of using body movements to recognize gestures [2–4], emotional categories [5–7], emotional dimensions defined in continuous spaces [8] or movement qualities [9–11]. However, to our very best knowledge, only a few papers proposed effective models of gesture descriptors [12–15] and such models are often incomplete, poorly intuitive, focused on very specific gestures, and rarely refer to perceptual studies on relevant motion features for gestures interpretation. In a certain sense, there is a lack of a generic computational model of gestures that can be exploited for high-level semantic interpretation.

In this paper, we propose a set of body motion descriptors, inspired from the Laban movement analysis model [16] and aiming at characterizing the pertinent features involved in a dynamic gesture. The rest of the paper is organized as follows: Sect. 2 presents a state of the art review of existing models and approaches involved in various gesture analysis applications. In Sect. 3, we introduce the proposed gesture description model, which is described in detail. Sec-

A. Truong (✉) · H. Boujut · T. Zaharia
ARTEMIS Department, Institut Mines-Telecom,
Telecom SudParis, CNRS UMR 8145-MAP5,
9 rue Charles Fourier, 91 011 Évry Cedex, France
e-mail: arthur.truong@telecom-sudparis.eu

H. Boujut
e-mail: hugo.boujut@telecom-sudparis.eu

T. Zaharia
e-mail: titus.zaharia@telecom-sudparis.eu

tion 4 deals with our first experiment based on MSRC-12 dataset which consists of gestures recognition and presents our results. In Sect. 5, we first introduce our orchestra conductor gestures database as well as its annotation protocol. Then, we present the results of emotional content recognition on this dataset. Finally, Sect. 6 concludes the paper and opens perspectives of future work.

2 Related work

A tremendous research effort has been dedicated in the past two decades to the fields of gesture/posture/action recognition and gesture-based affect analysis. A first category of approaches aims at directly interpreting gestures as actions, without considering the emotional/expressivity aspects.

2.1 Recognition of gestures as specific actions

Let us first mention approaches based upon global representations, which encode the visual observation as a whole. In [2], Wang et al. use Radon transform on silhouettes extracted from video sequences to feed HMM and recognize 5 human activities with high recognition rates (up to 98 %). In [17], the authors propose new silhouette-based descriptors for action recognition. After having performed foreground extraction, they localize the object silhouette per frame and convert it to a 1D time series. Then, they compute the symbolic aggregate approximation (SAX) of the time series, which consists of reducing the length of the time series to a small number of segments and sampling the values to a given number of symbols. They use these silhouette representations to feed Random Forests. On the Weizmann dataset, the method yields an accuracy rate of 89 %. In [18], Chen et al. use the angles of a "star" based on the head and the 4 limbs extracted from 2D images sequences as input vector of HMMs so as to recognize body postures with very satisfying results (98 % recognition rate). In [19], Cimen et al. propose 3 types of global motion descriptors (posture, dynamic and frequency descriptors) based on the position of end effectors (wrists, ankles, and head) throughout the motion. These descriptors are applied to gestures embodying 4 different affective states after space and time warping. Emotion recognition is performed with SVM for the 7 possible descriptor type combinations, with recognition rates reaching 91 %. In [20], a new time warping method is applied to skeleton-based gesture comparison. Each motion sequence is described by body joint angle series recorded by a depth sensor (e.g. Kinect). For initial joint angle sequences of length m, the aim is to obtain warped sequences of length c, comprising segments with a temporal length in the range of $[\lambda - \delta, \lambda + \delta]$, where the segment size λ and the slack size δ are configured. The set of segment-wise warping degrees is computed by maximizing an objective function taking into account Pearson's linear correlation coefficients (PCC) between reference and initial joint trajectories, as well as the different weights associated with each body joint. The normalized distance and correlation results obtained for the warping of different actions from Carnegie Mellon University motion dataset outperform the results obtained with other warping techniques (uniform time warping, dynamic time warping, etc…). The approach of Huang and Xu [3] is based on different plane sections of a silhouette at a particular frame and on the projection of these sections onto the landmarks attached to two orthogonal cameras. This "envelop shape" vector is used as input of a HMM to recognize 9 actions performed by 7 actors, with recognition rates superior to 95 and 83 % in the respective cases of subject-dependent and subject-independent recognition. In [21], Singh and Nevatia propose a combination of Dynamic Bayesian Action Networks (DBAN) with intermediate 2D body parts models for both pose estimate and action recognition tasks. Composite actions are decomposed into sequences of primitives based upon the variations between the 3D key poses of consecutive frames. The 3D poses are mapped onto 2D parts models so that a state of the DBAN at time t is denoted by a composite action ce_t, a primitive action pe_t, the time elapsed since the primitive action started d_t and a 2D pose p_t. The method is tested on a hand gesture dataset including about 500 gestures, with recognition accuracy results around 85–90 % for each action. In [22], the authors compute local binary patterns (LBP) from 3 orthogonal planes (corresponding to x, y, and temporal dimensions) all along video sequences. For each sequence, it results in features computed all over a "bounding volume" that are reduced to a histogram used to feed HMM model whose hidden states consist of activity classes. The approach is tested on the 10 different activities of Weizmann dataset [23] and yields recognition results superior to 95 %. In [24], Jiang et al. introduce a hierarchical model for action recognition based upon body joints trajectories recorded by a Kinect. A first step consists of assigning each gesture to a group according to the body parts motions during action performance. Then, for each motion-based group, a KNN classifier is trained which takes joints motions and relative positions as input. Bag-of-words are used for dimensional reduction and each word of the codebook is given a weight. At the test stage, the gesture is first given a motion-based group and the appropriate KNN classifier is used to give it the right label. This method is tested on UCFKinect action dataset, with accuracy results close to 97 %, and also on MSRC-12 gesture dataset (which will be presented below) with recognition rates reaching 100 % for certain gestures, even though confusions remain present for certain actions (worst rates close to 82, 86 %).

Other approaches are based upon local representations of motion, where the observation is described as a collection of patches or local descriptors. In [25], the saliency of the points

is evaluated relatively to the information contained in a spatiotemporal neighborhood, and a time-warping technique is applied to normalize 152 samples consisting of aerobic exercises performed by amateur participants. In [26], the saliency is obtained by a minimization of an energy function putting at stake color, intensity, and orientation conspicuities. The use of Nearest-Neighbor classifier on KTH dataset [27], consisting of 6 types of human action, gives a global recognition rate of 88.3 %. The approach is also tested on HOHA dataset [28] containing video samples from 32 movies labeled according to at least one of 8 action categories. In [29], histograms of gradient (HOG) descriptors are exploited to construct a local motion signature. The method is applied to characters present in 2D videos. In [30], Li et al. introduce a new action recognition method based upon bag-of-words and spatiotemporal features. For each spatiotemporal interest point (STIP) p_i detected in a video frame, they compute its corresponding local spatiotemporal features f_i, as well as its coding coefficients v_i represented as the posterior probability of the local spatiotemporal feature f_i belonging to each visual word d_j. The originality of the approach lies in the fact that the coding coefficients v_i do not only take into account the spatial similarities, but also limit the effects of quantization errors due the approximation of local spatiotemporal features by visual words. 3D embedded cuboids are defined in the neighborhood of each STIP. These "context" cuboids are divided into subparts, and density histograms are computed over each one of these subparts, according to the points that belong to it. A resulting density histogram describing STIP contextual structure is used as feature vector to feed multi-class nonlinear SVM. The testing of the approach on four benchmark datasets gives accuracy results globally higher than 92 %. In [31], Wu et al. introduce a gesture recognition framework based upon video segmentation. For an input video stream, we consider all the possible sub-segments of a predefined length L. The temporal structure of such sub-segments is modeled by a "hypothesis" consisting of a temporal cells pyramid (the total number of cells is equal to 8 and comprises a "zero" level cell corresponding to the whole video sample). Each cell is described by four descriptors (trajectory shape, Histograms of Oriented Gradients, Histograms of Optical Flow, Motion Boundary Histogram), so that the sub-segment feature vector consists of a concatenation of these 8 cell-descriptions. For each action class i, M temporal models are employed, each one of them having a different length L_m (and consequently a different number of hypothesis H_m). The resulting feature vector, for a given model and a given hypothesis, is used as input of a latent SVM method. The approach is tested on the Olympic Sports and HMDB51 datasets and gives recognition rates reaching 84 %. Always in a 2D video context, an interesting framework for hand gesture recognition, based on gesture segmentation into sub-gestures, is proposed in [32]. The methods are tested on two databases, respectively, composed of hand-signed digits and of American Sign Language occurrences. Finally, in [33], a new framework is proposed for recognition of both hand poses and hand gestures. After the acquisition of the depth image by a Kinect sensor, a first algorithm implementing mean-shift segmentation and a palm detection procedure is applied and provides hand segmentation. A hand pose recognition system using an SVM classifier is trained with descriptors based on Gabor filtering. The results obtained on American Sign Language (ASL) dataset show recognition rates reaching 97 % for certain letters. Gesture recognition task requires the decomposition of the motion into "hold-movement-hold" sequences. For each sequence, the hand centroid trajectory defines an angle series which is used to feed HMM. The method is tested on a database that included 16 gestures performed by ten users. At least 5 gestures are above 90 % of recognition rates, while 3 are below 50 %.

The analysis of the previous approaches shows that the descriptors used for action or posture recognition are often dedicated to visual indices of motion. Thus, they usually fail to take into account the semantic aspects of motion (intersubjectivity, expressivity, intentionality).

Another field of research aims at defining high-level emotions, affects or expressivity in gestures, and tackles the issue of semantic characterization of the body motions. In this case, the underlying hypothesis is that the content of a gesture cannot solely refer to a specific action, but must also relate to its intentional and communicative aspects. These related methods proposed in the literature are presented in the following section.

2.2 Emotion recognition and expressivity analysis

Laban movement analysis (LMA [16]), proposed by choreograph and dancer Rudolf Laban, provides a consistent representation of gestures expressivity. The underlying principle consists of characterizing the body motion in terms of a fixed number of qualities independently on the specific articulation of the motion and will be introduced in details in Sect. 3.1. Laban showed that the analysis of expressivity is the key to understand gestures intentional and communicative aspects. LMA has become a reference framework for different types of approaches.

A first category of approaches aims at characterizing gestures in terms of Laban qualities. Such methods require the use of machine learning techniques to infer expressive representations from low/mid-level features. In [9], a Bayesian fusion approach is used that fuses body motion features for identifying the Shape movement quality from dancers improvisations. In [10], four neural networks are exploited. They are trained with motion features notably based on curvature, torsion, swivel, and wrist angles, so as to characterize gestures with four Laban Effort sub-qualities (Fig. 1). Laban's

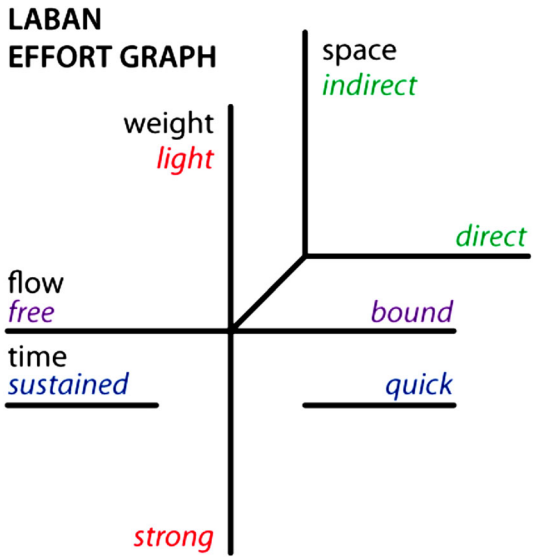


Fig. 1 Laban effort sub-qualities representation

model is also used in [34], where LMA features are computed using a collection of a neural networks with temporal variance aiming at creating a classifier that is robust with regard to input boundaries.

A different family of approaches consists of inspiring from Laban concepts to build expressive motion descriptors. Usually, the resulting mid-level features are used to determine higher-level features, like emotions of affective states, with the help of machine learning techniques. In [6], Camurri et al. investigate the possibility to use decision trees to classify motions of dancers and musicians in a discrete set of 4 emotional categories (joy, anger, fear, grief) with the help of mid-level features modeling the motion expressivity. However, the proposed descriptors are limited to some global energy characterization of the motion. The authors compare the recognition results obtained to those reported in [35] from spectators watching dancers and characterizing emotions expression. In [13], Glowinsky et al. propose a so-called minimal motion description, based on head and hands trajectories in 2D portrayal videos. The objective is to classify gestures in a continuous emotional 2D space corresponding to a valence-arousal representation [36]. After having reduced features' dimensionality to four clusters and performed recognition, they show that the major part of the emotion portrayals used can be associated with one of the clusters. Finally, let us cite the work of Bernhardt and Robinson [37], where energy profiles are used for performing a motion-based segmentation. Each segment is described by its trajectory. A k-means clustering approach is used for deriving a set of primitive trajectories. Such primitives are then classified using a standard SVM approach. Only four emotion categories are here considered.

A third category of approaches aims at quantifying Laban qualities. In such cases, the expressive characterization is directly determined as a function of dynamic features and is compared to the annotation carried out by experts. In [38], Nakata et al. propose a set of motion descriptors, each one referring to a LMA component, and apply these descriptors to 5 dancing robots gestures annotated with the help of 4 emotional categories. Factor analysis is used to establish causality between Laban qualities and emotions. In [39], Hachimura et al. implement similar descriptors. The processed results are compared to specialists' annotation, and the matching occurs only for certain qualities. In [40] other expressive features are defined for each frame of a gesture, to index various gestural contents with local motion "keys". Such keys are used for database querying purposes. Let us finally quote the work of Samadani et al. [11] who were inspired by [38,39] to propose different Laban features quantifications and apply their descriptors to pre-defined gestures involving hands and head, designed by motion professionals and annotated both in terms of LMA factors (on 5 points Likert scales) and emotions (6 categories). "Weight" and "Time" LMA dimensions show high correlation coefficients between annotations and quantification, which allows representing each emotion in the space generated by these two qualitative dimensions.

The analysis of the state of the art shows that the expressive character of gestures is crucial in the designing of a gesture description model. If the approaches presented in Sect. 2.2 satisfy this criterion, several of them require the help of motion experts to annotate datasets or to validate the quantification of LMA components for gestures of a very specific type [9,10,37]. In addition, the utility of Laban qualities quantifications [11,38–40] is not demonstrated for the recognition of other types of content than those specifically designed for recognition of Laban qualities(dance). Moreover, the global characterization approaches [11,38,39] fail to capture the local aspects of motion.

In this context, the challenge is to design a new expressive model of gesture, aiming at quantifying abstract concepts related to motion expressivity and able to incorporate temporal and local aspects of the motions. These quantifications of expressive features will be exploitable directly for characterizing high-level contents like actions and emotions in a machine learning framework, without explicitly determining the underlying abstract concepts.

In order to derive intuitive and meaningful gesture descriptions, we have considered the Laban Movement Analysis model, briefly recalled in the following section.

3 Proposed approach

3.1 Our framework: laban movement analysis

Rudolf Laban was a Hungarian dancer, choreograph and dance theoretician who developed a movement analysis

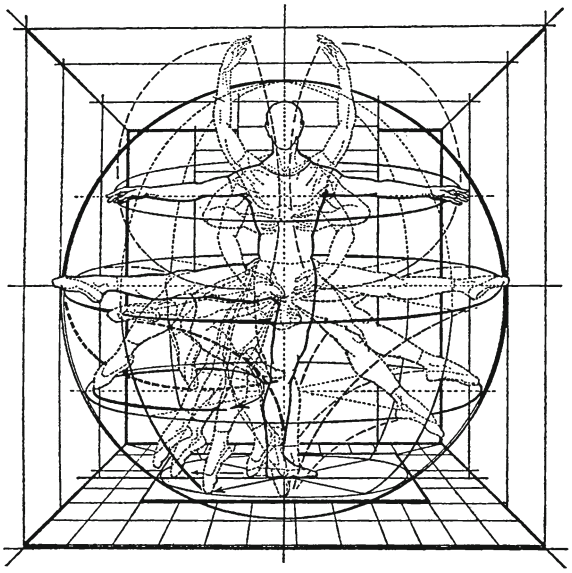


Fig. 2 Laban kinesphere representation

method, called "Laban Movement Analysis" (LMA [16]). The principle consists of describing movement in terms of qualities relating to different characterizations of the way this movement is performed, but independently on its precise trajectory in space. The conceptual richness of this gesture analysis model, originally designed for dance teaching, permitted its extension to the study of all types of movements.

More precisely, the LMA model includes five major qualities: Body, Relationship, Space, Effort, and Shape [41].

The Body component deals with body parts' usage, coordination, and phrasing of the movement.

The Relationship component refers to the relationships between individuals and is particularly suited in the case of group performances.

The Space component refers to the place, direction, and path of the movement, and is based on the concept of kinesphere ncluding the body throughout its movement (Fig. 2).

These three first qualities relate to the structural characterization of the movement.

The Effort component depicts how the body concentrates its effort to perform the movement and deals with expressivity and style. The Effort is further decomposed into the following 4 factors (Fig. 1):

- *Space* (not to be confused with space quality), which defines a continuum between direct (or straight) movements and indirect (or flexible) movements,
- *Time* which separates movements between sudden and sustained (or continuous) ones,
- *Flow* which describes movements as constrained or free,
- *Weight*, to distinguish between heavy and light movements.

The *Shape* description is decomposed into three sub-components:

- *Shape flow* sub-component describes the dynamic evolution of the relationships between the different body parts,
- *Directional movement* sub-component describes the direction of the movement toward a particular point,
- *Shaping* sub-component refers to body forming and how the body changes its shape in a particular direction: rising/sinking, retreating/advancing and enclosing/spreading oppositions are, respectively, defined along the directions perpendicular to the horizontal, vertical, and sagittal planes.

The effort and shape qualities refer to the qualitative aspect of body motion.

Some first studies [9,10,34] attempt to identify and classify the Laban qualities, based on visual gesture descriptors, with the help of supervised classification approaches. Some other approaches validated quantifications of Laban qualities in very specific contexts [11,38,39]. Such techniques show that a mid-level Laban representation can be obtained starting from visual descriptors.

In our case, the adopted approach is slightly different. Our objective is to define a set of descriptors able to characterize the individual Laban qualities and then to exploit them directly for gesture recognition purposes and emotional analysis in a machine learning framework, without explicitly determining the underlying Laban qualities. To our very best knowledge, such a goal has never been achieved before.

For this purpose, we have retained solely the space, effort, and shape qualities, which are mostly used and can be applied for generic applications. In effect, relationship and body qualities are strongly oriented towards motion structural aspects. Thus, their contribution to the designing of an intermediary gesture model which aims at characterizing various high-level contents (not only pre-determined actions, but also emotions, affects or metaphoric contents) does not seem to be useful. Moreover, among sub-qualities of effort, we discarded the space component since, as explained in [6,35], it is often characterized by features related to shape or space qualities.

Let us now describe the dedicated descriptors proposed for each of the retained qualities.

3.2 Descriptor specification

The proposed descriptors are based on 3D trajectories associated with the body skeleton joints that can be recorded with a depth sensor (i.e., Kinect camera) at a rate of 30 frames per second. The Kinect sensor provides a maximum number of 20 joints (Fig. 3), corresponding to the following body parts: center of the hip, spine, center of the shoulders, head,

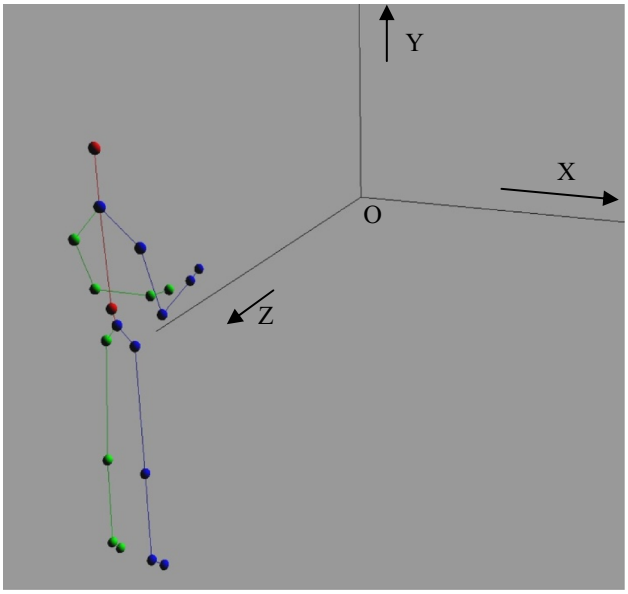


Fig. 3 Body skeleton joints at a particular frame

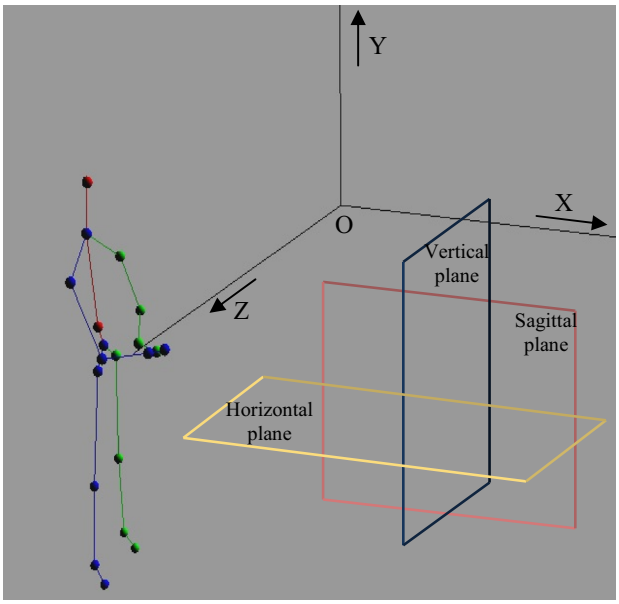


Fig. 4 Skeleton joints' new position after transforms application and body planes representation

left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand, left hip, left knee, left ankle, left foot, right hip, right knee, right ankle, and right foot.

Each body joint trajectory i is represented as a sequence of $(x_{i,t}, y_{i,t}, z_{i,t})_{t=0}^{N-1}$ coordinates in a 3D Cartesian system of coordinates ($Oxyz$). Here, N denotes the total number of frames of the considered gesture. For each gesture, before the computation of such characteristics, several elementary transforms are applied to the body at each frame of its trajectory. The objective is to set each body joint in a new position at frame t $(x_{i,t}^{\text{trans}}, y_{i,t}^{\text{trans}}, z_{i,t}^{\text{trans}})_{t=0}^{N-1}$ so that the $(x\text{O}y)$, $(y\text{O}z)$ and $(z\text{O}x)$ planes (Fig. 4), respectively, correspond to sagittal, vertical, and horizontal body planes. The aim of such transforms is to put the shoulders and the hip center in a same plane parallel to $(y\text{O}z)$ plane and put both shoulders at the same height. Thus, for each gesture and for each frame, we apply the following transforms:

- First of all, we translate the body to set the hip center at the origin of the landmark.
- Second, we apply a rotation around the y axis to the body to set left and right shoulders in a plane parallel to $(y\text{O}z)$ plane.
- Then, we perform a rotation around the z axis to the body to set shoulder and hip centers in a plane parallel to $(y\text{O}z)$ plane.
- A final rotation around the x axis consists of setting left and right shoulders in a plane parallel to $(z\text{O}x)$ plane.
- Finally, we translate the body to put the hip center at its initial position.

Figure 4 illustrates the result of this body alignment process.

Let us now present the features retained for each Laban quality.

The space quality is described with the help of 9 features. The first feature is the total length of the head trajectory. Then, we compute several values related to the forward–backward global motion, i.e. the motion in the direction perpendicular to vertical plane (Fig. 4). For this purpose, we retain

– the number of zero crossings of the first derivative of the head's component in this direction (which measures the number of head's retreats/advances)

$$\left(n_{\text{ZC}}\left(x'^{\text{trans}}_{\text{Head},t}\right)_{t=0}^{N-1}\right), \tag{1}$$

– the amplitude of the head movement following this direction:

$$\left(\max_{t=0}^{N-1}\left(\left|x^{\text{trans}}_{\text{Head},t}\right|\right) - \min_{t=0}^{N-1}\left(\left|x^{\text{trans}}_{\text{Head},t}\right|\right)\right), \tag{2}$$

– the relative temporal instant:

$$t_{\max}/N, \tag{3}$$

of maximum's reaching in this direction, or "forward maximum".

Then, we consider the forward tilt angle Φ defined for each frame as the angle between the vertical direction y and the axis binding the center of the hip and the head, expressed in

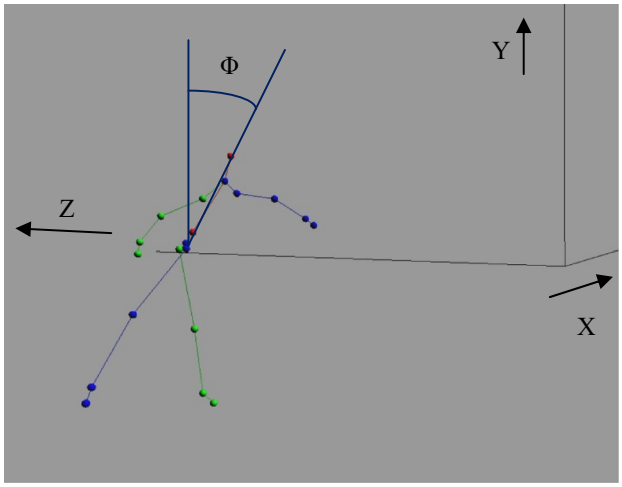


Fig. 5 Illustration of forward tilt angle at a particular frame (the angle is positive in this case)

radians (Fig. 5). The resulting tilt angle sequence is described by the following 5 parameters: mean, standard deviation, ratio between the global minimum and maximum values, number of local maxima, and relative temporal instant of the global maximum.

The time subcomponent of the effort quality is characterized by 8 features. We first considered the total gesture duration. Then, we compute features based on the total kinetic energy sequence defined in [13] for head and left and right hands:

$$E_c(t) = \sum_i (m_i . v_i^2(t)), \quad t \in [0; N-1], \tag{4}$$

where i indexes the joints retained (hands and head), m_i is the mass of i-th joint, and v_i its velocity. The values associated to head and hands masses correspond to averages computed over a set of individuals proposed in [42]. Each frame in the gesture sequence is categorized into two phases, corresponding to either high/medium activity or low activity (pause). Thus, frames with kinetic energy inferior to 1/10 of the maximal energy reached throughout the gesture are considered as pauses. Given this segmentation of the gesture between activity periods and pauses, we compute the percentage of low-activity frames relatively to the whole sequence, as well as the mean, standard deviation, and maximum value of the kinetic energy sub-sequences for both high/medium and low-activity components.

The flow subcomponent of the effort quality is described with the help of the third-order derivative of the left- and right-hands trajectories, so-called jerk. The jerk series are statistically described by the following entities: mean, standard deviation, ratio between the maximal and mean values, number of local maxima, and the relative temporal instant of the global maximum. This results in a 10 features vector (5 components for each hand).

For the weight subcomponent of effort quality, we consider the vertical components of the velocity and acceleration sequences (i.e., $y'_{.,t}$ and $y''_{.,t}$ signals) associated with 3 joints: the center of the hip, the left hand, and the right hand. The following 5 features are retained: mean, standard deviation, maximal amplitude, number of local minima, and relative temporal instant of the global minimum value. Such an approach makes it possible to characterize the vertical motion of the gesture sequence. The 5 features computed for the velocity and acceleration signals, and for both hands and for the center of the hip, lead to a 30 features representation.

Finally, we propose 24 features for describing the *Shape* quality. A first part of them aims at globally characterizing the spatial dissymmetry between the two hands over the whole sequence. For each frame, we associate a dissymmetry measure defined as described by the following equation:

$$\text{Dys} = \frac{d_{\text{left,center}}}{d_{\text{left,center}} + d_{\text{right,center}}}, \tag{5}$$

where $d_{\text{left/right,center}}$ denotes the distance between the left/right hand and the center of the shoulders.

The Dys measure takes values within the [0, 1] interval. For a perfectly symmetric gesture, Dys equals 0.5. The Dys_t sequence is globally described by 6 parameters: mean, standard deviation, global maximum/mean ratio, global minimum/mean ratio, number of local extrema, and the relative temporal position of the global extremum. In addition, the same parameters are associated with 3 other sequences, respectively, associated with global body amplitudes in the directions perpendicular to vertical, horizontal, and sagittal planes (Fig. 4), respectively, denoted by A^x, A^y, and A^z and defined by the following equations:

$$A_t^x = \left(\max_i \left(\left|x_{i,t}^{\text{trans}}\right|\right) - \min_i \left(\left|x_{i,t}^{\text{trans}}\right|\right)\right)_{t=0}^{N-1}, \tag{6}$$

$$A_t^y = \left(\max_i \left(\left|y_{i,t}^{\text{trans}}\right|\right) - \min_i \left(\left|y_{i,t}^{\text{trans}}\right|\right)\right)_{t=0}^{N-1}, \tag{7}$$

$$A_t^z = \left(\max_i \left(\left|z_{i,t}^{\text{trans}}\right|\right) - \min_i \left(\left|z_{i,t}^{\text{trans}}\right|\right)\right)_{t=0}^{N-1}, , \tag{8}$$

where i indexes the skeleton joints.

The above-described approach leads to a total number of 81 features for describing gestures. Let us now investigate how such features can be exploited for gesture recognition purposes.

4 Experiment I: gesture recognition

4.1 Microsoft Research Cambridge-12 database

A first experiment (Experiment I) consisted of testing our expressive model on specific actions. For this purpose, we used the Microsoft Research Cambridge-12 (MSRC-12)

Fig. 6 Examples of motions tracked for MSRC-12 datasets constitution

Table 1 Statistical parameters based upon the values taken by 4 Laban features for 3 different actions of MSRC-12 dataset.

Action	Start music/raise volume (A1)	Take a bow to end music session (A2)	Change weapon (A3)
F1			
Mean	0.498	0.515	0.555
SD	0.008	0.045	0.039
F2			
Mean	0.953	0.675	0.976
SD	0.027	0.101	0.027
F3			
Mean	0.048	0.157	0.089
SD	0.067	0.084	0,050
F3			
Mean	0.599	0.448	0.323
SD	0.190	0.249	0.108

dataset [43] (Fig. 6) that is publicly available and presents the advantage to provide two different types of gestures captured by a Kinect camera. The data set includes 6 categories of iconic gestures (which basically represent actions/objects: crouch or hide, shoot a pistol, throw an object, change weapon, kick and put on night vision goggles), and 6 metaphoric ones (more related to higher level, abstract concepts: start music/raise volume, navigate to next menu, wind up the music, take a bow to end music session, protest the music, and move up the tempo of the song). These two categories relate to McNeill's gestures taxonomy [1]. As proposed in [44], we have first performed the recognition separately on gestures of each given type (iconic, metaphoric). Then, we have considered globally all the 12 gesture categories, in order to test for scalability.

Table 1 briefly illustrates the values taken by 4 features of our gesture descriptors and presents these values for 3 actions of MSRC-12 dataset: start music/raise volume (A1), take a bow to end music session (A2) and change weapon (A3).

These 4 features (among 81) derived from our expressive model are

- F1: the mean of body dissymmetry sequence (quantifying the shaping subcomponent of shape quality),
- F2: the mean of the sequence corresponding to the body amplitude in the direction perpendicular to the horizontal plane, describing how the body changes its shape by rising/sinking (quantifying the shaping subcomponent of shape quality),
- F3: the amplitude of the head movement in the direction perpendicular to vertical plane, measuring forward/backward global body motion (quantifying the space quality),
- F4: the mean duration of high activity motion subsequences, so called "motion bells" (quantifying the time quality).

For each one of the 3 actions retained, we consider the different performances of this action in the database, and based upon the corresponding subset of gestures, we compute the mean and standard deviation of the values taken by our 4 expressive features.

In the case of F1 feature, the mean and scattering values computed for the 3 actions show that this feature allows easy distinctions between A2 and A3 occurrences or between A1 and A3 occurrences, but confusion risks remain possible between A1 and A2 actions. When it comes to F3, similar considerations prove that the only distinction problem can occur between A1 and A3 actions. Discrimination based upon F2 feature will make it difficult to distinguish between A1 and A3 actions, because mean value and standard deviation are very close in the case of these two actions. On the contrary, the high distance with A2 mean value should make it easy to recognize A2 action among occurrences of A1 and A3 actions. Finally, a study of mean and scattering of values taken by F4 feature shows that this last characteristic can only help to make clear distinctions between A1 and A3 actions.

These examples based on for 4 expressive features show that the use of combinations of several expressive features will help recognizing different actions present in the dataset.

4.2 Classification methods

We compared the performance of two different methods. The first one uses support vector classifiers (SVC) [45] with the one versus one strategy [46]. This multiclass learning strategy trains a classifier for each pair of classes. At the testing stage, the class collecting the highest score, i.e. the highest number of classifier votes, is retained.

The second classification method is the extremely randomized trees (extra trees) [47,48]. This classifier consists in several tree classifiers trained independently. At each tree node, a component of the feature vector and a threshold are randomly chosen at each tree node. Thus training data are split between the left and right child nodes. When all leaves contain only one training sample the class is associated and the process is stopped. At prediction time, the feature vector is processed by all the trees, and the class collecting the highest number of votes is retained.

For the both classification methods, we have used the "scikit-learn" [49] python toolbox implementation.

4.3 Evaluation protocol

We performed three different recognition runs (cf. Sect. 4.1): one only with iconic gestures, another only with metaphoric gestures, and a last with both types together.

We have applied a 5-fold cross-validation scheme, with a training/testing ratio of 80/20 % and 5 cross-validation steps. The cross-validation has been achieved by splitting the data into 5 blocks preserving the initial class distribution.

We use the F-score [50] as performance measure, defined as

$$F - \text{score} = 2\frac{\text{precision} \,.\, \text{recall}}{\text{precision} + \text{recall}}, \tag{9}$$

where precision refers to the percentage of samples truly positive among all samples classified as positive, and recall to the percentage of samples classified as positive among all samples effectively positive.

The results are compared to the ones reported in [44] for balanced distributions, Thus, for each classification strategy (SVC and extra trees), we report the mean F-scores obtained on each gesture class involved.

4.4 Experimental results

The classification results are presented for each recognition run (with only iconic, only metaphoric, and global strategies) and by class in Figs. 7, 8 and 9. The mean F-scores obtained on all the gesture classes involved are presented in Fig. 10. These results correspond to the best obtained among different parameters combinations associated with each classification strategy (SVC or extra trees) and obtained through an optimization strategy.

The different SVC parameters which needed to be optimized are listed below with the different tested values:

- penalty parameter C of the error term: 10^{-2}, 10^{0}, 10^{2}, 10^{3};
- kernel type: polynomial, Gaussian ("rbf"), linear, sigmoid;
- degree of kernel function: 10^{1}, 10^{2}, 10^{3}, 10^{4};
- kernel coefficient (when necessary) gamma: 10^{-5}, 10^{-2}, 10^{1}.

Likewise, the different extra trees parameters and optional values are as follows:

- the number of trees of the forest: 10^{1}, 10^{2}, 10^{3};
- the criterion function measuring the quality of a split: entropy, gini impurity;
- the number of features among N descriptor values to consider when looking for the best split: N ("None"), $\sqrt{N}$ ("sqrt" or "auto"), $\log_2 N$ ("log2").

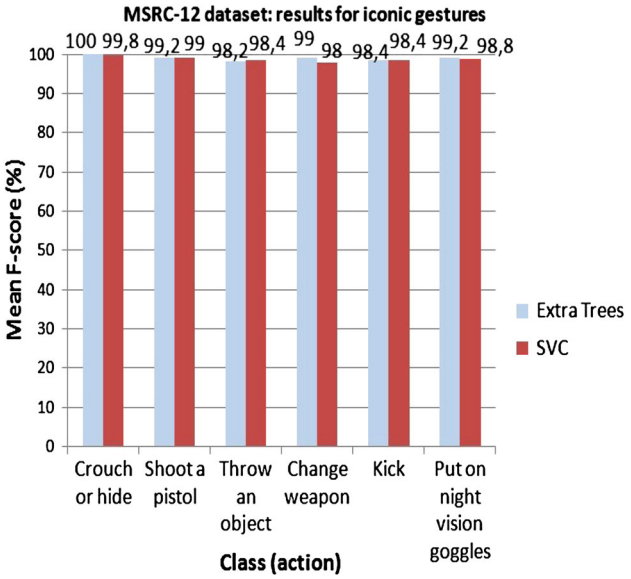


Fig. 7 F-scores (in %) obtained by class for the various classification methods retained in Experiment I and only with iconic gestures

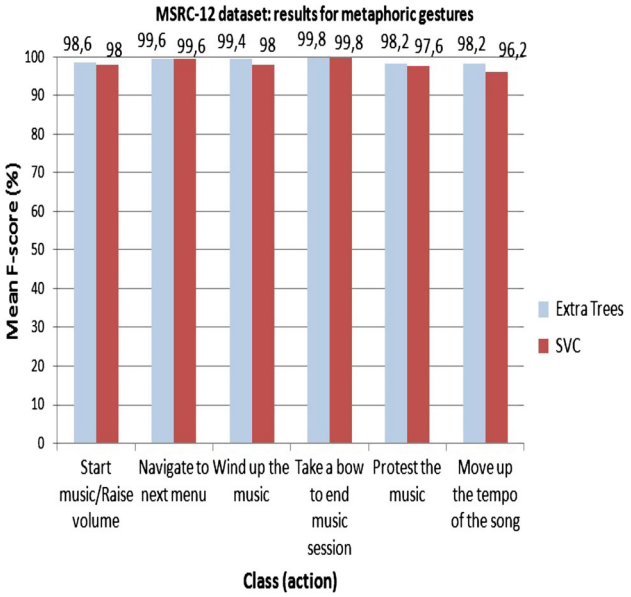


Fig. 8 F-scores (in %) obtained by class for the various classification methods retained in Experiment I and only with metaphoric gestures

The optimal parameters combinations obtained are summarized in Table 2 (for SVC) and Table 3 (for Extra Trees).

For more detailed descriptions of the various parameters involved, please refer to [49].

The mean F-measures obtained are in all cases superior to 97 %, whatever the classification strategy involved.

The Extra Trees performs slightly better than SVM with one-to-one strategy, with a gain in F-score of about 1 %.

The recognition rates are slightly superior for the iconic gestures than for the metaphoric ones SVC using.

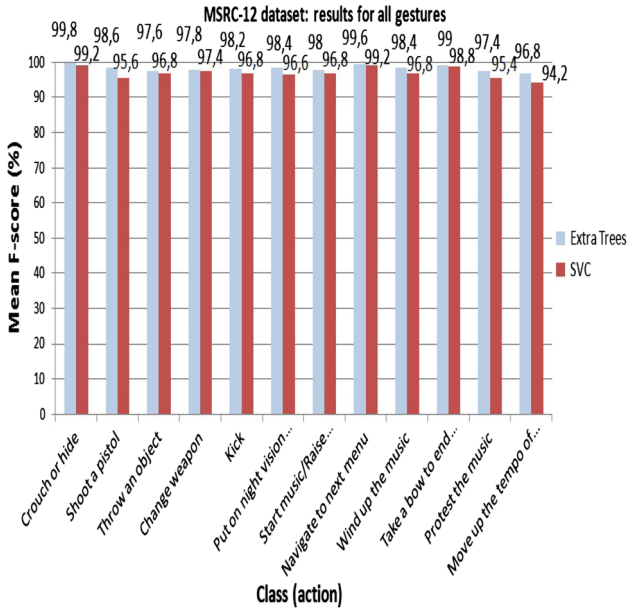


Fig. 9 F-scores (in %) obtained by class for the various classification methods retained in Experiment I and with all gestures

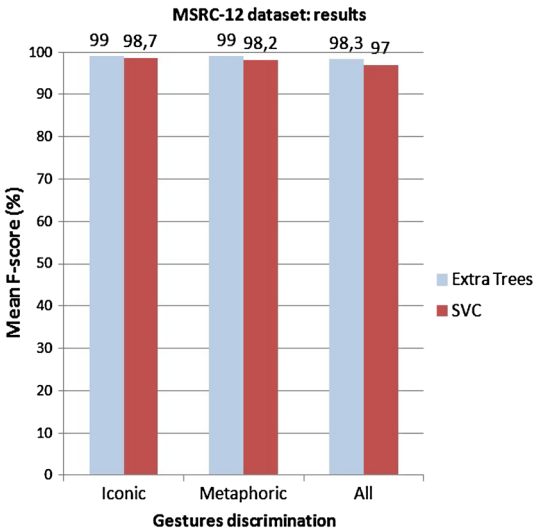


Fig. 10 Mean F-scores (in %) obtained for the various classification methods in Experiment I

When both iconic and metaphoric gestures are considered all together, the performances slightly degrade (1.4 and 0.7 % respectively for SVM and extra trees approaches), but still the F-score is greater than 97 %.

In any case, the proposed method outperforms the ones introduced in [44], where the best F-scores obtained separately with iconic and metaphoric gestures are, respectively, of 95 and 81 %. This demonstrates the pertinence of the proposed approach and the capacity of the Laban descriptors to efficiently capture the salient features of both iconic and metaphoric gestures.

Table 2 Best parameter combinations for SVC according to gestures discrimination in Experiment I (MSRC-12)

Gestures	Iconic	Metaphoric	All
C	0.01	0.01	1000
Kernel	"poly"	"poly"	"poly"
Degree	4	3	3
Gamma	10.0	10.0	0.01

Table 3 Best parameter combinations for Extra Trees according to gestures discrimination in Experiment I (MSRC-12).

Gestures	Iconic	Metaphoric	All
n_estimators	1,000	1,000	1,000
Criterion	"entropy"	"entropy"	"gini"
max_features	"auto"	"auto"	"None"

5 Experiment II: emotion recognition

In order to study the relevance of our model of expressive gesture, we also tested our descriptors on a corpus composed of pre-segmented orchestra conductors' gestures annotated with emotional discrete categories. In the following section, we present the stakes underlying the composition of such a corpus as well as its annotation protocol.

5.1 Orchestra conductor gestures database

Among the areas interested in gestures analysis, music is a field where gesture's expressivity is decisive. As explained in [51,52], the presence of motion at the origin of any sound implies that there is a relationship between sound and motion which is crucial for musical expressivity. For several decades, contemporary composers and artists have been getting inspired by these considerations on sound and motion. Some of them gave birth to "temporal semiotic units", introduced in [53], which are classes of sound units supposed to present common acoustic properties and embody semiotic contents independently on the context in which they appear. The description of these classes borrows its vocabulary and metaphors from the motion analysis field ("turning", "stationary", "suspended", "advancing"…), which confirms that musical impressions are based upon motion suggestions.

Musical performances put at stake various types of motion suggestions: breath, gestural preparation of sound emission, empathic movements, metaphorical references, and anticipations… Such suggestions are crucial, irrespective of whether they concern relationships between the performers and the audience or communication between the artists.

Orchestra conductors occupy a very specific position with regard to this relationship, since their objective is to embody

the music solely with the help of gestures, expressing physically and metaphorically what they anticipate from the inside.

Here, we concentrate on the emotional content of orchestra conductors' gestures. Conductors confirmed that thanks to a certain amount of confidence in their musicians, they continually forget the material production of the sound in order to embody the idea/sentiment that they want to transmit. We focus on intentional motions and put aside any consideration related to individual's psychology or mental states [54].

We made the hypothesis that there is a basis of common representations allowing these embodiments to be understood by musicians and that this inter-subjectivity is based upon minimal emotional representations.

Following the advice of orchestra leaders and musicologists, we built a lexicon of emotional categories inspired by both common musical feelings and universal emotions. Our approach is similar to those of Grewe et al. [55] who had asked participants to listen classical music excerpts and characterize them with emotions. It resulted in 17 classes: calm, agitated, dynamic, tense, serene, magical, mysterious, happy, sad, easy, surprising, troubled, melancholic, disquieting, wrathful, tragic, and inexpressive.

It has to be noted that during our discussions with professional and semi-professional orchestra conductors, some of them listed the features they estimated to be pertinent in terms of expressive characterization of their own movement. We have observed that several of these features match perfectly some of the Laban descriptors that we propose. Notably, they relate to symmetry or dissymmetry of the body, spatial extent, weight, directness of the motion, bending over, quantity of motion forward, and dynamics of the non-dominant hand (usually related to the expressive character of the music, whereas the other hand beats the tempo)…

We have recorded 8 different rehearsals with depth cameras (e.g., Kinect), which provided us gestures of individuals leading a lot of different music styles (classical, movie music, jazz…). Each session gave birth to several excerpts which we decided to segment manually to create gesture samples where conductors' motion and attitude can be considered as a whole and where the intentional communication with musicians is clearly interpretable and susceptible to be qualified by emotions. We have thus obtained a number of 892 gesture sequences of various durations (2–20 s).

Then, we asked professional or semi-professional musicians to annotate each gesture's segment by choosing emotions among the categories proposed in the emotional lexicon retained. Each participant was given the possibility to watch every gestural segment as much as he wanted and was asked to choose up to 3 emotions of the lexicon to characterize conductors' expressive intention. A number of 15 volunteers participated in this annotation task.

We did not give them any other instruction; we let them define their own strategy for "composing" their description with the emotional categories of the vocabulary. In other words, we did not take into account the personal handling of the protocol or the individual use of the words. Thus, some of the participants certainly preferred to "equally" consider the 17 emotions to determine the ones that best described the gestures in question. On the contrary, certain annotators probably focused on a small number of general categories (calm, dynamic…) to globally describe gestures intentionality and then used more specific words (melancholic, disquieting…) to complete and deepen their description. We put aside the study of such possible annotation strategies to concentrate on the annotation results.

Eventually, we have taken into account the fact that the participants would probably handle the emotional lexicon while they were annotating the corpus. In other words, we have guessed that the contribution of each word of the lexicon to the modeling of conductors' expressive intentionality would not be obvious to the annotator from the beginning, and would consequently vary in function of annotator's perception of corpus variability. Thus, we let the participants the possibility to change all their choices until the end of their annotation, to let them check the consistency of their own lexicon use.

Figure 11 illustrates an example of online personal space designed for the corpus annotation. The categories previously

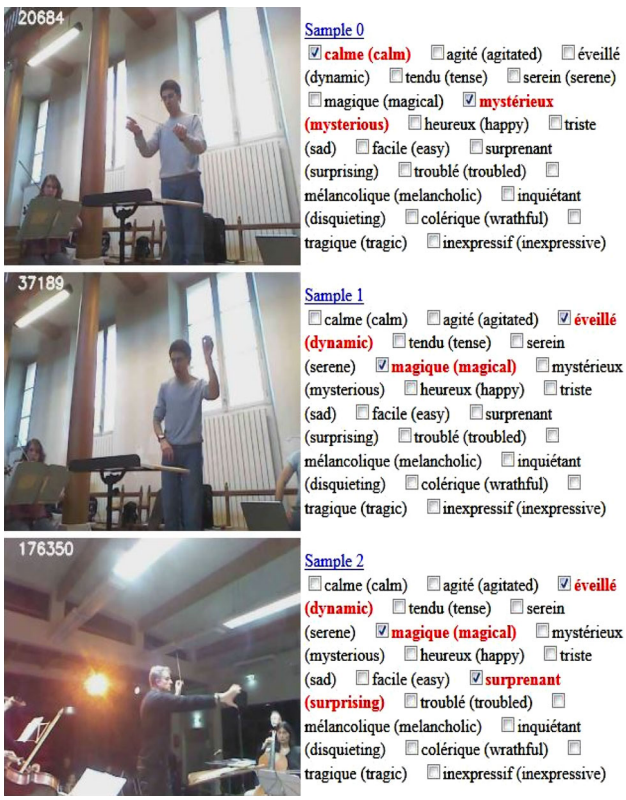


Fig. 11 Visualization of our annotation space

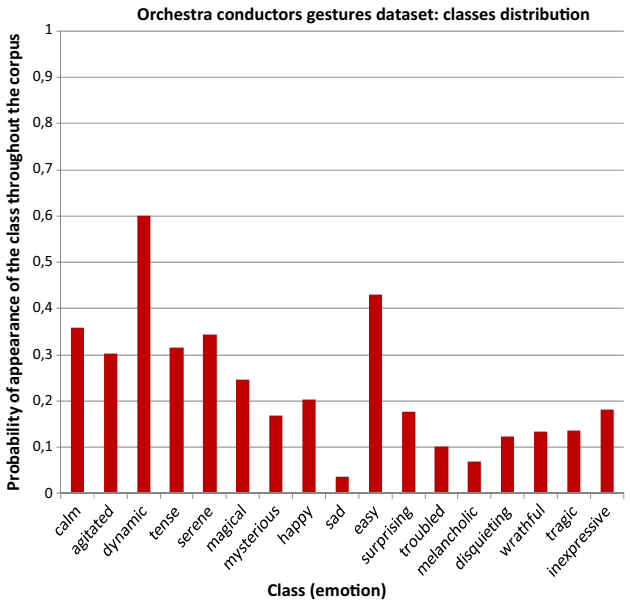


Fig. 12 Emotional category distribution all over the orchestra conductors gestures dataset

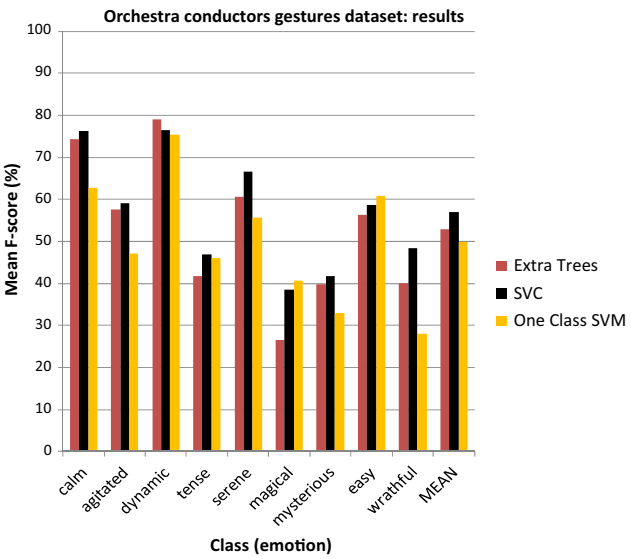


Fig. 13 F-scores (in %) obtained by class for the various classification methods retained in Experiment II: illustration

selected by the participant were displayed in red as a recall. Check boxes allowed the participant to select or to deselect each category for gesture's description..

The contribution of all the participants to the annotation resulted in emotion histograms for each gestural segment, where each histogram class represents an emotion.

Finally, for each gesture, we retained the 3 categories most chosen by the ensemble of the participants and kept them as ground truth. The global distribution of this "truncated" annotation is presented in Fig. 12. Here, each value corresponds to the probability of the corresponding emotional class to be kept among the 3 categories most chosen by the ensemble of the participants. Let us observe that certain categories are sub-represented and thus cannot be object of statistical studies. Among these ones, we only kept mysterious and wrathful, because the variability in their potential expressions interested use, and decided to eliminate from the classes list the other categories whose probability of appearance throughout the corpus is inferior to 0.2. We also put aside happy class because it would not have been relevant to keep it in the absence of sad class. We have finally retained the following 9 categories for our experiment: calm, agitated, dynamic, tense, serene, magical, mysterious, easy, and wrathful. This process resulted in a corpus of 882 annotated gestures that we have further considered for emotion recognition experiments.

5.2 Classification methods and evaluation protocol

Contrary to the first experiment, we have used classification strategies considering each class (emotion) independently of the others, to deal with the multi-labeling problem on a relatively reduced data set (882 gesture sequences).

We compared the performance of three different classification methods, including SVC, extra trees, and One Class SVM [56]. For the SVM and extra trees approaches, a classifier has been constructed for each of the 9 classes independently. The implementation of these classification methods were also provided by the "scikit-learn" python toolbox.

We used the same 5-Fold cross-validation scheme and the F-score [50] measure as in the first experiment.

5.3 Experimental results

The results obtained (F-scores) are summarized in Fig. 13 and Table 4.

As in Experiment I, a parameter optimization strategy has also been applied here. We have already introduced the parameters to be optimized for Extra Trees and SVC (cf. Sect. 4.4). For One Class SVM, the parameters and respective optional values are the same as SVC ones, but also comprise a "nu" parameter, corresponding to an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors which takes the following values: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}.

The optimized parameters are presented in Tables 5, 6 and 7, for extra trees, SVC and One Class SVM, respectively. A per-class optimization strategy has been considered here.

The results obtained show that the SVC approach is superior to the other methods with a mean best F-score equal to 56.9 %, against 52.8 % for Extra Trees and solely 49.9 % for One Class SVM. SVC method gives better results on all classes except for dynamic, magical and easy categories.

Table 4 F-scores (in %) obtained by class for the various classification methods retained in Experiment II.

Method	Extra trees	SVC	One class SVM
Calm	74.3	76.2	62.7
Agitated	57.6	59.1	47.1
Dynamic	79.1	76.5	75.3
Tense	41.8	46.8	46.0
Serene	60.6	66.7	55.7
Magical	26.5	38.6	40.6
Mysterious	39.8	41.8	32.9
Easy	56.3	58.7	60.8
Wrathful	40.0	48.5	28.1
MEAN	52.9	57.0	49.9

Table 5 Optimized parameters for extra trees

Parameter	n_estimators	Criterion	max_features
Calm	1,000	"entropy"	"None"
Agitated	100	"gini"	"None"
Dynamic	100	"gini"	"auto"
Tense	1,000	"entropy"	"None"
Serene	1,000	"entropy"	"None"
Magical	10	"entropy"	"None"
Mysterious	1,000	"gini"	"None"
Easy	100	"gini"	"sqrt"
Wrathful	1,000	"entropy"	"None"

Table 6 Optimized parameters for SVC

Parameter	C	Kernel	Degree	Gamma
Calm	1	"linear"	3	0
Agitated	100	"poly"	1	10
Dynamic	100	"poly"	4	0.01
Tense	100	"rbf"	3	0.01
Serene	1,000	"poly"	1	0.01
Magical	100	"rbf"	3	0.01
Mysterious	100	"sigmoid"	3	0.01
Easy	100	"poly"	2	0.01
Wrathful	1	"poly"	1	10

Table 7 Optimized parameters for one class SVM

Parameter	Kernel	Gamma	nu
Calm	"sigmoid"	10	0.1
Agitated	"rbf"	0.00001	0.1
Dynamic	"rbf"	0.01	0.0001
Tense	"rbf"	0.01	0.01
Serene	"sigmoid"	10	0.1
Magical	"rbf"	0.01	0.01
Mysterious	"sigmoid"	10	0.1
Easy	"rbf"	0.01	0.01
Wrathful	"rbf"	0.01	0.0001

Globally, SVC results exceed Extra Trees ones by 4.1 %. The difference on tense, serene, and wrathful categories is at least of 5 %, and reaches more than 12 % for magical class.

The SVC outperforms One Class SVM with a mean F-score difference of about 7 %. On calm, agitated, and serene classes, this difference between SVC and One Class SVM results is greater than 11 %. The biggest gap of 20.4 % is obtained for wrathful class.

When comparing the extra trees and One Class SVM approaches, the difference is more relative, with a global gain of 2.9 % for the extra trees approach. The highest gaps are obtained for calm, agitated, and wrathful classes (with gains of 11.6, 10.5 and 11.9 %, respectively).

However, the One Class SVM results are superior of more than 4 % for tense and easy emotions. The result obtained for magical category exceeds the one obtained with Extra Trees by more than 14.1 %.

The results obtained with One Class SVM are globally the worst, except for tense, magical, and easy categories. For magical, which is the less represented category in the corpus, the One Class SVM is the best classifier.

Extra Trees only give the best result for one class (dynamic), which demonstrates the higher efficiency of SVM methods in classification problems where the number of samples is relatively small.

Globally, the F-score obtained in this experiment for such high-level, emotional gestures are, without surprise, significantly lower than those obtained for gestures representing specific actions such as those considered in Experiment I (cf. Sect. 4), Improvements remain to be done on our gesture model, by adapting it to non pre-segmented gestures and taking into account additional motion characteristics (its repetitive character for instance…). However, the relatively good scores obtained for certain emotional categories such as calm, agitated, dynamic, serene or easy are quite remarkable (with respective F-scores of 76.2, 59.1, 79.1, 66.7, and 60.8 %). Thus, the proposed approach yields recognition rates of the same order as the one reached in other experiments aiming at using expressive features to characterize emotional content of gestures acted especially for this purpose [6,57].

6 Conclusion and perspectives

In this paper, we introduced a gesture description approach, based on a set of descriptors dedicated to the various entities defined in the Laban movement analysis model.

We put aside the problematic related to the definition and tracking of the relevant body parts, using a depth sensor. Thus, we concentrated on the design of intermediary gesture descriptors based upon lower-level features consisting of body joint trajectories provided by the camera.

The underlying model of gesture relates to different abstract notions which all contribute to a global and qualitative description of body motions, inspired from the LMA (Laban Motion Analysis) model. As a result, a 81 values feature vector has been retained as a global gesture descriptor. The components of the feature vector are each dedicated to the quantification of a Laban quality or subquality.

The experimental results, obtained on the MSRC-12 data set, show very high recognition rates (more than 97 % in terms of F-score) whatever the classification strategy retained (SVM or extra trees). In a second experiment, we have attempted to recognize emotions from orchestra conductors' gestures. The obtained results show promising recognition rates, with a global F-score superior to 57 % when an SVC classification approach is used. The experimental results obtained demonstrate the relevance of the proposed approach and notably the interest of considering qualitative descriptions of the body motion for designing a new gesture model. The resulting mid-level Laban features are efficient when it comes to distinguish between different actions or recognize emotions expressions. It confirms that expressivity and intentionality are crucial gestures components.

Our perspectives of future work first concern the improvement of the Laban quality quantifications, as well as the specification of additional qualities. To this purpose, we need to go further into the analysis of Laban qualities, to avoid the semantic overlapping between the expressive concepts that have been introduced in Sect. 3.1 and have to be quantified. For instance, it seems that body quality, dedicated to the coordination between the implied body parts, and shape flow subcomponent of flow quality, dealing with the dynamic evolution of the relationships between the different body parts, share some concepts in common.

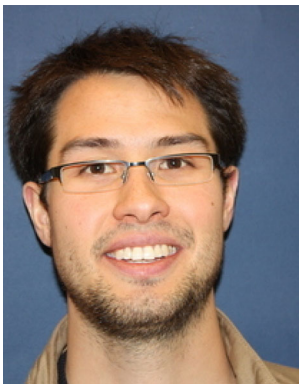
Another perspective deals with the testing of our gesture model on other 3D datasets, notably on corpus where the different occurrences of a given category of gestures show more variations than the ones of MSRC-12.

Finally, we plan to investigate the possibility to redefine our descriptors on spans of variable temporal widths, to achieve real-time, joint segmentation, and recognition of sequences of gestures. Being able to analyze the evolution of embodied emotional profile or recognize successive actions for a given motion period could be useful for the temporal and structural apprehension of gestuality.

References

1. McNeill, D.: Language and gesture. Cambridge University Press, Cambridge (2000)
2. Wang, Y., Huang, K., Tan, T.: Human activity recognition based on r transform. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR'07, IEEE, pp. 1–8 (2007)
3. Huang, F., Xu, G.: Viewpoint insensitive action recognition using envelop shape. In: Computer Vision-ACCV 2007. Springer, Berlin Heidelberg, pp. 477–486 (2007)
4. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011). IEEE, pp. 500–506 (2011)
5. Balomenos, T., et al.: Emotion analysis in man-machine interaction systems. Machine learning for multimodal interaction, pp. 318–328 (2005)
6. Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., Volpe, G.: Multimodal analysis of expressive gesture in music and dance performances. In: Gesture-based communication in human-computer interaction, pp. 20–39 (2004)
7. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. J. Netw. Comput. Appl. **30**(4), 1334–1345 (2007)
8. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
9. Swaminathan, D., et al.: A dynamic bayesian approach to computational laban shape quality analysis. Adv. Hum. Comput. Interact. pp. 1–17 (2009)
10. Zhao, L., Badler, N.I.: Acquiring and validating motion qualities from live limb gestures. Graph. Models **67**(1), 1–16 (2005)
11. Samadani, AA., Burton, S., Gorbet, R., Kulic, D.: Laban effort and shape analysis of affective hand and arm movements. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, pp. 343–348 (2013)
12. Braem, P.B., Bräm, T.: A pilot study of the expressive gestures used by classical orchestra conductors. J. Conduct. Guild **22**(1–2), 14–29 (2001)
13. Glowinski, Donald, et al.: Toward a minimal representation of affective gestures. IEEE Trans. Affect. Comput. **2**(2), 106–118 (2011)
14. Boutet, D.: Une morphologie de la gestualité: structuration articulaire. Cahiers de linguistique analogique, vol. 5, pp. 81–115 (2008)
15. Luo, P., Neff, M.: A perceptual study of the relationship between posture and gesture for virtual characters. In: Motion in Games. Springer, Berlin, pp. 254–265 (2012)
16. Laban, R.: La Maîtrise du Mouvement. Actes Sud, Arles (1994)
17. Junejo, I.N., Junejo, K.N., Al Aghbari, Z.: Silhouette-based human action recognition using SAX-Shapes. Vis. Comput. **30**(3), 259–269 (2014)
18. Chen, H.S., Chen, H.T., Chen, Y.W., Lee, S.Y.: Human action recognition using star skeleton. In: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks. ACM, pp. 171–178 (2006)
19. Cimen, G., Ilhan, H., Capin, T., Gurcay, H.: Classification of human motion based on affective state descriptors. Comput. Animat. Virtual Worlds **24**(3–4), 355–363 (2013)
20. Etemad, S.A., Arya, A.: Correlation-optimized time warping for motion. Vis. Comput. (2014)
21. Singh, V.K., Nevatia, R.: Simultaneous tracking and action recognition for single actor human actions. Vis. Comput. **27**(12), 1115–1123 (2011)

22. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. BMVC, pp. 1–10 (2008)
23. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005, IEEE, vol. 2, pp. 1395–1402 (2005)
24. Jiang, X., Zhong, F., Peng, Q., Qin, X.: Online robust action recognition based on a hierarchical model. Vis. Comput. **30**, 1021–1033 (2014)
25. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. Syst. Man Cybern. Part B Cybern. **36**(3), 710–719 (2005)
26. Rapantzikos, K., Avrithis, Y., Kollias, S.: Dense saliency-based spatiotemporal feature points for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, IEEE, pp. 1454–1461 (2009)
27. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004, IEEE, vol. 3, pp. 32–36 (2004)
28. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008, IEEE, pp. 1–8 (2008)
29. Kaâniche, M., Brémont, F.: Recognizing gestures by learning local motion signatures of HOG descriptors. IEEE Trans. Pattern Mach. Intell. (2012)
30. Li, Y., Ye, J., Wang, T., Huang, S.: Augmenting bag-of-words: a robust contextual representation of spatiotemporal interest points for action recognition. Vis. Comput. (2014)
31. Wu, J., Hu, D., Chen, F.: Action recognition by hidden temporal models. Vis. Comput. **30**, 1395–1404 (2013)
32. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1685–1699 (2009)
33. Pedersoli, F., Benini, S., Adami, N., Leonardi, R.: XKin: an open source framework for hand pose and gesture recognition using kinect. Vis. Comput. **30**, 1107–1122 (2014)
34. Bouchard, D., Badler, N.: Semantic segmentation of motion capture using laban movement analysis. In: Intelligent virtual agents. Springer, Berlin Heidelberg, pp. 37–44 (2007)
35. Camurri, A., Lagerlöf, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. Int. J. Hum. Comput. Stud. **59**(1), 213–225 (2003)
36. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image Vision Comput. (2012)
37. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Affective Computing and Intelligent Interaction. Springer, Berlin Heidelberg, pp. 59–70 (2007)
38. Nakata, T., Mori, T., Sato, T.: Analysis of impression of robot bodily expression. J. Robot. Mechatron. **14**(1), 27–36 (2002)
39. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. In: IEEE International Workshop on Robot and Human Interactive Communication. ROMAN 2005, IEEE, pp. 294–299 (2005)
40. Kapadia, M., Chiang, I.K., Thomas, T., Badler, NI., Kider J.T. Jr: Efficient motion retrieval in large motion databases. In: Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. ACM, pp. 19–28 (2013)
41. Laban, R.: Espace Dynamique. Contredanse (2003)
42. Clauser, C.E., McConville, J.T., Young J.W.: Weight, volume and center of mass of segments of the human body. Wright-Patterson Air Force Base, Ohio (AMRL-TR-69-70): ANTIOCH COLL YELLOW SPRINGS OH (1969)
43. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems. ACM, pp. 1737–1746 (2012)
44. Song, Y., Morency, L.P., Davis, R.: Distribution-Sensitive Learning for Imbalanced Datasets. In: 2013 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2013). IEEE (2013)
45. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
46. Milgram, J., Cheriet, M., Sabourin, R.: "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? In: Tenth International Workshop on Frontiers in Handwriting Recognition (2006)
47. Breiman, L.: Random forests. **45**(1), 5–32 (2001)
48. Geurts, Pierre, Ernst, Damien, Wehenkel, Louis: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
49. Scikit-learn. (bibOnline). http://scikit-learn.org/stable/
50. Hripcsak, G., Rothschild, A.S.: Agreement, the f-measure, and reliability in information retrieval. J. Am. Med. Inf. Assoc. **12**(3), 296–298 (2005)
51. Adrien, J.M: Une approche polyvalente: Direction Musicale Dansée, Captation Gestuelle Causale. In: Actes des Journées d'Informatique Musicale. Rennes (2010)
52. Jorland, G., Thirioux, B.: Note sur l'origine de l'empathie. Revue de métaphysique et de morale. février, no. 58, pp. 269–280 (2008)
53. Hautbois, X.: Les Unités Sémiotiques Temporelles : de la sémiotique musicale vers une sémiotique générale du temps dans les arts. In: ICMS 8. Gestes, formes et processus signifiants en musique et sémiotique interarts, Paris, Huitième Congrès International sur la Signification Musicale (2004)
54. Shove, P., Repp, BH.: Musical motion and performance: theoretical and empirical perspectives. In: Rink, J. (ed.) The practice of performance. Cambridge University Press, pp. 55–83 (1995)
55. Grewe, O., Kopiez, R., Altenmüller, E.: L'évaluation des sentiments musicaux: une comparaison entre le modèle circomplexe et les inventaires d'émotions à choix forcé", Musique, langage, émotion. Approche neuro-cognitive, pp. 49–73 (2010)
56. Chen, Y., Zhou, X.S., Huang, T.S.: One-class SVM for learning in image retrieval. In: 2001 International Conference on Image Processing Proceedings IEEE, vol. 1, pp. 34–37 (2001)
57. Kanluan, I., Grimm, M., Kroschel, K.: Audio-visual emotion recognition using an emotion space concept. In: 16th European Signal Processing Conference. Lausanne (2008)



Arthur Truong received an engineering degree in computer science from Telecom SudParis (Evry, France) in 2012. After a first experience in linguistic research for CNRS "Rhapsodie" project, dealing with spoken language/written language interface, he started a thesis in October 2012 to work on affective computing and expressive gestures. Besides, he has been playing music for a long time and received in June 2014 a pre-professional musical studies qualification for clarinet practicing.

Hugo Boujut received his master's degree in computer science from the University of Bordeaux (France) in 2008. In 2009, he joined the Audemat Worldcast System Company and obtained a Ph.D. degree in mathematics and computer science from the University of Bordeaux in 2012. He then joined the ARTEMIS Department at the Institut Mines-Telecom/Telecom SudParis (France) as a postdoctoral researcher. His research interests concern facial feature detection and tracking, virtual character animation, and visual content indexing and coding, including feature extraction, visual saliency modeling, and video quality assessment.

Titus Zaharia received his engineering degree in electronics and master's degree in electronics from University POLITEHNICA (Bucharest, Romania) in 1995 and 1996, respectively. In 2001, he obtained his Ph.D. degree in mathematics and computer science from University Paris V-René Descartes (Paris, France). He then joined the ARTEMIS Department at TELECOM & Management SudParis as a research engineer and became an associate professor in 2002. His research interests concern visual content indexing and coding, feature extraction, image and video segmentation, motion detection and estimation, 2D/3D reconstruction, virtual character modelling and animation, virtual/augmented reality, digital interactive TV, calibration techniques, and color image processing.