Truong and Zaharia *EURASIP Journal on Image and Video Processing* (2017) 2017:52
DOI 10.1186/s13640-017-0202-5

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Laban movement analysis and hidden Markov models for dynamic 3D gesture recognition

Arthur Truong and Titus Zaharia*

Abstract

In this paper, we propose a new approach for body gesture recognition. The body motion features considered quantify a set of Laban Movement Analysis (LMA) concepts. These features are used to build a dictionary of reference poses, obtained with the help of a k-medians clustering technique. Then, a soft assignment method is applied to the gesture sequences to obtain a gesture representation. The assignment results are used as input in a Hidden Markov Models (HMM) scheme for dynamic, real-time gesture recognition purposes. The proposed approach achieves high recognition rates (more than 92% for certain categories of gestures), when tested and evaluated on a corpus including 11 different actions. The high recognition rates obtained on two other datasets (Microsoft Gesture dataset and UTKinect-Human Detection dataset) show the relevance of our method.

Keywords: Laban movement analysis, Gesture recognition, Body motion descriptors, Soft assignment, Hidden Markov model

1 Introduction

Gestures are generally defined as motions of the body that contain meaningful information [1]. Within this framework, action analysis is a highly challenging issue that involves 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.

The emergence of general public, affordable depth cameras (e.g., Kinect) facilitating 3D body tracking can explain the recent growth of interest for gesture analysis [2–7]. In effect, gesture analysis and interpretation is highly useful for numerous applications: e-health, video games, artistic creation, video surveillance, 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 and recognition algorithms. Only a few papers propose effective models of gesture descriptors [8–11] and such models rarely refer to perceptual studies on relevant motion indices for gestures interpretation.

In this paper, which extends our preliminary work presented in [12], we introduce a new approach for dynamic gesture recognition, inspired from the Laban movement analysis (LMA) model [13]. The main contributions proposed are the following: (1) a new set of local descriptors based on LMA; (2) a mid-level representation of such features based on a soft assignment procedure that maps the LMA descriptors onto a reduced set of reference poses, automatically extracted from a learning gesture data set (3) a HMM-based learning approach that exploits the soft assignment representation for dynamic, real-time gesture recognition.

The rest of the paper is organized as follows. Section 2 presents a state of the art review of existing action recognition methods. Section 3 introduces details the methodology proposed in this paper. In Section 3.1, we first detail the local LMA descriptors considered. Then, in Section 3.2, we propose a dynamic gesture recognition approach which exploits a set of LMA features that makes it possible to build a dictionary of reference poses exploited within a HMM recognition framework. Section 4 describes first the evaluation dataset and protocol considered. Then,

* Correspondence: titus.zaharia@telecom-sudparis.eu
ARTEMIS Department, Institut Mines-Telecom, Telecom SudParis, CNRS UMR 8145 - MAP5 et SAMOVAR 9 rue Charles Fourier, 91 011 EVRY Cédex, France

we present and discuss the gesture recognition results obtained. Finally, Section 5 concludes the paper and opens perspectives of future work.

2 Related work

In the last two decades, a tremendous research effort has been dedicated to the field of gestures analysis. Within this context, one of the main challenges concerns the definition of relevant motion features for gesture recognition. As in our case we tackle the issue of gesture recognition from 3D acquired with a Kinect camera, we will mainly focus on methods that are dealing with 3D gestures, either acquired directly with the help of 3D devices or estimated from 2D videos with the help of 2D/3D pose estimation techniques.

A first family of approaches, described in the following section, is based on silhouette/posture/pose extraction and representation methodologies.

2.1 Posture/pose representations

In [14], 3D upper body pose is estimated in a multi-hypothesis Bayesian inference framework following a generative model-based approach. Body pose is used for hands search and left/right hand assignment. Hand poses are used in a SVM framework trained using HOG features. The method is tested on the NATOPS aircraft handling signals database. This dataset exploits the official gesture vocabulary for the U.S. Navy aircraft carrier environment, which defines a variety of body-and-hand signals that that carrier flight deck personnel use to communicate with the US Navy pilots. The FA scores obtained are superior to 89% for the 4 gestures considered.

In [15], Singh and Nevatia propose a combination of Dynamic Bayesian Action Networks (DBAN) with intermediate 2D body parts models for both pose estimation and action recognition tasks. Composite actions are decomposed into sequences of primitives based on the variations between the 3D key poses of consecutive frames. The 3D poses are finally mapped onto 2D parts models. The method is tested on a hand gestures dataset including about 500 gestures, with recognition accuracy results around 85–90% for each action.

In the method introduced in [7], Li et al. use an Action Graph to model the dynamics of body motions based on salient postures. Such a model is described by a set of salient postures ω_m, the set of actions A_l to be analyzed, transition probabilities $a_{i,j,l}$ from one posture to another for each action A_l., transition probabilities of all actions $b_{i,j}$, and observations emission probabilities $p(x/\omega_m)$ for each salient posture ω_m. These last distributions are modeled as Gaussian probabilities, and for each body motion frame, the observation consists of 3D points. These points are computed with the help of 3 orthogonal Cartesian plane projections of a depth map representing the 3D surface of the body pose; onto the 3 plane projections, contours are extracted, sampled, and used to retrieve 3D points. The method is tested on MSR-Action3D dataset [7] and almost all categories are recovered with rates superior to 90%.

The introduced approaches show that silhouette or posture shapes are relevant features for action recognition. In addition, they show that body posture or pose characterization are pertinent mid-level representation gestures structural aspects.

A second family of approaches, based on so-called local patterns, describes the body motion in a more local manner, through a body parts segmentation procedure.

2.2 Local patterns methods

In [3], Jiang et al. introduce a hierarchical model for action recognition based on 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. The features considered here are the joints motion and relative positions. A bag-of-words model is used for dimensional reduction and to each word of the codebook is allocated a weight. At the test stage, the gesture is first assigned to a motion-based group and the appropriate KNN classifier is used to yield the classification label.

The method is tested on UTKinect-HumanDetection dataset [6], with accuracy results close to 97%, and also on MSRC-12 gesture dataset [5] with recognition rates reaching 100% for certain gestures, even though confusions remain present for certain actions (worst rates close to 82 and 86%).

In [4], Hussein et al. propose a method for action recognition based on the covariance matrix for skeleton joints locations over time. Covariance matrices are computed over hierarchical sub-sequences inspired by the idea of spatial pyramid matching. The number of layers for sub-sequences hierarchy layers can be parameterized. The action descriptors proposed are based on such covariance matrices computed at different scales. They are tested on MSR-Action3D [7], MSRC-12 [5] and HDM05-MoCap [16] datasets.

The possibility offered by the tracking of body joints provides new keys for mid-level features construction or pose extraction. Still, for recognition purposes, an appropriate local characterization of the body motions is required, at each frame of a sequence.

Moreover, the greatest part of the body motion features introduced suffer from a lack of expressive characterization, and are often dedicated to visual indices of motion. Thus, they usually fail to take into account the semantic aspects of motion (inter-subjectivity,

expressivity, intentionality) and remain focused on the structural descriptions.

The Laban Movement Analysis (LMA) [13], proposed by choreograph and dancer Rudolf Laban, provides a consistent representation of gestures expressivity (Section 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. Since our work strongly relies on the LMA model, let us analyze how this representation is taken into account in the state of the art.

2.3 Expressivity and style

A first category of approaches consists of inspiring from Laban concepts in order 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 [17], Camurri et al. investigate the possibility to use decision trees in order to classify motions of dancers and musicians in a discrete set of four emotional categories (joy, anger, fear, grief) with the help of mid-level features modeling the motion expressivity. The authors compare the recognition results obtained to those reported in [18] from spectators watching dancers and characterizing emotions expression.

In [9], 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 [19]. 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 [20], 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 by using a standard SVM approach. Only four emotion categories are here considered.

In spite of their considerations on expressivity, the proposed methods are limited to some global energy characterizations of the motion, and would require some unification in a model able to capture the keys of the gesture.

A different family 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 [21], a Bayesian fusion approach is used that fuses body motion features for identifying the shape movement quality from dancer improvisations. In [22], 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 (cf. Section 3.1). Laban's model is also used in [23], 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.

The main inconvenient of such approaches is that they require the help of Laban concepts experts, in order to annotate the corpuses and come to a ground truth.

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 [24], Nakata et al. propose a set of motion descriptors, each one referring to a LMA component, and apply these descriptors to dancing robots gestures annotated with the help of four emotional categories. Factor analysis is used to establish causality between Laban qualities and emotions.

In [25], Hachimura et al. implement similar descriptors. The processed results are compared to specialists' annotation, and the matching occurs only for certain qualities. In [26], expressive features aiming at quantifying Body, Effort and Shape LMA qualities (cf. Section 3.1) are defined locally (i.e., for each frame of a gesture), in order to index various gestural contents with local motion "keys". Such motion keys are used for database querying purposes. For a given motion clip, key motion states are extracted, which represent its most salient properties. Let us finally quote the work of Samadani et al. [27] who inspired from [28] [24] and [25] 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-point Likert scales) and emotions (six 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.

Such works validate quantifications of Laban concepts and show that a mid-level LMA-based representation can be obtained starting from visual descriptors. The main problem is that they have validated the use of LMA on gestural corpuses that have been specifically designed for expressivity analysis. Until then, the use of LMA for characterizing generic contents, like actions or spontaneous emotions, has not been proved. Still, in the above-mentioned approaches, the features are generally computed over the whole gesture, as global descriptors, as we have done in our previous work [29], where we have quantified several Laban qualities to directly exploit the features extracted for gesture recognition in a

machine learning framework, without explicitly determining the underlying Laban components. Statistical parameters (up to second order moments) were computed over the entire numerical series to characterize the whole gesture as a global vector.

In this paper, we consider a different approach which attempts to obtaining a more local description, appropriate for on-the-fly gesture recognition. Our challenge is to design an expressive model of gesture aiming at characterizing each frame of a gestural sequence. This requires re-visiting and extending the previous LMA representation, in order to set up a set of local descriptors, appropriate for on-the fly recognition purposes, The LMA descriptive model proposed is introduced in the following section.

3 Methods

The proposed model of descriptors is detailed in the following section.

3.1 Model of descriptors

Let us first describe the LMA framework retained.

3.1.1 LMA framework

Rudolf Laban was a Hungarian dancer, choreograph and dance theoretician who developed a movement analysis method called Laban Movement Analysis (LMA) [13]. 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 the gesture analysis model, originally designed for dance teaching, permitted its extension to the study of all types of movements. The LMA model includes five major qualities that are Relationship, Body, Space, Effort, and Shape [30].

The *Relationship* component refers to the relationships between individuals and is particularly suited in the case of group performances. Thus, its contribution to the designing of an intermediary gesture model which aims at characterizing various high-level contents does not seem to be useful. The *Body* component deals with body parts usage, coordination, and phrasing of the movement. The *Space* component refers to the place, direction and path of the movement, and is based on the concept of *kinesphere* including the body throughout its movement. 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 four elements:

- *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 free or constrained,
- *Weight*, to distinguish between heavy and light movements.

The *Shape* description is decomposed into three subcomponents:

- *Shape flow*: describes the dynamic evolution of the relationships between the different body parts,
- *Directional movement*: describes the direction of the movement toward a particular point,
- *Shaping:* 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.

Each gesture frame t is described by a vector of P components:

$$v(t) = \big(v_1(t), v_2(t), v_3(t), ..., v_p(t)\big), \tag{1}$$

where each component $v_i(t)$ is dedicated to one Laban quality or sub-quality.

Such a local characterization of body motions may satisfy the context of real-time classification, where the data are processed dynamically, on-the-fly before the end of the entire gesture and without any pre-segmentation.

Let us now detail the local descriptors proposed.

3.1.2 Local 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, corresponding to the following body parts: *Center of the Hip*, *Spine*, *Center of the Shoulders*, *Head*, *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* (Fig. 1).

Each body joint trajectory i is represented as a sequence of $\left\{P_{i,t} = \left(x_{i,t},,y_{i,t},,z_{i,t}\right)\right\}_{t=0}^{N-1}$ coordinates in a 3D Cartesian system of coordinates ($Oxyz$) where N

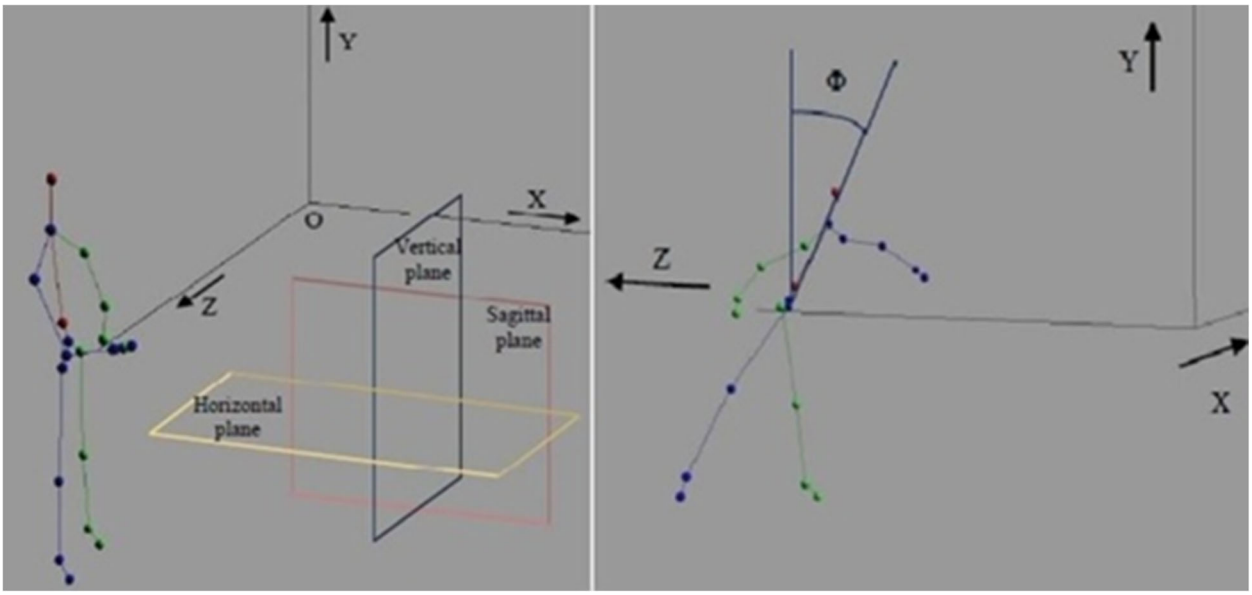

Fig. 1 Body skeleton joints new positions after transforms application and body planes representation (**a**); illustration of forward tilt angle (**b**)

denotes the number of frames of the sequence. As a pre-processing, normalization step, several elementary transforms are applied to the body at each frame of its trajectory before the computation of descriptors. The objective is to set each body joint i in a new position $P_{i,t}^{trans} = \left(x_{i,t}^{trans}, y_{i,t}^{trans}, z_{i,t}^{trans}\right)$ at frame t so that the (xOy), (yOz) and (zOx) planes (Fig. 1), 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 (yOz) 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 in order to set the hip center at the origin of the landmark.
- Secondly, we apply a rotation around the y axis to the body in order to set left and right shoulders in a plane parallel to (yOz) plane.
- Then, we perform a rotation around the z axis to the body in order to set shoulder and hip centers in a plane parallel to (yOz) plane.
- A final rotation around the x axis consists of setting left and right shoulders in a plane parallel to (zOx) plane.
- Finally, we translate the body in order to put the hip center at its initial position.

Once these elementary transforms applied at each frame t, the feature vector $v(t)$ *is* computed. Its various components are introduced in the following.

The *Space* quality is described with the help of two values.

The first one, defined as the x component of the head position $x_{Head,t}^{trans}$, characterizes the head forward-backward motion. The second value is the forward tilt angle $\Phi(t)$ defined as the angle between the vertical direction y and the axis binding the center of the hip and the head, expressed in radians (Fig. 1).

The *Flow* sub-component of the *Effort* quality is described with the help of the third order derivative modules of the left and right hands trajectories, so-called jerk.

For the *Weight* sub-component of *Effort* quality, we consider the vertical components of the velocity and acceleration sequences (i.e., $y'_{.,t}$ and $y''_{.,t}$ signals) associated to 3 joints: the center of the hip, the left and the right hand. These six new values describe the vertical motion of the gesture.

The Shape flow sub-component of *Shape* is described by an index characterizing the contraction of the body, as defined in Eq. (2):

$$C(t) = \frac{\left(\left\|P_{Hip\ Center,t} - P_{Left\ Hand,t}\right\| + \left\|P_{Hip\ Center,t} - P_{Right\ Hand,t}\right\|\right)}{2}, \tag{2}$$

and has been inspired by the contraction index suggested in [18].

Shaping sub-quality of *Shape* is quantified by three values corresponding to the amplitudes in the directions perpendicular to vertical, horizontal and sagittal planes

(Fig. 1), respectively, denoted by A_t^x, A_t^y, and A_t^z and defined by the following equations:

$$A_t^x = \left(max_i\left(\left\{x_{(i,t)}^t rans\right\}\right) - min_i\left(\left\{x_{(i,t)}^t rans\right\}\right)\right), \tag{3}$$

$$A_t^y = \left(max_i\left(\left\{y_{(i,t)}^t rans\right\}\right) - min_i\left(\left\{y_{(i,t)}^t rans\right\}\right)\right), \tag{4}$$

$$A_t^z = \left(\max\left\{_i\left(z_{(i,t)}^t rans\right)\right\} - min_i\left(\left\{z_{(i,t)}^t rans\right\}\right)\right), \tag{5}$$

where i indexes the skeleton joints.

Finally, the *Body* component is quantified with the help of three features.

The first one is an index characterizing the spatial dissymmetry between the two hands and has been inspired by the symmetry index proposed in [9]. This dissymmetry index is defined as described by the following equation:

$$Dys(t) = \frac{d_{left,center}(t)}{d_{left,center}(t) + d_{right,center}(t)}, \tag{6}$$

where $d_{left/right\,,\,center}$(t) denotes the distance between the left/right hand and its projection onto the trunk (e.g., axis binding the center of the hip 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 second and third values are respectively the distance between left hand and left shoulder and the distance between right hand and right shoulder. These parameters are used as a characterization of *Body* quality in [28]:

$$D_t^g = \left\|P_{Left\ Shoulder} - P_{Left\ Hand}\right\|, \tag{7}$$

$$D_t^d = \left\|P_{Right\ Shoulder} - P_{Right\ Hand}\right\|, \tag{8}$$

The above-described approach leads to a total number of $P = 17$ features for describing a gesture at each frame t.

Let us now investigate how such features can be used for dynamic gesture analysis.

3.2 Dynamic gesture analysis

The proposed dynamic gesture analysis method is illustrated in Fig. 2, with both off-line (learning) and on-line (classification) stages.

Our descriptors are used for a poses extraction stage consisting of sub-representing each class of gestures on key-words, which means that a gesture is considered as a path through different key states (cf. Section 4.1). The temporal aspects of the body motion remain implicitly present in such keys, because it is present in the content of our LMA descriptors (cf. 1st, 2nd and 3rd order derivatives). Then, we use Hidden Markov Models (HMM)

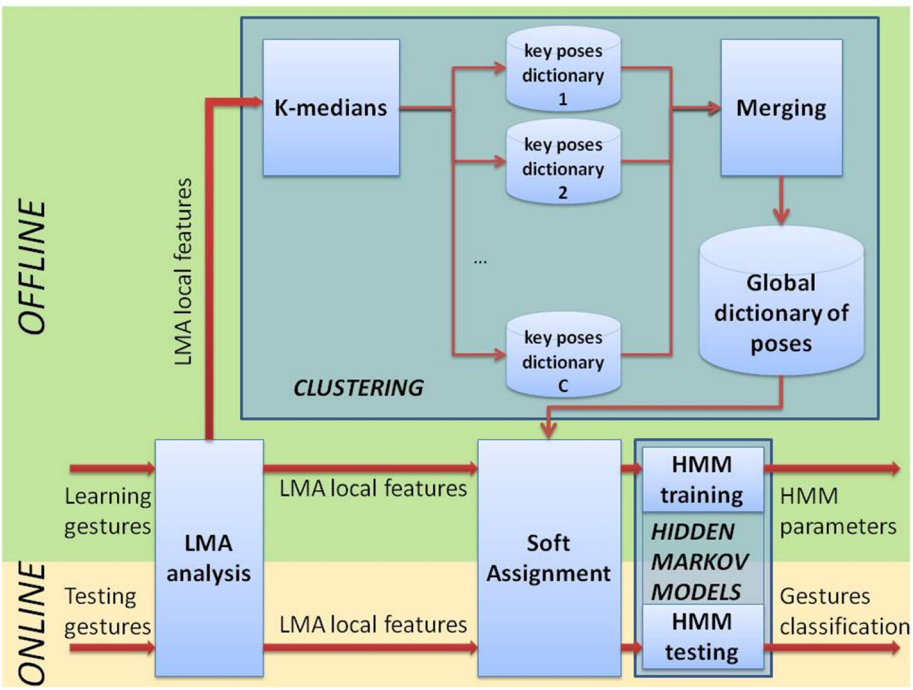

Fig. 2 Overview of the dynamic gesture analysis approach proposed

[31] in a soft assignment based approach (cf. Section 4.2) where each HMM observation consists of a projection on spatiotemporal words (e.g., the key-poses); thus we supposed that these HMM will manage gesture structuration and actions succession (e.g., transitions between HMM states). In our case, a gesture is not interpreted a succession of states, but as a continuous projection of spatiotemporal reference keys.

A first step of our approach consists in determining a set of reference, key poses for the whole set of gestures, as described in the following section.

3.2.1 Reference pose extraction

On the off-line stage, the objective is to determine a lexicon of distinctive reference poses that can conduct to a simplified representation of the body gestures. We suppose a learning database of gestures to be available, categorized into a set of gesture classes (see Section 5.1 for a description of the data set considered in our work).

A set of reference poses is first determined for each gesture category, independently on the other classes.

Let us consider the set of all gesture sequences included in the *G* category:

$$\left(S_1^G, S_2^G, ..., S_{|G|}^G\right), \tag{9}$$

where |G| is the number of gestures in class *G*.

Each instantiation *i* is represented by a series of frame descriptors:

$$S_i^G = \left(v^{G,i}(1), v^{G,i}(2), ..., v^{G,i}\left(T^{G,i}\right)\right), \quad (10)$$

where $T^{G,i}$ denotes the number of frames of the i^{th} gesture sequence.

The key poses are computed over all sequences $\left(S_1^G, S_2^G, ..., S_{|G|}^G\right)$ by using a k-medians clustering algorithm [32] with random initialization. The k-medians algorithm ensures that the reference poses determined correspond to real, existing poses from the training data set.

At each iteration, the relevance of the current clusters is evaluated with the help of the validity measure (*validity*) defined as the ratio between intra and inter class dissimilarity:

$$validity = \frac{intra}{inter}, \tag{11}$$

where *intra* is the intra-class compactness measure:

$$intra = \frac{1}{N_G}\sum\nolimits_{k\in[1,K]}\left(\sum\nolimits_{x\in C_k} d\left(x, \mu_k\right)2\right) \tag{12}$$

and extra is the inter-cluster distance:

$$extra = min_{k\in[1,K-1],l\in[k+1,K]}\left(d\left(\mu_k, \mu_l\right)2\right) \tag{13}$$

In the notations above, N_G is total number of frames in the gesture learning data set for the considered category G, K is the number of clusters, C_k is k^{th} cluster, μ_k its centroid and $d(x_1, x_2)$ is the normalized Euclidean distance [33] between vectors x_1 and x_2 in the LMA descriptors space.

The k-medians algorithm aims at iteratively minimizing the validity measure and stops when its variation between two successive iterations is below a given threshold or when a maximum number of iterations is achieved.

At the end of the clustering process, we obtain a vector of poses $\left(P_1^G, P_2^G, ..., P_K^G\right)$ for each gesture class G. Each P_j^G in the dictionary consists of the skeleton pose parameters and the associated LMA feature vector.

This per-gesture category reference pose calculation strategy offers the advantage of representing each category with a reduced number of distinctive key-poses. However, it is likely to obtain similar key-poses for different categories. In order to eliminate such redundant elements, a final inter-category key-pose merging process is applied. The principle consists of iteratively merging centroid key-poses whose distance is lower than a pre-defined threshold ϱ. This makes it possible to obtain a global reference pose dictionary $\left(P_1^{ref}, P_2^{ref}, ..., P_M^{ref}\right)$, for the whole learning set. Here, *M* denotes the final number of centroids retained. The empirical choice of the threshold ϱ is analyzed and discussed in Sections 4.2 and 4.3.

The availability of a set of reference key-poses makes it possible to obtain a reduced representation of each gesture sequence, able to handle the variability of gestures performed by different individuals. This can be simply achieved by assigning each frame to its closest prototype in the dictionary. However, such a hard classification may suffer from significant vector quantization errors, notably in the case of a reduced number of prototypes. For this reason, we have considered instead a more gradual representation, based on a soft assignment approach.

3.2.2 Soft assignment

The soft assignment method [34] is used to locate a feature vector among a set of prototypes.

For each feature vector v(t) = $(v_1(t), \ldots, v_P(t))$ at frame t, we compute the distance $d_j(t)$ of $v(t)$ to every key pose P_j^{ref} of global dictionary:

$$\forall j\in\{1, ..., M\}, d_j(t) = d\left(v(t), P_j^{ref}\right), \tag{14}$$

Table 1 Gesture categories, number of sequences per class and symbolic labels considered

Gesture category	N_{seq}	L
Say "thank you" in ASL	53	A
Tie shoelaces	57	B
Draw a circle with the right arm	54	C
Rotate on oneself	49	D
Catch an object	48	E
Juggle	51	F
Throw an object in front	49	G
Cover one's ears	53	H
Rub one's eyes	44	I
Kneel	54	J
Stretch out	53	K

where d (.,.) denotes a distance in the feature space. In our case, we have considered simply a L_2 distance between feature vectors.

The soft assignment vector $o(t)$ at frame t is defined as the set of normalized distances:

$$o(t) = \left(d_1^{'}(t), d_2^{'}(t), ..., d_M^{'}(t)\right), \tag{15}$$

where.

$$\forall j \in \{1, ..., M\}, d_j^{'}(t) = \frac{d_j(t)}{\sum_{i \in \{1,...,M\}} \left(d_i(t)\right)}, \tag{16}$$

The soft assignment vector $o(t)$ describes the relative position of the vector in the space drawn by the key poses at frame t.

The following section describes how the vector sequence $o(t)$ can be used as observations within the framework of a HMM (Hidden Markov Model) recognition approach.

3.2.3 Hidden Markov Models framework

In our approach, the gesture categories are used as hidden states in a HMM formulation [31]. The o(t) vectors resulting from the soft assignment stage (Eq. 19) are used as frame observation sequences.

The HMM parameters to be estimated are:

- initial probabilities (π_i , $i \in \{states\}$),
- transitions probabilities from one state to another (a_{ij} , $(i, j) \in \{states\}2$),
- observation emissions probabilities given a state ($b_j(o)$, $j \in \{states\}$).

The various model parameters are globally stored in a vector Λ.

Emission probabilities are modeled as Gaussian distributions whose parameters are the mean vector and covariance matrix ((μ_j, Σ_j) , $j \in \{states\}$).

We consider a number of S gesture sequences in a given learning set. Let us denote by $(\mathcal{O}_1, , \mathcal{O}_2, ..., , \mathcal{O}_S)$ and $(\mathcal{S}_1, , \mathcal{S}_2, ..., , \mathcal{S}_S)$ the observations sequences and the corresponding hidden states series, respectively.

We train the HMM using Baum-Welch algorithm [31] which consists of maximizing the observation expectation:

$$\Lambda^{opt} = \arg max_{\Lambda}\left(\sum_{s \in [1,S]} logP(\mathcal{O}_s|\Lambda\right)), \tag{17}$$

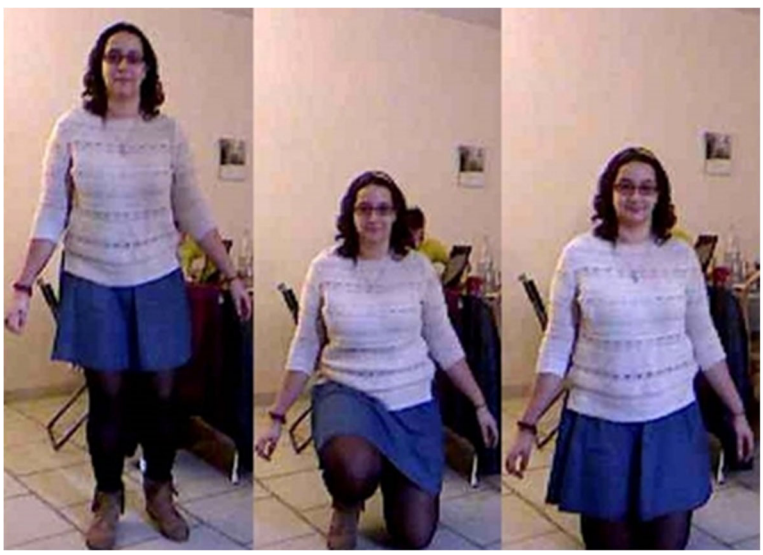

Fig. 3 Execution of kneel gesture by a student

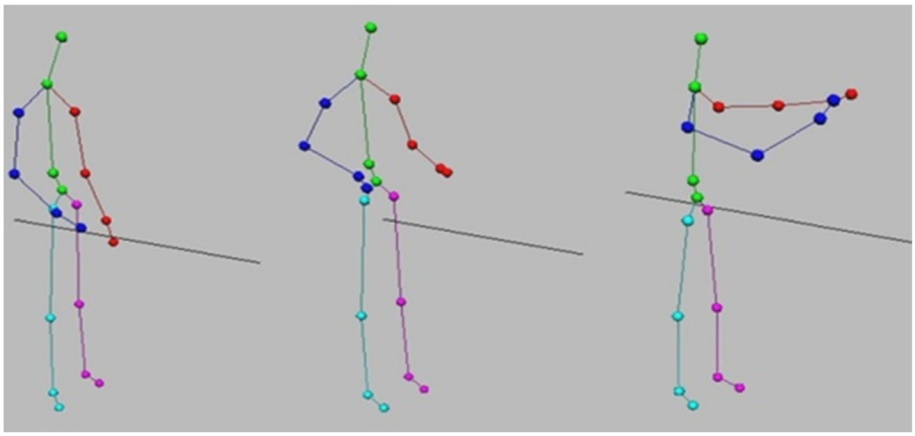

Fig. 4 Examples of key poses retained for the *Catch an object* category

The decoding stage, i.e., determining the most probable state sequence given an observation, is performed with the help of Viterbi decoding procedure [31] which maximizes the a posteriori probability:

$$\mathcal{S}^{opt} = \arg max_{\mathcal{S}}(logP(\mathcal{S}|O^{opt})) \tag{18}$$

Following the recommendations presented in [35], we implemented our own HMM framework by using logarithms of probabilities, in order to handle in a numerically stable manner the extremely small probabilities that appear for relatively long observation sequences.

The HMM initial and transitions probabilities are initialized as recommended in [36]. Concerning the emission distributions parameters, they are computed according to GMM (Gaussian Mixture Models) formula [37]:

$$\mu_j = \frac{\sum_{s\in[1,S]}\sum_{t\in[1,Ts]}(\mathcal{O}_s(t).I(j,t))}{\sum_{s\in[1,S]}\sum_{t\in[1,Ts]}(I(j,t))}, \tag{19}$$

$$\Sigma_j = \frac{\sum_{s\in[1,S]}\sum_{t\in[1,Ts]}\left(\left(\mathcal{O}_s(t)-\mu_j\right)^*\left(\mathcal{O}_s(t)-\mu_j\right)^T.I(j,t)\right)}{\sum_{s\in[1,S]}\sum_{t\in[1,Ts]}(I(j,t))}, \tag{20}$$

where $I(j, t)$ is equal to 1 if the hidden state at time t is j, zero otherwise.

In the following section, we present and analyze the experimental results obtained.

4 Results and discussion

4.1 HTI 2014–2015 dataset

Among the most popular, publicly available 3D gesture databases, let us cite MSR-Action3D [7], and MSRC-12

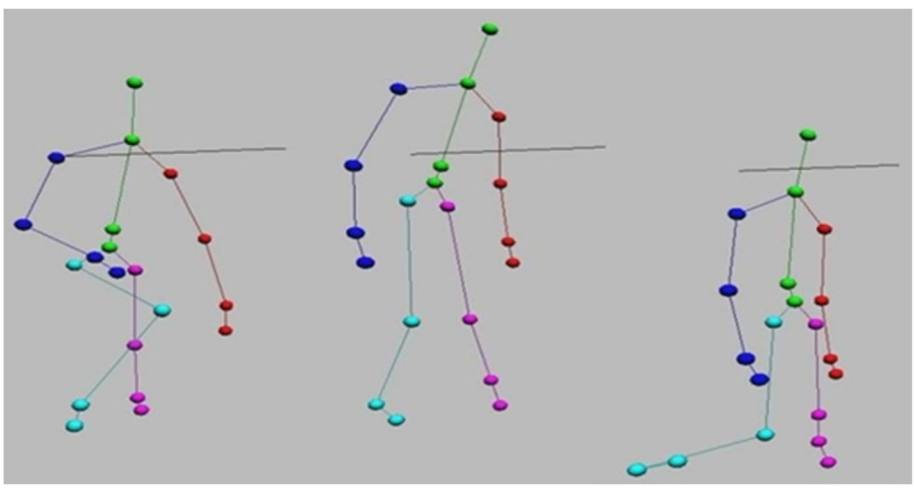

Fig. 5 Examples of key poses retained for the *Kneel* category

Table 2 Global dictionary size in function of threshold ϱ

Threshold ϱ	Global dictionary size M
4.5	20
4.0	31
3.5	40

dataset [5]. However, they include only individual gestures and not sequences of multiple gestures.

The only available 3D corpus proposing successions of individual actions in a single sequence is the UTKinect Human Detection dataset [6]. In addition to its low quality (e.g., 15 fps), the main drawback of this dataset is that it provides identical successions of the same actions (i.e., in the same order). In order to evaluate the ability of our approach to dynamically characterize a gestural content composed of several actions, we have created a new dataset. The gestures were acquired with the help of a Kinect camera, thus providing 3D skeleton joints trajectory sequences.

To our knowledge, our corpus is currently the only one to propose various successions of various actions with a reliable quality (e.g., 30 fps, no occlusions). For this reason, we plan to provide it to the research community.

We asked 11 students from High-Tech Imaging (HTI) major in Télécom SudParis to perform of a set of actions following the instruction given in [5] relatively to the elaboration of MSRC-12 dataset.

We have considered 11 categories of gestures, summarized in Table 1. Figure 3 illustrates the execution of a kneel gesture.

All the students were asked to execute pre-defined sequences of six different gestures selected over the lexicon. This pre-definition provided us an implicit segmentation of multi-gestures sequences into temporal spans of a single action. Between each succession actions, the individuals were supposed to return to a resting state for an undefined period. This constant returning to an artificial resting state has been employed for avoiding the difficulty of gesture pre-segmentation, which is usually required for actions sequences decoding.

All the students were asked to execute pre-defined sequences of six different gestures selected over the lexicon. This pre-definition provided us an implicit segmentation of multi-gestures sequences into temporal spans of a single action. Between each succession actions, the individuals were supposed to return to a resting state for an undefined period.

Table 1 gives the number of segmented gestures per category, with the number of resulting sequences (*Nseq*) and symbolic labels assigned (*L*).

We have finally obtained 107 gestures sequences of various durations (from 16 s to 1 min 7 s), including a total number of 565 individual gestures. The total corpus length is of 48 min 54 s. The individual gestures durations vary from 1 to 15 s.

4.2 Representative gesture clustering

We followed the protocol described in Section 3.2.1 to build class lexicons and gather the retained poses in a global dictionary.

After the observation of the sequences and the variability of class instantiations, we decided to keep the $K = 10$ most representative poses for each class. Figures 4 and 5,

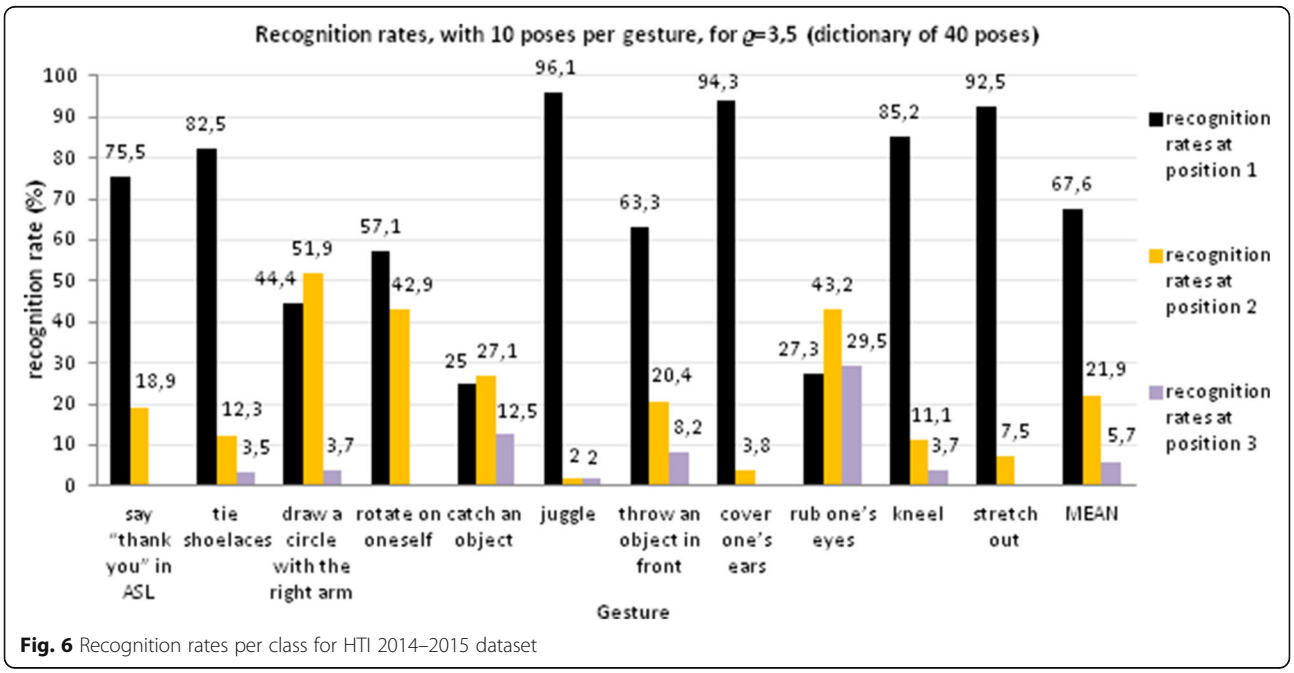

Fig. 6 Recognition rates per class for HTI 2014–2015 dataset

Table 3 Cumulative recognition rates for HTI 2014–2015 dataset

	A	B	C	D	E	F	G	H	I	J	K	Mean
$RR_{cum(1)}$	75.5	82.5	44.4	57.1	25	96.1	63.3	94.3	27.3	85.2	92.5	67.6
$RR_{cum(2)}$	94.4	94.8	96.3	100	52.1	98.1	83.7	98.1	70.5	96.3	100	89.5
$RR_{cum(3)}$	94.4	98.3	100	100	64.6	100	91.9	98.1	100	100	100	95.2

respectively, show different poses retained in the corpus for the *Catch an object* and *Kneel* categories.

Concerning the choice of parameter K, which defines the number of clusters considered in the k-medians algorithm (cf. Section 3.2.1), let us underline that this issue is not critical. Intuitively, the parameter K should be big enough to ensure that all the representative key-poses can be well taken into account. A too small value of K would lead to a risk of missing relevant key-poses. The principle consists of setting K to a value that can ensure a slight over-representation of the key poses. The redundant poses will then be eliminated in the fusion stage (cf. key-pose merging process described in Section 3.2.1). However, a too big value of K would penalize the algorithm in terms of computational burden. In our experiments, taking into account the length (e.g., several hundreds of frames) of the gestures considered in our experiments a value of $K = 10$ proved to be sufficient.

We can observe that the reference poses obtained sample the 3D pose space in a salient manner, while drastically reducing the variability of poses present in a given gesture.

The threshold ϱ (cf. Section 3.2.1) used for merging the per-class dictionaries obtained was determined experimentally. Table 2 gives the obtained dictionary size in function of ϱ.

In the results presented in the following section, we have privileged a value of $\varrho = 3.5$, which corresponds to 40 clusters. This value corresponds to an average number of 3.6 reference poses per gesture category, which seems to be a reasonable guess for the considered corpus. An analysis of the impact of the ϱ parameter on the recognition performances will also be presented in Section 4.3.

4.3 Evaluation results

For the HMM learning procedure, we have applied a fivefold cross-validation scheme, with a training/testing ratio of 80/20% and five cross-validation steps. The cross-validation has been achieved by splitting the data into five blocks preserving the initial class distribution.

At the testing stage, each observation sequence is decoded as a sequence of gesture labels corresponding to the recognized category (Table 1). A distinct label is assigned to each frame, according to the Baum-Welch algorithm (cf. Section 3.2.3). In order to evaluate the recognition performances, for each gesture sequence $\mathcal{O}_{test}$, we identify the 3 most represented classes ordered by their relative frequency of apparition (with respect to the number of frames of the given gesture). Let us denote them by $\mathcal{S}_{\mathcal{O}_{test}}(1), \mathcal{S}_{\mathcal{O}_{test}}(2)$and $\mathcal{S}_{\mathcal{O}_{test}}(3)$. Intuitively, they correspond to the three most probable categories that can be associated to the given sequence.

In order to globalize the information over the entire test database and obtain an objective recognition score for each gesture category G, we compute three recognition rates defined as the percentages of gestures where the *correct* (with respect to the considered ground truth) category has been identified at first, second and third positions. They are, respectively, denoted by RR(1), RR(2) and RR(3) and defined as follows.

$$RR(i) = \frac{\sum_{s=1}^{N_{test}} I_i^s(G)}{\sum_{s=1}^{N_{test}} I^s(G)}, \forall i \in \{1,2,3\}, \tag{21}$$

where:

$$I_i^s(G) = \begin{cases} 1 \text{ if } \mathcal{S}_{\mathcal{O}_s}(i) = G \text{ and } GroundTruth(\mathcal{O}_S) = G \\ 0 \text{ if not} \end{cases}, \tag{22}$$

Table 4 Average recognition rates RR(1)

Threshold ϱ	Average RR(1)
4.5	63.3
4.0	62.2
3.5	67.6

Table 5 Confusion matrix (HTI 2014–2015 dataset)

	A	B	C	D	E	F	G	H	I	J	K
A	*56..6*	0	2.4	2.6	0.1	3.5	33.8	0	0.1	0.9	0
B	21.7	*51.4*	0.4	4.5	0.1	0.1	0.9	0	0.1	20.8	0
C	38.2	0	*40.5*	5.1	1.1	0.4	12.4	0	0	0.2	2.1
D	*42..4*	1.7	5.2	40.8	0.2	1	2.8	0.1	0.2	5.6	0
E	*29..2*	0.1	12.4	3.2	23.7	5	12	0.5	7.9	0.8	5.2
F	20.4	0.3	2.2	2.5	4.8	*60.5*	4.7	0	3.8	0.6	0.2
G	32.3	0.5	6.4	5.6	7.1	2.6	*41.2*	0	0.7	0.7	2.9
H	18.1	0.4	0.2	3.4	1.9	1.4	1.8	*59.9*	11.6	0.5	0.8
I	21.8	0.6	0.4	3.7	1.3	1.8	1.7	*39.6*	28.1	0.8	0.2
J	17.1	14.9	0.6	2.6	1.8	0.5	1.3	0	0	*61.1*	0.1
K	16.4	0.1	1.6	4.3	5	1.2	3.5	3.2	4.3	0.7	*59.7*

In italics: the best recognition scores for each row (category). The (*i*, *j*) entry in this matrix represents the percentage of frames of a sequence in the *i* category that have been recognized as frames from the *j* category

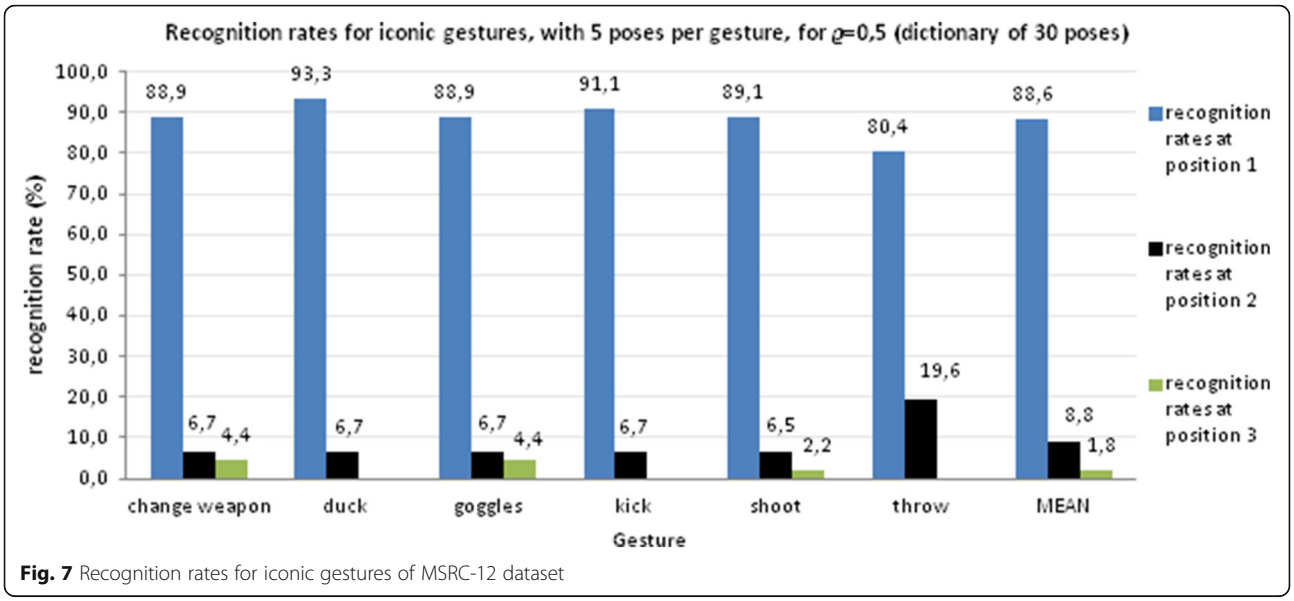

Fig. 7 Recognition rates for iconic gestures of MSRC-12 dataset

$$I^s(G) = \begin{cases} 1 \text{ if } GroundTruth(\mathcal{O}_S) = G \\ 0 \text{ if not} \end{cases}, \quad (23)$$

In the equation above, N_{test} denotes the number of test sequences in the data set and $\text{GroundTruth}(\mathcal{O}_S)$ the class to which the gesture sequence $\mathcal{O}_S$ belongs to.

We also considered the cumulated scores, i.e. rates where the correct category is recognized within the top first, second and third positions, defined as:

$$\forall i \in \{1,2,3\}, RR_{cum} = \sum_{k=1}^{i} RR\,(i), \quad (24)$$

Figure 6 presents the recognition results obtained. Globally, the mean recognition score in first position RR(1), calculated over the entire gesture corpus is of 67.3%. When also considered the cumulative recognition rates $RR_{cum}(2)$ and $RR_{cum}(3)$, which are summarized in Table 3, we can observe the correct category is retrieved in the first two (resp. three) candidates in 89.5% (resp. 95.2%) of cases. Such high recognition scores show the pertinence of the proposed approach.

When analyzing the recognition rates on a per category basis, we can observe that three gestures yield RR(1) recognition rates greater than 92%: *juggle* (F), *cover one's ears* (H) and *stretch out* (K). Very good performances are also obtained for the *Tie shoelaces* (B), *Kneel* (J), gestures, which yield RR(1) recognition rates superior to 82%. For the *Say* "thank you" *in ASL* (A),

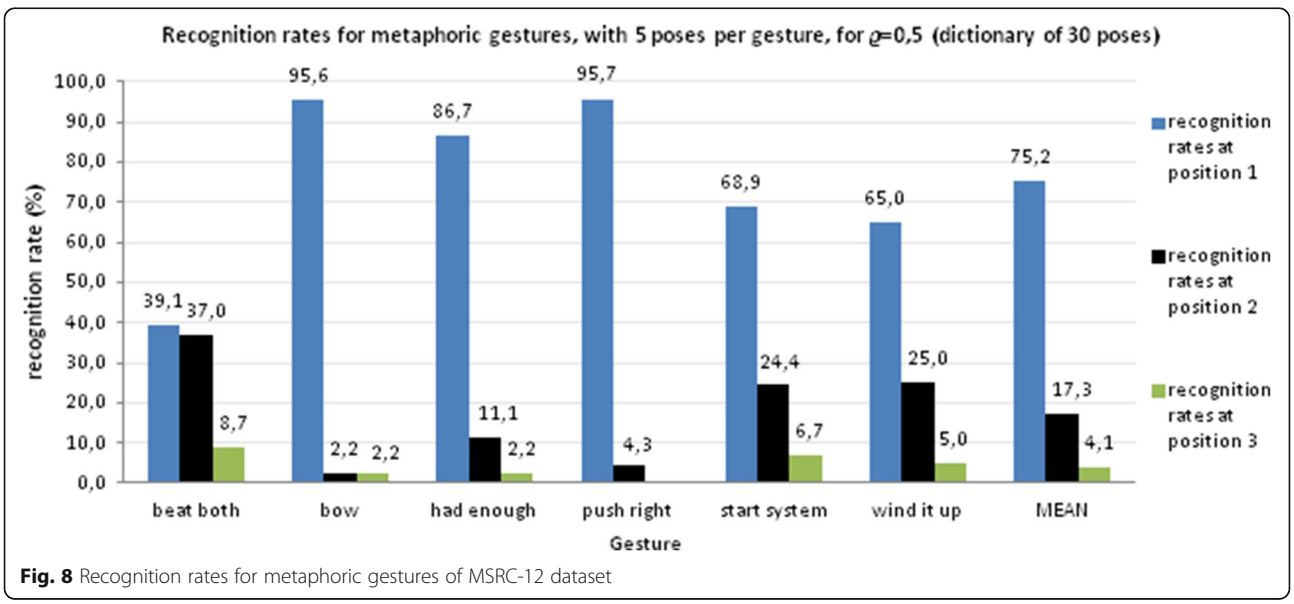

Fig. 8 Recognition rates for metaphoric gestures of MSRC-12 dataset

Table 6 Cumulative recognition rates for iconic gestures of MSRC-12 dataset

	Change weapon	Duck	Goggles	Kick	Shoot	Throw	Mean
$RR_{cum}(1)$	88.9	93.3	88.9	91.1	89.1	80.4	88.6
$RR_{cum}(2)$	95.6	100.0	95.6	97.8	95.7	100.0	97.4
$RR_{cum}(3)$	100.0	100.0	100.0	97.8	97.8	100.0	99.3

Rotate on oneself (D) and *Throw an object in front* (G) gestures, the RR(1) scores slightly lower (75.5, 57.1, and 63.3%, respectively). The lowest RR(1) scores concern the *Draw a circle with the right arm* (C), *Catch an object* (E), and *Rub one's eyes* (I) gestures. However, even in these cases, the correct category is obtained in more than 52.1% of cases in the top two candidates ($RR_{cum}(2)$) and in 64.6% cases in the top three ($RR_{cum}(3)$).

In order to evaluate the influence of the ϱ parameter, we have also computed and compared the recognition rates for values of ϱ of 3.5, 4.0, and 4.5. The average RR(1) rates are summarized in Table 4.

Globally, the results are quite stable, with an increase in recognition rates for $\varrho = 3.5$. This shows that disposing of a greater number of key-poses (40 for $\varrho = 3.5$ and only 20 for $\varrho = 4.5$) can ameliorate the discriminative power of the representation.

In order to further analyze the results, we have also considered per-frame recognition rates. That makes it possible to calculate the confusion matrix C presented in Table 5. Here, the $C(i, j)$ elements represent the percentage of times where a given frame from category i has been classified as category j. Theses scores have been computed over the entire set of frames of the whole corpus.

A relatively strong confusion is made between *say thank you* (A) and *throw an object in front* (G) gestures, which can be caused by the proximity between the involved body motions (notably the arm motion).

A strong confusion also occurs between *rub one's eyes* (I) and *cover one's ears* (H), which can also be explained by body motions similarities (only upper members are involved, and their motions are close in both cases).

The confusion between *draw a circle with the right arm* (C) and *say thank you* (A) is related to the fact that only the right arm is involved in both cases.

Finally, the confusion between *rotate on oneself* (D) and *say thank you* (A) may be explained by the pose normalization process (Section 3.2), since only the relative positions of articulations are taken into account by our model.

Such limitations would need the inclusion in the descriptive model of additional features, able to distinguish in a finer manner between gestures that remain globally similar.

However, the confusion matrix confirms the global recognition results presented in Fig. 6 and shows that in the majority of cases the correct categories are determined.

We also tested our method on other available corpuses. We obtained several results for different values of ϱ parameter, and we present here the best ones obtained.

We first introduce the results obtained on Microsoft Gesture dataset (MSRC-12 dataset, which we had used in our previous work [29]). The data set includes six categories of *iconic* gestures (which basically represent actions/objects: *duck (crouch or hide)*, *shoot [with 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 system (start music/raise volume)*, *push right (navigate to next menu)*, *wind it [the music] up*, *[take a] bow (to end music session)*, *had enough (protest the music)* and *beat both (move up the tempo of the song*).

In Figs. 7 and 8, the results, respectively, obtained on iconic gestures subset and metaphoric gestures. Our approach yields recognition rates close to the ones obtained on the same corpus [4, 38–40], excerpt for certain metaphoric gestures: *start system*, *wind it up*, and *beat both* (the worst score is reached for this last category, with 39.1% recognition rate). In contrast with our real-time recognition system, these approaches use non-dynamic supervised machine learning algorithms (such as SVM, Hidden Conditional Random Fields, Random Decision Forests…) and require the use of global descriptors (e.g., dealing with the entire gesture duration).

Table 7 Cumulative recognition rates for metaphoric gestures of MSRC-12 dataset

	Beat both	Bow	Had enough	Push right	Start system	Wind it up	Mean
$RR_{cum}(1)$	39.1	95.6	86.7	95.7	68.9	65.0	75.2
$RR_{cum}(2)$	76.1	97.8	97.8	100.0	93.3	90.0	92.5
$RR_{cum}(3)$	84.8	100.0	100.0	100.0	100.0	95.0	96.6

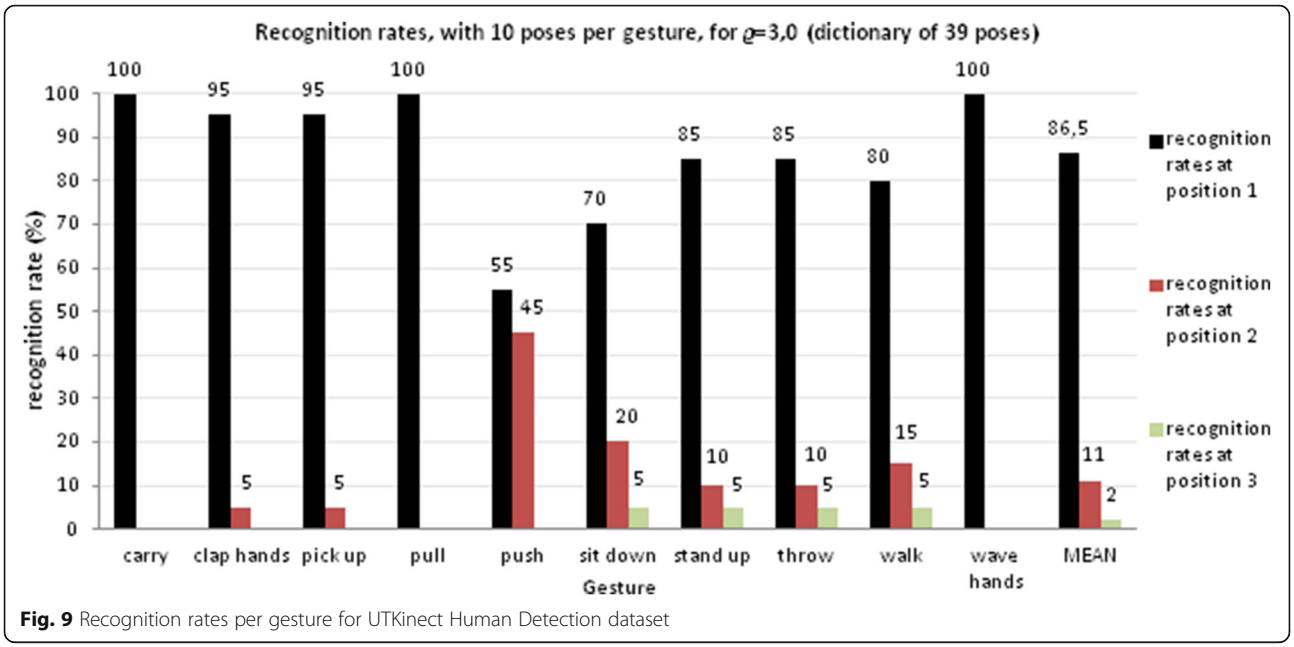

Fig. 9 Recognition rates per gesture for UTKinect Human Detection dataset

Tables 6 and 7 present the cumulated scores (Eq. 24) for iconic and metaphoric gestures. Globally, the mean recognition scores in first position RR(1), respectively calculated over the iconic gestures and metaphoric gestures subsets, are of 88.6 and 75.2%. When also considered the cumulative recognition rates $RR_{cum}(2)$ and $RR_{cum}(3)$, the correct category is retrieved in the first two (resp. three) candidates in 97.4% (resp. 99.3%) of cases for iconic gestures, in 92.5% (resp. 96.6%) of cases for metaphoric gestures.

We also tested our method on UTKinect-Human Detection dataset which is described exploited in [6] and puts at stake the following 10 gestures: *carry*, *clap hands*, *pick up*, *pull*, *push*, *sit down*, *stand up*, *throw*, *walk*, and *wave hands*. The results in Fig. 9 show that half of the gestures reach recognition rates greater than 95%. There is only one category whose score is less than 70%, which shows the relevance of our approach.

Table 8 presents the cumulated scores for the whole dataset. Globally, the correct category is retrieved in first position RR(1), in the first two candidates, and in the first three candidates, respectively, in 86.5, 97.5, and 99.5% of cases.

Such results can be compared to the ones obtained in [6] on the same corpus, where the authors yield 90.9% mean recognition rate, and five categories of gestures reach 96.5% at least. In [41], different approaches based upon global representations of gestures are tested, and yield mean recognition rates between 80.8 and 91.9%.

We have proved that the characterization of expressivity is a key for gestures understanding. Still, the confusion between close gestures at recognition stage has showed that the structural aspects of the motion are necessary to discriminate actions whose body parts coordination and correlation are similar. For this purpose, we need to fuse or aggregate the information coming both from intentionality, communication and style, and purely visual aspects like motion structuration or more local indices.

The most important computational burden concerns the HMM learning stage: 23 min 45 s for the entire database on an Intel Xeon CPU E5–1620 0 3.60GHz processor with 16.0 Go RAM. However, this is only an off-line stage and thus with no impact on the real-time recognition process. At the testing stage, the classification (including LMA feature extraction, soft assignment and HMM decision) is achieved at a rate of 3.2 ms per frame, which is obviously fast enough for real-time applications.

Table 8 Cumulative recognition rates for UTKinect Human Detection dataset

	Carry	Clap hands	Pick up	Pull	Push	Sit down	Stand up	Throw	Walk	Wave hands	Mean
$RR_{cum}(1)$	100	95	95	100	55	70	85	85	80	100	86.5
$RR_{cum}(2)$	100	100	100	100	100	90	95	95	95	100	97.5
$RR_{cum}(3)$	100	100	100	100	100	95	100	100	100	100	99.5

5 Conclusion and perspectives

In this paper, we introduced a gesture analysis approach, based on a set of local descriptors dedicated to the various entities defined in the LMA framework. Our approach yielded high recognition rates on a dataset of 11 actions. At short time, our perspectives of future work concern the evaluation of our approach on sequences of actions, or in different applications frameworks, that can involve gaits, affective states or emotions analysis. At a longer term, we plan to enrich the LMA representation with additional features, able to take into account relatively small details and thus to discriminate better between globally similar, but different gestures.

Acknowledgements
Not applicable.

Funding
Part of this work has been supported by the ITEA2 Empathic Products project (https://itea3.org/project/empathic.html).

Authors' contributions
AT implemented the LMA-based dynamic gesture recognition framework proposed in this paper and conducted the experiments.TZ supervised the research work and reviewed the paper. Both authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 19 April 2017 Accepted: 21 July 2017
Published online: 01 August 2017

References

1. Gordon Kurtenbach and Eric A. Hulteen, "Gestures in human-computer communication," *The art of human-computer interface design*, (1990) pp. 309–317
2. S. Ali Etemad and Ali Arya, "Correlation-optimized time warping for motion," *The Visual Computer*, (2014)
3. X. Jiang, F. Zhong, Q. Peng, X. Qin, Online robust action recognition based on a hierarchical model. Vis. Comput. **30**, 1021–1033 (2014)
4. Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, (Press, AAAI, 2013), pp. 2466–2472
5. Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin, "Instructing people for training gestural interactive systems,: in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, (ACM, 2012), pp. 1737–1746
6. Lu Xia, Chia-Chih Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (IEEE, 2012), pp. 20–27
7. Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (IEEE, 2010), pp. 9–14
8. P.B. Braem, T. Bräm, A pilot study of the expressive gestures used by classical orchestra conductors. J. Conductor's Guild **22**(1–2), 14–29 (2001)
9. D. Glowinski et al., Toward a minimal representation of affective gestures. IEEE Trans. Affect. Comput. **2**(2), 106–118 (2011)
10. Dominique Boutet, "Une morphologie de la gestualité: structuration articulaire," *Cahiers de linguistique analogique*, no. 5, pp. 81–115, 2008
11. Pengcheng Luo and Michael Neff, "A perceptual study of the relationship between posture and gesture for virtual characters," in *Motion in Games*, (Springer, 2012), pp. 254–265
12. Arthur Truong, Titus Zaharia "Dynamic Gesture Recognition with Laban Movement Analysis and Hidden Markov Models", Proceedings of the 33rd Computer Graphics International conference (CGI'16), pp. 21–24. Heraklion, Greece — June 28 - July 01, doi: 10.1145/2949035.2949041
13. R. Laban, *La Maîtrise du Mouvement* (Actes Sud, Arles, 1994)
14. Yale Song, David Demirdjian, and Randall Davis, "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, 2011), pp. 500–506
15. V.K. Singh, R. Nevatia, Simultaneous tracking and action recognition for single actor human actions. Vis. Comput. **27**(12), 1115–1123 (2011)
16. Meinard Müller et al., Documentation mocap database hdm05, Citeseer ed., 2007
17. Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe, "Multimodal analysis of expressive gesture in music and dance performances," *Gesture-based communication in human-computer interaction*, (2004), pp. 20–39
18. A. Camurri, I. Lagerlöf, G. Volpe, Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. Int. J. Human-Comput. Stud. **59**(1), 213–225 (2003)
19. Hatice Gunes and Björn Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, (2012)
20. Daniel Bernhardt and Peter Robinson, "Detecting affect from non-stylised body motions," in *Affective Computing and Intelligent Interaction*, (Springer Berlin Heidelberg, 2007), pp. 59–70
21. Dilip Swaminathan et al., "A dynamic bayesian approach to computational laban shape quality analysis," *Advances in Human-Computer Interaction*, vol. 2009, pp. 1–17, 2009
22. L. Zhao, N.I. Badler, Acquiring and validating motion qualities from live limb gestures. Graph. Model. **67**(1), 1–16 (2005)
23. Durell Bouchard and Norman Badler, "Semantic segmentation of motion capture using laban movement analysis," in *Intelligent Virtual Agents*, (Springer Berlin Heidelberg,2007), pp. 37–44
24. T. Nakata, T. Mori, T. Sato, Analysis of impression of robot bodily expression. J. Robotics Mechatronics **14**(1), 27–36 (2002)
25. Kozaburo Hachimura, Katsumi Takashina, and Mitsu Yoshimura, "Analysis and evaluation of dancing movement based on LMA," in *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005*, (IEEE, 2005), pp. 294–299
26. Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, and Joseph T Kider Jr, "Efficient motion retrieval in large motion databases," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, (ACM, 2013), pp. 19–28
27. Ali-Akbar Samadani, Sarahjane Burton, Rob Gorbet, and Dana Kulic, "Laban effort and shape analysis of affective hand and arm movements," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, (IEEE, 2013), pp. 343–348
28. Andreas Aristidou and Yiorgos Chrysanthou, "Feature Extraction for Human Motion Indexing of Acted Dance Performances," *GRAPP 2014 - International Conference on Computer Graphics Theory and Applications*, 2014
29. Arthur Truong, Hugo Boujut, and Titus Zaharia, "Laban descriptors for gesture recognition and emotional analysis," *The Visual Computer*, (2015) pp. 1–16
30. Rudolf Laban, *Espace Dynamique*, Contredanse, Ed. Bruxelles, 2003
31. L.R. Rabiner, B.-H. Juang, An introduction to hidden Markov models. ASSP Magazine IEEE **3**(1), 4–16 (1986)
32. Alfons Juan and Enrique Vidal, "Comparison of four initialization techniques for the k-medians clustering algorithm," in *Advances in Pattern Recognition*, (Springer, 2000), pp. 842–852
33. E. Szmidt, J. Kacprzyk, Distances between intuitionistic fuzzy sets. Fuzzy Sets Syst. **114**(3), 505–518 (2000)
34. Lingqiao Liu, Lei Wang, and Xinwang Liu, "In defense of soft-assignment coding," in *2011 IEEE International Conference on Computer Vision (ICCV)*, (IEEE, 2011), pp. 2486–2493

35. Tobias P. Mann, "Numerically stable hidden Markov model implementation," in *An HMM scaling tutorial.*, (2006), pp. 1–8
36. Pauline Larue, Pierre Jallon, and Bertrand Rivet, "Modified k-mean clustering method of HMM states for initialization of Baum-Welch training algorithm," *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 951–955, 2011
37. Douglas Reynolds, "Gaussian mixture models," *Encyclopedia of Biometrics*, (2015) pp. 827–832
38. Yale Song, Louis-Philippe Morency, and Randall Davis, "Distribution-Sensitive Learning for Imbalanced Datasets," in *2013 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2013)*, (IEEE, 2013)
39. Farhood Negin, Firat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil, "A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras," in *Image Analysis and Recognition*, (Springer, 2013), pp. 648–657
40. Xin Zhao, Xue Li, Chaoyi Pang, and Xi Zhu, "Online human gesture recognition from motion data streams," in *Proceedings of the 21st ACM international conference on Multimedia*, (ACM, 2013), pp. 23–32
41. Yu Zhu, Wenbin Chen, and Guodong Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (IEEE, 2013), pp. 486–491