Dealing with Variability When Recognizing User's Performance in Natural 3D Gesture Interfaces

Article *in* International Journal of Pattern Recognition and Artificial Intelligence · October 2013

DOI: 10.1142/S0218001413500237

4 authors:

Sorel Anthony
Université de Rennes 2
17 PUBLICATIONS 50 CITATIONS
SEE PROFILE

Richard Kulpa
Université de Rennes 2
82 PUBLICATIONS 905 CITATIONS
SEE PROFILE

Emmanuel Badier
University of Geneva
2 PUBLICATIONS 3 CITATIONS
SEE PROFILE

Franck Multon
Université de Rennes 2
146 PUBLICATIONS 1,422 CITATIONS
SEE PROFILE

International Journal of Pattern Recognition and Artificial Intelligence
© World Scientific Publishing Company

Dealing with variability when recognizing user's performance in natural 3D gesture interfaces

ANTHONY SOREL

M2S - University Rennes 2, Artefacto,
UFR-APS, Avenue Charles Tillon
Rennes, France
anthony.sorel@univ-rennes2.fr

RICHARD KULPA

M2S - University Rennes 2,
UFR-APS, Avenue Charles Tillon
Rennes, France
richard.kulpa@univ-rennes2.fr

EMMANUEL BADIER

M2S - University Rennes 2,
UFR-APS, Avenue Charles Tillon
Rennes, France
emmanuel.badier@gmail.com

FRANCK MULTON

M2S - University Rennes 2,
UFR-APS, Avenue Charles Tillon
Rennes, France
franck.multon@univ-rennes2.fr

Recognition of natural gestures is a key issue in many applications including videogames and other immersive applications. Whatever is the motion capture device, the key problem is to recognize a motion that could be performed by a range of different users, at an interactive frame rate. Hidden Markov Models (HMM) that are commonly used to recognize the performance of a user however rely on a motion representation that strongly affects the overall recognition rate of the system. In this paper, we propose to use a compact motion representation based on Morphology-Independent features and we evaluate its performance compared to classical representations. When dealing with 15 very similar upper limb motions, HMM based on Morphology-Independent features yield significantly higher recognition rate (84.9%) than classical Cartesian or angular data (70.4% and 55.0% respectively). Moreover, when the unknown motions are performed by a large number of users who have never contributed to the learning process, the recognition rate of Morphology-Independent input feature only decreases slightly (down to 68.2% for a HMM trained with the motions of only one subject) compared to other features (25.3% for Cartesian features and 17.8% for angular features in the same conditions). The method is illustrated through an interactive demo in which three virtual humans have to interactively recognize and replay the performance of the user.

Each virtual human is associated with a HMM recognizer based on the three different input features.

Keywords: motion recognition, Hidden Markov Models, natural interaction, virtual human, Morphology-Independent representation

1. Introduction

Nowadays, a variety of interaction devices allows the user to interact with a virtual environment in a natural manner. Low-cost systems have been widely used in the videogame industry to directly interact with the game by using natural motions, such as moving a device (Nintendo Wii or Sony Motion Controller) or moving the full body (Microsoft Kinect). Modern mobile devices open a new way to interact in 3D space in their camera's field of view[39]. In serious games and virtual reality, more accurate systems are generally used, such as magnetic sensors or optoelectronic systems. Whatever the 3D motion capture device, one of the main challenges consists in recognizing user's actions and computing an appropriate reaction of the virtual environment at interactive frame rate.

Prior motion recognition methods have generally been applied to a set of very discriminated actions, such as walking, crouching, grasping[38], or to actions exhibiting strongly constrained spatiotemporal patterns[32]. However, navigating and interacting in immersive environments require to deal with very similar upper limb motions, such as manipulation tasks, pointing, grasping, hitting, pushing, pulling, punching. In that case, the intrinsic properties of the motions are very similar as they involve the same limited number of degrees of freedom in the same subspace. As most previous methods generally rely on geometric features to classify the user's performance, the selection of the most relevant features is a key point to address motion recognition. The selection of inappropriate features lead to failures in classifying similar gestures because of inaccuracy in capturing the intrinsic properties of each gesture. Indeed, motion capture data are strongly linked to the user's anthropometric data: long arms will provide larger displacements of sensors than smaller arms even if the movement is supposed to be the same. Relevant features would also make it possible to deal with motion variability: grasping performed by two different users at several different target positions in space should be all recognized as the same motion.

Most related works in this field is based on Hidden Markov Models (denoted HMM). However, HMMs rely on features that strongly affect the recognition performance, especially for very similar movements. In this paper we propose an original alternative to classical Cartesian position of body joints[17] or Euler angles[28,18] in order to limit the impact of morphological variations among users: the Morphology-Independent feature. This feature is based on a normalized representation of the human motion that is less sensitive to variation of morphologies. When used as input of a classical HMM classifier, this feature exhibits significantly higher recognition rates and seems to be promising in discriminating very close gestures such as slapping with the palm or with the back of the hand. The resulting recognition system has been tested on two-arm motions which are most commonly used in videogames where the user is carrying devices in his hands to interact with the virtual environment.

2. Related work

3D gesture recognition is used in videogames or serious games either to drive avatars or to interact with simulated worlds. For the former type of application, many researchers have worked on using accurate motion capture data to animate an avatar[2,26]. Such methods compute the required information to animate each joint of the virtual human while correcting some inaccuracies. Motion reuse has also been widely explored to adapt motion capture data to characters with various sizes and with different kinematic constraints[10,34,20]. However, in many cases and especially in the game industry, low-cost devices are used and generate poor and noisy information. With these systems, directly animating an avatar is difficult and alternative solutions based on gesture recognition have been proposed.

Performance-driven animation generally relies on retrieving the motion that best corresponds to the user's performance by searching through a wide database of prerecorded motions. For example, some authors proposed to control an avatar in an iconic and intentional manner[15]. This approach used simple metaphors based on the displacement of a plush doll. Such a performance-driven interface allows more natural interaction facilities. With a few cameras and markers, it is thus possible to animate the avatar of the user[5]. The low-dimensional control signals are transformed into full-body motions by constructing a series of local models from a motion capture database. Some of these methods take dynamics into account[14]and yield impressive results for motions that are very different from each other. Slyper and Hodgins introduced a simple metric to seek a database of motions in order to retrieve the closest one to the user's performance[36] without considering the semantics of the motion. On the opposite, pattern recognition techniques make it possible to retrieve the semantic information of the motion in order to control an avatar[35,22]. However, this type of system is generally limited to simple and significantly different motions such as swinging arms or legs.

Whatever the type of interactive application, designing a recognition system to automatically deal with a wide range of user morphologies and situations remains challenging. Another problem is due to the high-dimensional and noisy nature of motion capture data. Data reduction (Principal Components Analysis[24]) and filtering methods (Hidden Markov Models[6,25], Finite-State Machines[12,13], Kalman filters[31], or more advanced particle filters and condensation algorithms[16,21]) have been used to address this problem in computer vision, in inertial sensing or in other application domains dealing with noisy sensors and/or noisy experimental conditions.

Gesture recognition performed using 3D captured motions of users classically uses raw Euler orientations[28] of the body segments despite their inherent non-linearity and their direct link with the morphology of the user[18]. For example, two users with different morphologies have indeed significantly different angles between their arm and forearm even when performing the same task, such as clapping the hands or putting the arms in the pockets. An alternative consists in defining relevant geometric features that capture some of the semantic information. Most of these techniques discretize the input space into clusters or areas[27,23,9], or use Laban notation[40] to encode complex motions in a compact and meaningful manner. Whatever the method, comparing two motions consists in computing the distance between them using a specific metric generally based on sequences of discrete features. Some of these geometric features are independent of the morphology of the user, such as checking if a hand is over a shoulder or if a foot is in front of the character[27]. These features are then efficiently combined to retrieve a set of motions that satisfy predefined geometric constraints. Nevertheless, the discretization of these features leads to inaccuracy at the boundary between two discrete values. This inaccuracy of each feature combined to the high dimensionality of the feature vector can produce conflicting action classifications. As a consequence, the recognition systems may fail to distinguish very similar motions, such as throwing an object or punching someone. Besides, scientific literature is sparse in evaluating recognition systems on similar gestures. To reduce feature dimensionality while preserving essential motion information, some authors proposed to boost the classifiers[25], so that the system automatically selects the features that provide the highest performance for each classification task. This technique is time consuming and involves long machine learning processes and huge databases of training samples. Raamana et al. [29] also analyzed the relevance of various features in 3D gesture recognition. Their paper introduced the *shoulder to wrist* feature: the direction of the wrist in the shoulder reference frame. It demonstrated that systems relying on this feature outperformed those relying on angular features in a table-top scenario. Although their gesture database exhibited intra-class variability (displacement in several directions and heights) and inter-class similarity (similar type of gestures), the authors did not evaluate the ability of the feature to deal with morphological variability. Furthermore, the study was limited to 3 gesture classes.

Defining a generic set of features that would lead to high recognition rates and reliable results despite morphology and style variations is still a difficult task. To our knowledge, although a lot of studies have tried to preprocess the raw data (discretization, projection on more subtle subspaces), only a few contributions explicitly addressed the problem of reducing inter-subject variability in the context of natural 3D motion recognition. Yet, the reduction of the inter-individual morphological variability at motion representation stage should improve the downstream classification result. The classification stage can therefore closely focus on motion class variability since the morphological variability is dealt with before. As Turaga et al. discussed in their survey[37], dealing with anthropometric variations is still an important challenge and requires careful attention.

In computer animation, motion retargeting has proposed some solutions to efficiently tackle this problem, such as solving kinematic constraints[10,7] or designing Morphology-Independent representations [20,11]. In this paper we propose to adapt such Morphology-Independent representation to gesture recognition and demonstrate that this type of features could efficiently address the problem of multi-user motion recognition based on the most common method, namely HMM. The following section recalls the general principle of HMM and how it has been used in this work.

3. HMM-based recognition system

HMM are stochastic models that have been widely used to encode time series as piecewise stationary processes. In fact, their Markovian nature that links the most recent observations to the future ones makes them very suitable for sequential data modeling. Basically a time-varying feature (for instance a trajectory) is modeled as a state automaton in which each state stands for a range of possible observation values of the feature while the transitions between states can model time (Figure 1). The feature observation values and the transitions between states are driven by probabilities, which makes HMMs very robust to spatiotemporal variations.

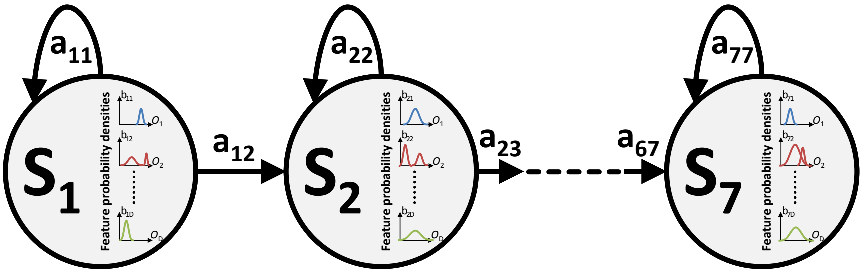

a_{11}
a_{22}
a_{77}
a_{12}
a_{23}
a_{67}
S_1
S_2
S_7
Feature probability densities
Feature probability densities
Feature probability densities

Fig. 1. Graphical representation of our Bakis HMM with 7 states $\{S_i\}$ with diagonal covariance distribution matrix. Feature probability densities b_{ik} are represented inside each state. A Bakis HMM only allows self-transition and transitions from $\{S_i\}$ to $\{S_{i+1}\}$. These state transition probabilities are given by a_{ij}.

In this work, each gesture is encoded as a time-varying D-dimensional feature vector with continuous values (typically a 3D trajectory). D stands for the number of features. Such a gesture is modeled as a feature vector of length T formally noted $\boldsymbol{O} = \boldsymbol{O}(1), \ldots, \boldsymbol{O}(t), \ldots, \boldsymbol{O}(T)$, where each feature vector is $\boldsymbol{O}(t) = (O_1(t), \ldots, O_d(t), \ldots, O_D(t))$ for a given frame t. Note that bold letters stand for vectors.

To recognize a class of gesture m, its spatiotemporal dynamic has to be modeled as a HMM λ_m. This is the training phase. To fully encompass the huge variability of the gesture class, the training algorithm requires a great number of repetitions with variability of gestures (in term of joint trajectories, velocities and amplitudes) and in users' morphologies. This training is performed thanks to an expectation-maximization (EM) procedure: the iterative Baum-Welch method[30].

During the classification phase, an unknown motion is classified among the trained movement models $\Lambda = \{\lambda_1, \dots, \lambda_M\}$. To this end, we concurrently compute to which degree an unknown motion observation sequence $\boldsymbol{O}$ matches each λ_m (see Figure 2). $\boldsymbol{O}$ is then associated to the class of gesture model which provides the maximum likelihood:

$$GestureClass(\boldsymbol{O}) = \arg \max_{m=1,\dots,M} P(\boldsymbol{O}|\lambda_m)$$

where $P(\boldsymbol{O}|\lambda_m)$ expresses the similarity between the motion model λ_m and the unknown motion $\boldsymbol{O}$.

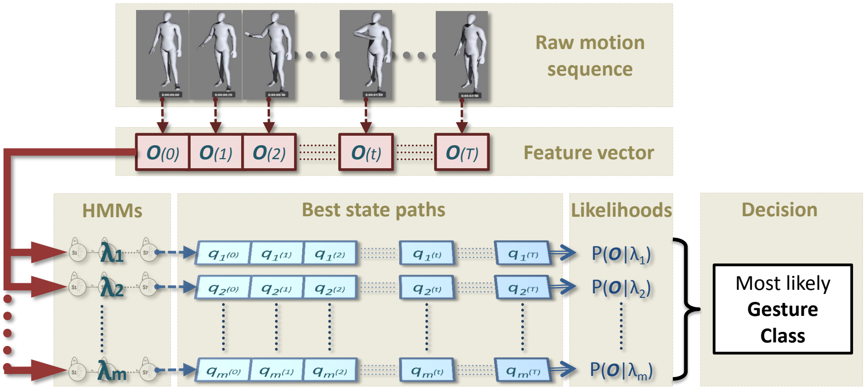


Fig. 2. Classification process. 1) Raw motion data corresponding to an unknown gesture serve as input. 2) Features are extracted. 3) For each HMM λ_m, the Viterbi algorithm determines the state path $(q_m(1,\dots,T))$ that best matches the feature sequence, and provides the corresponding likelihood. 4) The HMM associated to the greatest likelihood is selected as the gesture class that best represents the gesture to recognize.

In this study, each gesture model λ_m is encoded as a first-order continuous Bakis HMM with 7 states (Figure 1). The Bakis topology, also called left-right topology, requires transitions to be either self-transition or between states $\{S_i\}$ and $\{S_{i+1}\}$. In fact, Bakis topology seems well adapted to gesture modeling as it correctly models its sequential nature, especially when no cyclic motion occurs. Moreover Romaszewski and Glomb[33] showed that ergodic (or fully connected) and Bakis topologies yield best results. But the greater computational complexity of the ergodic topology makes it less interesting for real-time purpose. Bakis topology was thus selected in this paper.

All HMMs have 7 states with a mixture density of maximum 6 multivariate Gaussian distributions with diagonal covariance matrix. Automatically deciding the number of states and mixture components is hard in practice[1]. Although the

Akaike or Bayesian Information Criterion have been proven to be suitable to optimize the number of components in a Gaussian mixture, very few practical studies have been carried out to evaluate these criteria to select the best number of states number[4]. As suggested in the literature[25], we have made a pre-experiment to determine the number of states and mixture components that are the best trade-off between computational complexity and achievement of good classification. In this experiment, the number of states and of Gaussian components vary from 1 to 20 and the database was randomly half-splitted between training and validation sets. For each set of input features, using more states or more Gaussian components than the trade-off parameters (i.e. 7 states and 6 Gaussian components) leads to very little improvement of the recognition rate (less than 1%) of the validation set, while increasing the computation time. Moreover we checked that the recognition rates of each gesture class reached a plateau.

Once the HMM parameters have been selected, all HMMs λ_m are trained over repetitions of the respective gesture classes m. After this training phase, the real-time classification of any unknown gesture can be performed, as depicted in Figure 2.

4. Morphology-Independent features

Raw motion data are represented as a hierarchy of 19 body segments. Each body segment can move around 3 orthogonal degrees of freedom (DoF) corresponding to rotations, resulting in a total of 57 DoF. In this work we focused on upper limb motions and only the corresponding DoF are considered for recognition. Among the wide variety of kinematic representations, most researchers either rely on Cartesian data (e.g.[17]) or Euler angles (e.g.[19,28]). However, these parameters directly depend on the morphology of the user. The resulting HMM recognizer may thus be very sensitive to variations in the user's dimensions, and may lead to bad performance when dealing with new users. In this paper, we propose to evaluate if Morphology-Independent features can overcome this limitation. We consequently evaluate the performance of the HMM recognizer based on the three following representations:

- Euler-based $\boldsymbol{O}_{Euler}$ composed of 24 features: the 3 local orientation angles of each shoulder and elbow, plus the corresponding derivatives. These latter derivatives are included for each feature representation because, as highlighted by Campbell et al. [3], velocities should play a major role in recognition tasks.
- Cartesian-based $\boldsymbol{O}_{Cartesian}$ composed of 24 features: the 3 Cartesian positions of each wrist and elbow in the shoulder reference frame for both sides, plus the corresponding velocities. All the 3D coordinates are given in a local coordinate frame attached to the hips $\mathcal{R}_{Hips}$ so that root orientation is compensated. Features are thus invariant with respect to where the action is facing.
- and the Morphology-Independent $\boldsymbol{O}_{MI}$ composed of 12 features: the 3 nor-

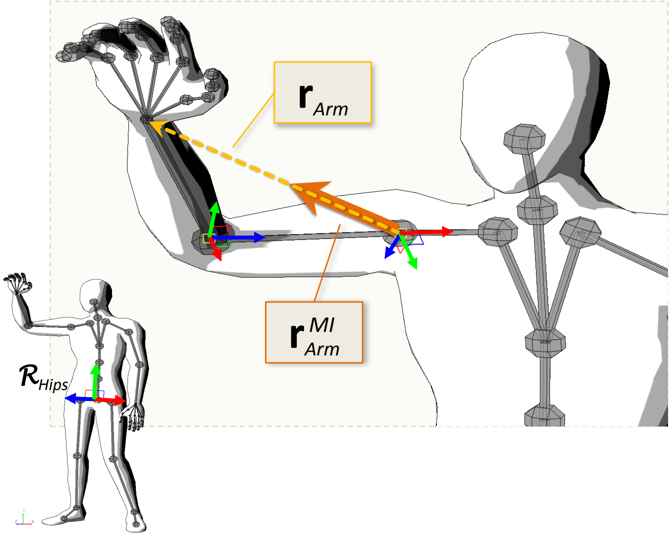


Fig. 3. The Morphology-Independent representation encompasses two 3D vectors $\mathbf{r}^{MI}_{Arm}$ linking each shoulder to its respective wrist expressed in the local frame $\mathcal{R}_{Hips}$, divided by the total length of the arm (dark orange arrow). Their velocities are also included. The light orange dotted arrow indicates the Cartesian vector linking each shoulder to its respective wrist $\mathbf{r}_{Arm}$. The Cartesian feature representation also includes the derivative $\dot{\mathbf{r}}_{Arm}$. The Euler features are based on the 3 rotations around local shoulder and elbow frames (red/green/blue frames), plus their derivatives.

malized positions of both wrists in the corresponding shoulder reference frame, plus the corresponding derivatives (see Figure 3). This motion representation is inspired from the work of Kulpa et al.[20] for motion editing in computer animation. We adapted this representation to gesture recognition by taking velocities into account and not considering elbow joints that are very sensitive to variation in morphology.

The assumption behind Morphology-Independent feature is that most of the human tasks deal with kinematic constraints expressed in world Cartesian positions, such as the position of an object during grasping or manipulation, or the position of the hands during clapping for instance. However, the length of the body segments may be different from one user to another, leading to different angular configurations for the same Cartesian task. For instance, clapping the same way for two different users may lead to different elbow angles. On the opposite, using Cartesian positions ensures an accurate evaluation of the final joint that is supposed to interact with 3D objects in the environment but the intermediate joints varied a lot depending on the morphology. To deal with various upper limb dimensions, this information has to be normalized by the length of the corresponding kinematic chain: arm and forearm. This Morphology-Independent representation of the upper body is thus a scaled version of the Cartesian-based representation: 3D vectors are normalized by their maximum extension, i.e. the arm length (see Figure 3).

$$\mathbf{r}^{MI}_{Arm}(t) = \frac{\mathbf{r}_{Arm}(t)}{\max \|\mathbf{r}_{Arm}\|}$$

This representation is expected to reduce the influence of the subject morphology: when the arm is totally extended $\|\mathbf{r}^{MI}_{Arm}\| = 1$ and when it is half extended $\|\mathbf{r}^{MI}_{Arm}\| = 0.5$ whatever the size of the subject. We also add the corresponding derivatives $\dot{\mathbf{r}}_{LArm}$ and $\dot{\mathbf{r}}_{RArm}$. As a result, at time t, the MI-based features are:

$$\boldsymbol{O}_{MI}(t) = (\mathbf{r}^{MI}_{LArm}, \dot{\mathbf{r}}^{MI}_{LArm}, \mathbf{r}^{MI}_{RArm}, \dot{\mathbf{r}}^{MI}_{RArm})(t)$$

The Morphology-Independent feature is different from the *shoulder to wrist* direction feature (S2W) introduced by Raamana et al. [29]. Firstly, their feature capture the direction of the wrist in the shoulder reference frame whereas the Morphology-Independent feature also capture the normalized distance between the shoulder and the wrist. This information is particularly important since it defines the extension of the arm. Indeed, a subject in T-pose or touching his shoulder with his fingers would roughly result in an equal S2W value if the direction of the wrist is unchanged. The MI feature discriminates such postures. Secondly, we added velocity information as suggested by Campbell et al. [3].

5. Evaluation method

In order to evaluate the relevance of using Morphology-Independent data as input of natural motion recognition systems, we tested this feature against the two types of features classically used in the literature and described above: Euler angles and Cartesian positions. These three features are used as inputs for a common HMM-based system to recognize natural 3D gestures.

As existing motion capture databases are not dedicated to similar upper limb motions, we created our own database with 15 different gestures (Figure 4): applause, crossing arms, slap with palm, slap with back hand, touch chin, throw something, hands on hips, hands in pocket, grasp something at hip level, grasp something high, grasp something at chest level, punch, hello high (with one hand above the head), hello head (with one hand at head height) and uppercut.

10 subjects (age 25±4 years old, height 171±11cm) performed each motion at least 5 times each side and were instructed to include high variability (in speed, location and amplitude). This population included 5 men and 5 women with different morphologies. We used an Optitrack system (product of Natural Point) to capture the motion of 34 reflective markers placed on the whole body of the subject. After post-processing (such as interpolation of missing data) the resulting data were stored in BVH files. The BVH files contain data available for the full body but only data linked to the two arms were used in the following experiments. This database is freely available online[a].

[a] *www.m2slab.com/data/gesture-recognition/database*

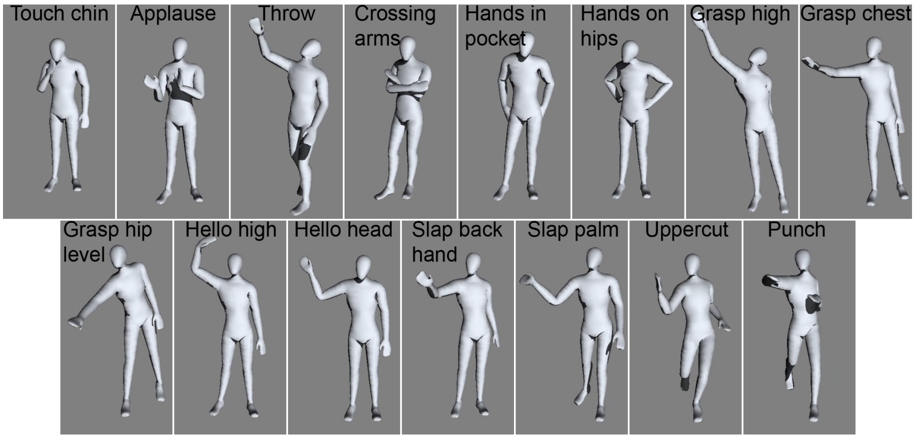


Fig. 4. The 15 upper limb motions used to evaluate the system. Some of the motions share very similar geometric properties to evaluate the performance of the recognition system.

Our database involves two-hand motions and some of them share very similar geometric properties, such as slapping with palm or with the back of the hand. In previously published works, most of the authors focused on very different motions such as walking, grasping, sneaking. In most interactive applications the user cannot move around a lot and has to mainly use his arms to interact with the environment. It is therefore important to be able to correctly recognize and distinguish these upper limb motions even if they seem to be very similar.

Several experiments have been carried out to evaluate the impact of morphology on the recognition rate of these three types of features. They consist in using the motions of a subset of subjects to train the recognition system while using the motions of the other subjects for validation. The subjects used for validation are new users for the system and this kind of experiment demonstrates the influence of new morphologies on the recognition rate. This experiment is known as "Leave-One-Out approach" and denoted L1O in the paper. The evaluation of the features is then made by successive experiments from the L1O (one subject not used for training) up to L9O in which only one subject is used for training and the system tries to recognize the motions of all the other subjects.

5.1. *Leave-One-Out approach*

When extracting the motions of only one subject in the database for training, the Morphology-Independent feature already shows significantly better performance than other features: 85% for Morphology-Independent data, 70% for Cartesian data and 55% for Euler data. A Friedman's ANOVA demonstrates the influence of features on recognition rate, $\chi^2(2, N = 150) = 52.91$, $p < 0.0001$. A post-hoc Wilcoxon signed rank test shows the significantly higher recognition rate for the features based on Morphology-Independent data compared to Cartesian data ($T = 0.343, p < 0.05$) and Euler angles ($T = 0.657, p < 0.05$). Moreover, the smaller standard deviation of Morphology-Independent data shows that it seems to be less sensitive to deal with new users contrary to other features that have high standard deviations. The latter are mainly due to large variations between subjects. For instance, the subject with the worst score is associated with a mean recognition rate of 35.3% when using Euler data. The subject with the highest score obtained a mean recognition rate of 76.2% when using the same data. The recognition rate becomes 66.0% and 99.3% respectively for the subjects with the lowest and highest scores when using Morphology-Independent data.

Table 1 provides the recognition rate for each type of motion for the three tested features. One can see that using Morphology-Independent data leads to higher performance (minimum is 63% for a *punch*) compared to Cartesian data (minimum is 47% for *grasping an object placed at a high position*) and Euler data (minimum is 17% for *grasping an object placed at a high position*).

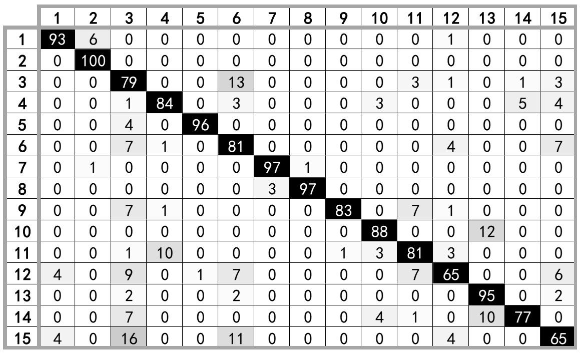
Motion	Morph.-Ind. (%)	Cartesian (%)	Euler (%)
1. Applause	**90.9**	88.9	69.9
2. Crossing arms	**100**	87.5	80
3. Slap with palm	**76.4**	61	52.3
4. Slap back hand	**80.2**	75.4	45.1
5. Touch chin	**96.4**	68	53
6. Throw	**77.5**	70.8	69.2
7. Hands on hips	**96.7**	87.3	88.3
8. Hands in pocket	**97.2**	90	68
9. Grasp hip level	**84.9**	46.7	17.1
10. Grasp high	**90.6**	70.9	89.3
11. Grasp chest	**84.2**	57.3	35.5
12. Punch	62.6%	**64.7**	51
13. Hello high	**94.5**	80.7	26.5
14. Hello head	**77.7**	33.4	24.2
15. Uppercut	64.2	**73.7**	55.7
Mean ± std	**84.9 ± 11.8**	70.4 ± 16.2	55.0 ± 22.7

Table 1. Recognition rate for each motion and each type of feature for the Leave-One-Out approach. Statistical analysis confirm the significant higher recognition rate of Morphology-Independent features compared to the others.

The confusion matrices for all 10 L1O experiments are presented in figures 5, 6 and 7 for each feature.

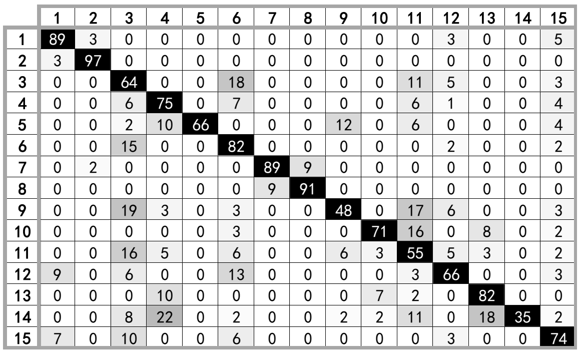
5.2. *Leave-n-Out approach*

In real situations, the recognition system cannot be trained on all the potential users. On the contrary, the training is performed on a small database compared



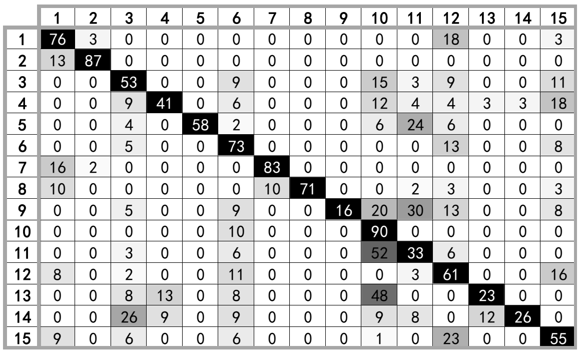
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	93	6	0	0	0	0	0	0	0	0	0	1	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	79	0	0	13	0	0	0	0	3	1	0	1	3
4	0	0	1	84	0	3	0	0	0	3	0	0	0	5	4
5	0	0	4	0	96	0	0	0	0	0	0	0	0	0	0
6	0	0	7	1	0	81	0	0	0	0	0	4	0	0	7
7	0	1	0	0	0	0	97	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	3	97	0	0	0	0	0	0	0
9	0	0	7	1	0	0	0	0	83	0	7	1	0	0	0
10	0	0	0	0	0	0	0	0	0	88	0	0	12	0	0
11	0	0	1	10	0	0	0	0	1	3	81	3	0	0	0
12	4	0	9	0	1	7	0	0	0	0	7	65	0	0	6
13	0	0	2	0	0	2	0	0	0	0	0	0	95	0	2
14	0	0	7	0	0	0	0	0	0	4	1	0	10	77	0
15	4	0	16	0	0	11	0	0	0	0	0	4	0	0	65

Fig. 5. Confusion matrix of Morphology-Independent feature for L1O. True motion classes appear in rows, recognized motion classes appear in columns. See table 1 for correspondance between numbers and motion class. Note that each subject did not perform the same number of motions (between 5 and 7) leading to a slight difference with table 1.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	89	3	0	0	0	0	0	0	0	0	0	3	0	0	5
2	3	97	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	64	0	0	18	0	0	0	0	11	5	0	0	3
4	0	0	6	75	0	7	0	0	0	0	6	1	0	0	4
5	0	0	2	10	66	0	0	0	12	0	6	0	0	0	4
6	0	0	15	0	0	82	0	0	0	0	0	2	0	0	2
7	0	2	0	0	0	0	89	9	0	0	0	0	0	0	0
8	0	0	0	0	0	0	9	91	0	0	0	0	0	0	0
9	0	0	19	3	0	3	0	0	48	0	17	6	0	0	3
10	0	0	0	0	0	3	0	0	0	71	16	0	8	0	2
11	0	0	16	5	0	6	0	0	6	3	55	5	3	0	2
12	9	0	6	0	0	13	0	0	0	0	3	66	0	0	3
13	0	0	0	10	0	0	0	0	0	7	2	0	82	0	0
14	0	0	8	22	0	2	0	0	2	2	11	0	18	35	2
15	7	0	10	0	0	6	0	0	0	0	0	3	0	0	74

Fig. 6. Confusion matrix of Cartesian feature for L1O approach



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	76	3	0	0	0	0	0	0	0	0	0	18	0	0	3
2	13	87	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	53	0	0	9	0	0	0	15	3	9	0	0	11
4	0	0	9	41	0	6	0	0	0	12	4	4	3	3	18
5	0	0	4	0	58	2	0	0	0	6	24	6	0	0	0
6	0	0	5	0	0	73	0	0	0	0	0	13	0	0	8
7	16	2	0	0	0	0	83	0	0	0	0	0	0	0	0
8	10	0	0	0	0	0	10	71	0	0	2	3	0	0	3
9	0	0	5	0	0	9	0	0	16	20	30	13	0	0	8
10	0	0	0	0	0	10	0	0	0	90	0	0	0	0	0
11	0	0	3	0	0	6	0	0	0	52	33	6	0	0	0
12	8	0	2	0	0	11	0	0	0	0	3	61	0	0	16
13	0	0	8	13	0	8	0	0	0	48	0	0	23	0	0
14	0	0	26	9	0	9	0	0	0	9	8	0	12	26	0
15	9	0	6	0	0	6	0	0	0	1	0	23	0	0	55

Fig. 7. Confusion matrix of Euler feature for L1O approach

to all the persons (and their corresponding morphology) that can use the system. We therefore made the same experiment with 2 subjects left out from the training phase (L2O : Leave-2-Out), 3 subjects (L3O), up to the case where only the motions of one subject is used for the training phase and the motions of all the 9 others for the classification phase (L9O).

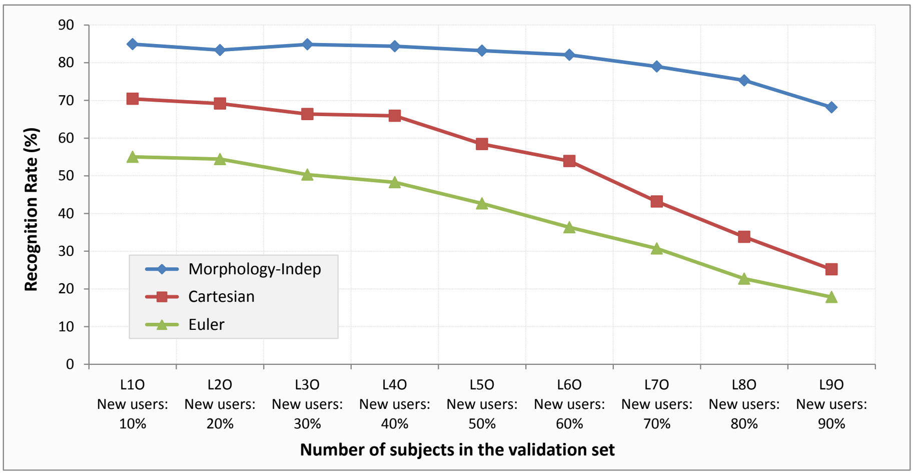


Fig. 8. For each feature, evolution of recognition rate depending on the evaluation approach. The more on the right of the graph, the greater is the number of new users who have to be recognized. Morphology-Independent feature recognition rate only slightly decreases when at least 60% of the users are unknown in the system.

The figure 8 shows the global recognition rate for all the type of motions depending on the evaluation method (from L1O to L9O). When the number of subjects used for training decreases (and thus the number of new users increases), the recognition rate of Cartesian and Euler features falls down more rapidly than Morphology-Independent features. The score of this latter indeed falls from 85% to 68% in the worst situation (the motions of only one subject are used for training). The recognition rate of both other features falls down to less than 30% for this worst situation. Moreover, the score of Morphology-Independent features only begins to fall when at least 60% of the motions to recognize are performed by new users (L6O).

The better performance of the Morphology-Independent feature, that was already significant for the L1O method, is still significant for all the other methods (from L2O to L9O). Friedman's ANOVA were again used to demonstrate the influence of features on recognition rate and post-hoc Wilcoxon signed rank tests confirmed each time the significantly higher recognition rate for the features based on Morphology-Independent data compared to Cartesian data and Euler angles.

The weak recognition performance obtained with classical features may not only be due to changes in morphology. A new user is not only a new morphology to deal with, but also a new style in performing motions. Let us consider two subjects J and K of same gender, with similar sizes and weights in our database. In L1O approach, the recognition rate of the *uppercut* motion is 20% for J, while it is 100% for K when using Cartesian data. It seems to demonstrate that the style of J is very different from the other subjects who were used to train the system. Conversely, K has a more "standard" style for *uppercut.* This kind of results also changes according to the motion class. For instance, *slapping with the palm* is 100% recognized for J and only 40% for K when using Cartesian data. In short, recognition rates depend on both the subject and the motion class. These results show that style is also a key problem when developing recognition systems. However Morphology-Independent data which were initially designed to be less sensitive to changes in morphology seem to be also more appropriate to deal with style than Cartesian or Euler data.

5.3. *Application to an interactive game*

In order to evaluate the three HMM recognizers in real conditions, we have designed a simple interactive game. The user was equipped with reflective markers at the same locations than those used for the database. The user was placed in front of a wide screen where three virtual humans are displayed (see Figure 9).

The user has to perform one of the 15 motions studied in this paper without any other constraint. To segment the continuous motion capture flow, we compute the distance between the rest pose and the current one of the subject. When this distance goes beyond a threshold, the user is supposed to perform one of the motions. When this distance returns back below the threshold, we assume that the motion is finished. This simple segmentation algorithm enabled us to start the recognition process after this last event occurred.

Three recognizers have been run concurrently with Cartesian-based, Euler-based and Morphology-Independent features respectively. Once each recognizer has computed the recognized motion, this latter is played by the corresponding virtual human. The user had a wireless mouse close to his hand at rest posture and could give a score to each virtual human: 1 (left-button) if the motion was correctly recognized and 0 (right-button) otherwise. This game was repeated 10 times (10 motions were randomly selected among 15 by the user) and the scores were summed for each virtual human leading to a global score between 0 and 10. The virtual human with the highest score won the game. We selected this application because it was impossible to predict what motion the user decided to perform.

The whole experiment was repeated 10 times. The results indicate a mean recognition rate of 8.8 ± 0.6 for Morphology-Independent features, 7.2 ± 0.8 for Cartesian-based features and 5.0 ± 1.1 for Euler-based features. These results confirm the significantly better recognition rates for Morphology-Independent features (Cochran's Q Test $\chi^2(2, N = 102) = 44.98$, $p < 0.001$).

The recognition process was running on a standard PC with an Intel Core 2

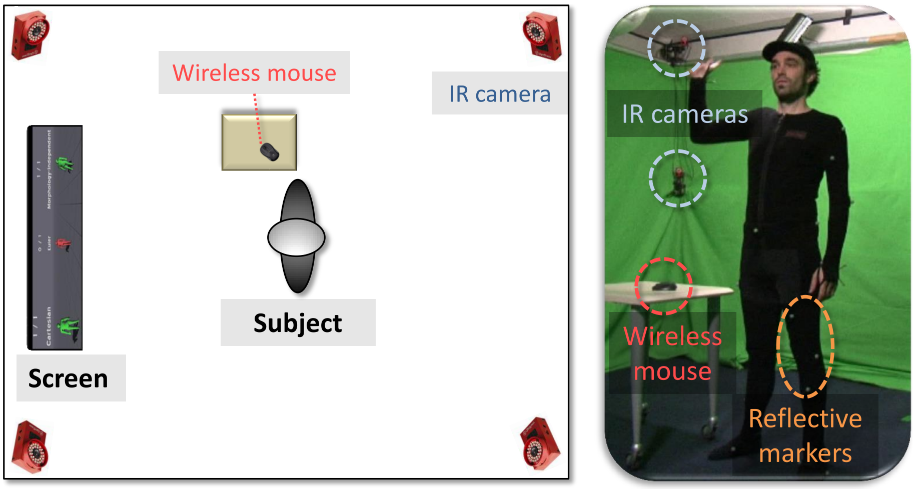


Fig. 9. Experimental set-up for the interactive application.

Quad CPU Q8400 2.66GHz with 4Gb memory. The average computation time is consistently less than 1% of the gesture duration and is thus compatible with interactive applications. Moreover, the computation time for Morphology-Independent data is twice smaller than for Cartesian or Euler ones. Indeed, even if the recognition rate of Morphology-Independent data is significantly better, it is only based on a 12 features while the others are based on 24. This has an obvious impact on the computation time which should remain as small as possible in the context of video games for example (concurrent need of rendering, animation, network, sound, etc.).

6. Conclusion

The main contribution of this paper is to propose and evaluate a new type of generic features for natural human motion recognition. The key idea is that important differences in morphology between users are difficult to deal with, especially for motion recognition systems used for the general public such as Nintendo Wii or Microsoft Kinect. The Morphology-Independent feature allows to tackle this problem at the motion representation stage and to maintain a high recognition rate, even if the HMM system has never been trained with the user's motions. Indeed dealing with new subjects may lead to new morphologies but also to new styles. We clearly demonstrate that the classical Cartesian or Euler-based features fail to address this problem for very similar motion, while using Morphology-Independent data leads to higher recognition rates.

This study has been performed on a database of 3D motions captured with an accurate system. An interesting perspective of this work is to evaluate whether the Morphology-Independent feature is robust to noisy signals measured by less accurate sensors.

This study was limited to the two upper limbs which are involved in most of the applications based on motion recognition. This work could be extended to a full-body representation of the motion, as suggested by Kulpa et al.[20] for computer animation purposes. Further experiments are required to check if this representation is appropriate for full-body motion recognition.

In this paper, the HMMs were chosen because of their popularity in the field of motion recognition. It could be interesting to use the Morphology-Independent features with other types of classifiers, such as Support Vector Machine[13] or Artificial Neural Networks[8]. Recent works tend to use machine learning to identify the most relevant combination of weak classifiers that provides the best recognition rate. These works use a huge amount of features and let the learning process define which ones are relevant. Such systems could also use Morphology-Independent data as basic features to tackle problems due to changes in morphology.

Another challenge of motion recognition systems is to address both segmentation and recognition in a real-time framework. In interactive games, the system indeed cannot wait the end of the user's motion to determine the action he performed and to start the appropriate reaction. Early recognition methods should then be explored. The dimension of the Morphology-Independent feature is smaller than the classical ones and allows faster gesture recognition. It is also impossible to invite the user to move during some imposed time windows. These two problems are key points for future developments. To address this complex problem a first step is to design a method that is robust even when only few data are available. Selecting the relevant features is one step in this direction.

By using more appropriate features it might be possible to recognize a motion with only a few information, such as using the early first frames. It is of great interest for applications which involve recognizing motions as rapidly as possible, such as animating avatars of the user or enabling real-time interaction with a dynamic environment. In such highly time-constrained applications, it is not possible to wait until the end of the user's performance to begin to react. Morphology-independent data seem to provide good recognition rate even for similar motions but new studies have to be carried-out to check if they also enable earlier recognition compared to other types of features. We should also analyze if it can help to segment the motion stream in order to jointly detect and recognize motions in such interactive applications.

References

1. T. K. Bhowmik, J-P. van Oosten, and L. Schomaker. Segmental k-means learning with mixture distribution for hmm based handwriting recognition. In *Proceedings of the 4th international conference on Pattern recognition and machine intelligence*, PReMI'11, pages 432–439, Berlin, Heidelberg, 2011. Springer-Verlag.
2. B. Bodenheimer, C. Rose, S. Rosenthal, and J. Pella. The process of motion capture: Dealing with the data. In *Eurographics Workshop on Computer Animation and Simulation*, pages 3–18, September 1997.

3. L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, and A. Pentland. Invariant features for 3d gesture recognition. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 157–162, 1996.
4. G. Celeux and J-B. Durand. Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564, October 2008.
5. J. Chai and J.K. Hodgins. Performance animation from low-dimensional control signals. *ACM Trans. on Graphics*, 24:686–696, 2005.
6. B. Chakraborty, M. Pedersoli, and J. Gonzàlez. View-invariant human action detection using component-wise hmm of body parts. In *8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08).*, pages 1–6, 2008.
7. K.J. Choi and H.S. Ko. Online motion retargetting. *The Journal Of Visualisation and Computer Animation 2000*, 11(5):223–235, December 2000.
8. A. Corradini and P.R. Cohen. Multimodal speech-gesture interface for handfree painting on a virtual paper using partial recurrent neural networks as gesture recognizer. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'02)*, volume 3, pages 2293–2298. IEEE, 2002.
9. Z. Deng, Q. Gu, and Q. Li. Perceptually consistent example-based human motion retrieval. In *Proceedings of the symposium on Interactive 3D graphics and games (I3D '09)*, pages 191–198, New York, NY, USA, 2009. ACM.
10. M. Gleicher. Retargetting motion to new characters. In *Proc. of ACM SIGGRAPH*, pages 33–42, July 1998.
11. C. Hecker, B. Raabe, R. W. Enslow, J. DeWeese, J. Maynard, and K. van Prooijen. Real-time motion retargeting to highly varied user-created morphologies. *ACM Transactions on Graphics (TOG)*, 27:27:1–27:11, August 2008.
12. P. Hong, T. S. Huang, and M. Turk. Gesture modeling and recognition using finite state machines. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG '00)*, pages 1–6, Washington, DC, USA, 2000. IEEE Computer Society.
13. N. Ikizler and D. Forsyth. Searching for complex human activities with no visual examples. *Int. J. Comput. Vision*, 80:337–357, December 2008.
14. S. Ishigaki, T. White, V.B. Zordan, and C.K. Liu. Performance-based control interface for character animation. *ACM Trans on Graphics*, 28(3):61:1–61:8, 2009.
15. M. P. Johnson, A. Wilson, B. Blumberg, C. Kline, and A. Bobick. Sympathetic interfaces: using a plush toy to direct synthetic characters. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 152–158, New York, NY, USA, 1999. ACM.
16. H. Kim, Y. Lee, and C. Lee. A study on the gesture recognition based on the particle filter. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4692 of *Lecture Notes in Computer Science*, pages 429–438. Springer Berlin / Heidelberg, 2007.
17. L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, 2002.
18. D. Kulic and Y. Nakamura. Comparative study of representations for segmentation of whole body human motion data. In *Proceedings of the IEEE/RSJ international conference on Intelligent Robots and Systems (IROS'09)*, pages 4300–4305, Piscataway, NJ, USA, 2009. IEEE Press.
19. D. Kulic, W. Takano, and Y. Nakamura. Representability of human motions by factorial hidden markov models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2388–2393. IEEE, 2007.
20. R. Kulpa, F. Multon, and B. Arnaldi. Morphology-independent representation of mo-

tions for interactive human-like animation. *Computer Graphics Forum, Eurographics 2005 special issue*, 24(3):343–352, 2005.

21. C. Kwok, D. Fox, and Meila M. Real-time particle filters. *Proceedings of the IEEE*, 92(3):469–484, 2004.
22. X. Liang, Q. Li, X. Zhang, S. Zhang, and W. Geng. Performance-driven motion choreographing with accelerometers. *Computer Animation and Virtual Worlds*, 20:89–99, 2009.
23. X. Liang, S. Zhang, Q. Li, N. Pronost, W. Geng, and F. Multon. Intuitive motion retrieval with motion sensors. In *Proceedings of Computer Graphics International (CGI), Istanbul - Turkey*, jun 2008.
24. W-L. Lu and J.J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *The 3rd Canadian Conference on Computer and Robot Vision.* IEEE, 2006.
25. F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Lecture Notes in Computer Science*, pages 359–372. Springer Berlin / Heidelberg, 2006.
26. T. Molet, R. Boulic, and D. Thalmann. Human motion capture driven by orientation measurements. *Presence*, 8(2):187–203, 1999.
27. M. Muller and T. Roder. Motion templates for automatic classification and retrieval of motion capture data. In *Proc. of Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*, pages 137–146, 2006.
28. P. Natarajan and R. Nevatia. Online, real-time tracking and recognition of human actions. In *IEEE Workshop on Motion and Video Computing (WMVC)*, pages 1–8. IEEE, 2008.
29. P. R. Raamana, D. Grest, and V. Krueger. Human action recognition in table-top scenarios: an hmm-based analysis to optimize the performance. In *Computer Analysis of Images and Patterns*, pages 101–108. Springer, 2007.
30. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
31. A. Ramamoorthy, N. Vaswani, S. Chaudhury, and S. Banerjee. Recognition of dynamic hand gestures. *Pattern Recognition*, 36(9):2069 – 2081, 2003.
32. M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '11)*, pages 147–156, New York, NY, USA, 2011. ACM.
33. M. Romaszewski and P. Glomb. The effect of multiple training sequences on hmm classification of motion capture gesture data. In *Computer Recognition Systems 4*, volume 95 of *Advances in Intelligent and Soft Computing*, pages 365–373. Springer Berlin / Heidelberg, 2011.
34. H.J. Shin, J. Lee, S.Y. Shin, and M. Gleicher. Computer puppetry: An importance-based approach. *ACM Trans. Graph.*, 20(2):67–94, 2001.
35. T. Shiratori and J.K. Hodgins. Accelerometer-based user interfaces for the control of a physically simulated character. *ACM Trans. on Graphics*, 27(5):123:1–123:9, 2008.
36. R. Slyper and J.K. Hodgins. Action capture with accelerometers. In *Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 193–199, 2008.
37. P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.
38. D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Under-*

standing, 115:224–241, February 2011.

39. S. Yousefi, F. A. Kondori, and H. Li. Camera-based gesture tracking for 3d interaction behind mobile devices. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(8), 2012.
40. T. Yu, X. Shen, Q. Li, and W. Geng. Motion retrieval based on movement notation language. *Computer Animation and Virtual Worlds*, 16(3-4):273–282, 2005.