Characterisation of gestural units in light of human-avatar interaction

Ilaria Renna, Sébastien Delacroix, Fanny Catteau, Coralie Vincent, Dominique Boutet

EAI Endorsed Transactions
on Creative Technologies

Research Article ICST.ORG

Characterisation of gestural units in light of human-avatar interaction

I. Renna[1,*], S. Delacroix[2], F. Catteau[1], C. Vincent[1], D. Boutet[1]

[1]Structures Formelles du Langage, UMR 7023 (CNRS / Université Paris 8)
[2]Laboratoire d'Analyse du Mouvement, Institut National de Podologie, Paris

Abstract

We present a method for characterizing coverbal gestural units intended for human-avatar interaction. We recorded 12 gesture types, using a motion-capture system. We used the markers positions thus obtained to determine the gestural units after stroke segmentation. We complement our linguistic analysis of gestures with an elaboration of our biomechanical hypotheses, our method of segmentation, our characterization hypotheses and the results obtained.

Keywords: Gesture units, stroke segmentation, gesture characterization.
Received on 22 May 2014, accepted on 18 September 2014, published on 2 June 2015

doi: 10.4108/ct.2.3.e5

1. Introduction

Characterization of the meaning of gestures is traditionally based on body-oriented descriptions [36] capturing the gestural elements according to a global description of bodily reference points. We aim to show that the meaning of different Gestural Units (GUs) can be defined on the basis of forms along the upper limb, using multiple reference points that are not limited to body-orientation, but oriented via each of the segments (hand, forearm, arm), thus facilitating an automatic characterization of gestures to exploit in a human-avatar context.

This work takes place in the CIGALE project, whose final goal is to create novel human-avatar interactions in the context of theatrical performances starting from the off-line analysis, characterization and classification of 4 different datasets of gestures we recorded using a motion-capture system; one of these datasets is presented (Sec. 3) and exploited in this work. Once the off-line study is completed, the project will move to the on-line analyses in order to evaluate the actual possible interaction between humans and the avatar, whose behavior has been built exploiting the off-line gesture analysis.

After presenting the state of the art from a linguistic and engineering point of view (Sec. 2), we describe the database and the adopted biomechanical model (Sec. 3). Section 4 provides an overview of the linguistic framework within which our semantic characterization of co-verbal gestures is presented. The gesture significant part (*stroke* [31], Sec. 5) segmentation represents a necessary preliminary to such characterization, since it is impossible to characterize the meaning of a gesture without knowing when it occurs. For this purpose, an automatic segmentation is presented and tested against the segmentation of two coders, which serves as ground truth, in line with the highest standard methodes of both domains: robotics and linguistics (Sec. 5). Once this operation is validated, automatic characterization relies on centering with respect to the motion variation of the degree of freedom (DOF) – the prono-supination (Sec. 6). We conclude summing up our results and presenting future work (Sec. 7).

*Corresponding author. Email: ilaria.renna@gmail.com

2. State of the art

Gesture segmentation in concatenated units inspired by Kendon's work [25, 26, 30] is widely used: a Gesture Unit consists of a series of Gesture Phrases which are themselves composed by Gesture Phases. Each latter unit includes the core-meaning of the gesture, named *stroke* (see Sec. 5). Other studies try to set up syntagmatic rules system for movements phases available for both gesture and sign of sign languages [31]. Minor differences exist between these approaches. In the linguistics part of the present study, we adopt Kendon's terminology.

According to their meaning or function, gestures are classified in several categories. McNeill [37], following overall Kendon's classification [27], differentiates gestures in beats, which punctuate the discourse, deictics which includes pointing gestures, iconics which are "images of concrete entities and/or actions" and metaphorics which show "images of the abstract".

Our study concerns gestures which belong to both iconic and metaphoric categories. We adopt Kendon's type gesture of quotable gestures defined as "those standardized gestures which have fairly stable meanings within a given community and which, on the whole, appear to serve in the place of a complete speech act of same sort" [28].

Gestures/actions segmentation is necessary to cut streams of motions into single instances that are consistent to the set of initial model hypotheses and that can be used as training sequences for recognition. In computer vision, different techniques are used to prepare data for gesture recognition and the segmentation concerns image processing methods to extract features [12, 13, 33, 48] to represent the spatial structure of gestures. Such features are then exploited to learn the temporal structure of gestures with different methods, e.g. Hidden Markov Models (HMMs) [7, 20, 50, 54], Baum-Welch [3], parametric models [21, 56] and others.

When temporal segmentation is needed, different kinds of methods can be used to investigate motion profiles and trajectories to recognize human gestures. A general approach for segmenting actions is based on *concatenating action grammars* to model transitions in a gesture or between consecutive gestures [53]. Concatenative grammars can be built, for instance, by joining all models in a common start and end node and by adding a loop-back transition between these two nodes; segmentation and labeling of a complex action sequence is then computed as a minimum-cost path trough the network using dynamic programming techniques. Some works [34, 43] use such networks for action recognition based on HMMs, others [38, 48] on Conditional Random Fields (CRFs) or on semi-Markov models [46].

Another strategy for recognizing gestures consists in dividing video sequences into multiple, overlapping segments, using a *sliding window*; classification is then performed sequentially on all the candidate segments, and peaks in the resulting classification scores are interpreted as gesture locations. Sliding window are used in many template-based representations [17, 59], in combination with dynamic time warping (DTW) [13, 39] and even grammars [4]. For example, Abdelkader et al. [1] propose a template-based approach using DTW to align the different trajectories using elastic geodesic distances on the shape space; the gesture templates are then calculated by averaging the aligned trajectories.

A common strategy is to use a generic segmentation method based on detecting *motion boundaries*, then separately classifying the resulting segments. Such motion boundaries are typically defined as discontinuities and extrema in acceleration, velocity, or curvature of the observed motions. For example, Ogale et al. [40] segment action sequences by detecting minima and maxima of optical flow inside body silhouettes; Zhao et al. [58] calculate velocity and treat local minima in the velocity as gesture boundaries; Wang et al. [51] treat local minima in acceleration as a gesture boundary, allowing them to construct a motion alphabet whose "characters" of this motion are then combined using a HMM; Kahol et.al. [24] tested a user centric gesture segmentation algorithm and developed observer profiles based on how individual users segment motion sequences, encoding gesture boundaries as a binary vector of hierarchically connected body segment activities. Boundary detection methods are attractive because they provide a generic segmentation of the video, which is not dependent on the gestures classes; some precautions

are needed because they are not stable across viewpoints and they are easily confused by the presence of multiple, simultaneous movements.

Movements primitives can also be extracted as joint trajectories using Principal Component Analysis (PCA) [18, 23]. In Lim et al. [32] each movement primitive is represented and stored as a set of joint trajectory basis functions that are then extracted via a PCA of human motion capture data. In [2], gestures computed from inertial sensors are defined by hand paths as a discrete time sequence in the Cartesian space; these are converted to training functional data by basis function expansions using B-splines (curve fitting), and then Functional Principal Component Analysis (FPCA) is performed on all the training data to determine a finite set of functional principal components (FPCs) that explain the modes of variation in the data.

Other sensors than those exploited in computer vision or in motion capture based approaches can be used: in [57], for example, accelerometers and multichannel electromyography (EMG) signals are used for segmentation.

As we want to validate our gesture characterization in an off-line situation, we decided to exploit a simple but robust boundary method for gesture segmentation based on arm movement considerations (see Sec. 5).

3. The dataset

The dataset examined is composed of 91 isolated coverbal symbolic gestures. These coverbal gestures are semantically autonomous and cover all DOF of the upper limb (see 4.1). Some gestures are performed using the entire upper limb, while others employ a subpart only (for example only the fingers) for a total of 150 different gestures reproduced by one person.

3.1. The marker-set and the biomechanical model

Gestures are collected using a 3D motion-capture system of digital infra-red cameras, which ensures the reliability of our understanding of the gesture and the effectiveness of the characterization of avatar motion. The system uses hemispherical reflectors glued to the skin and records their trajectory.

A list of cutaneous markers is established, in order to model the body segments in three dimensions (marker-set of 90 points, Fig. 1). This list references the anatomical positions that should be used in modelling each segment as a rigid body.

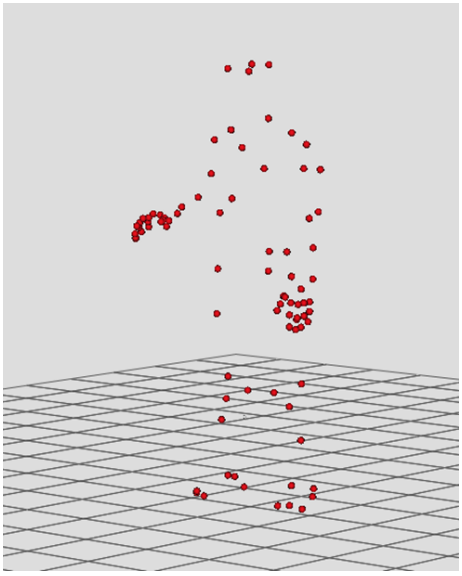

Figure 1. Marker-set visualization.

Generally, three non-aligned anatomical reference points are sufficient to define a segment. In our model (Fig. 2), the torso, arm, forearm, and hand segments have been defined based on coordinates of spatial perspective using a standardized method. This method enables the creation of three orthogonal axes for each system of segment coordinates [15, 55]. It involves calculation of the centers of the wrist, elbow, and shoulder joints, as well as those of the cervical and lumbar regions [15]. For the hand, the origin of the coordinate system is the centre of the wrist joint, Y is the unitary vector connecting the centre of the 2nd and 5th metacarpal heads to the origin, X is the normal unitary vector containing the origin and the 2nd and 5th metacarpal heads, Z is the vector result of axes X and Y. For the forearm, the origin of the coordinate system is the centre of the elbow joint, Y is the unitary vector connecting the centre of the wrist joint to the origin, X is the unitary vector normal to the plane containing the origin and the styloid processes of the ulna and radius, Z is the vector result of axes X and Y. For the arm, the origin of the coordinate system is the centre of the shoulder joint, Y is the unitary vector

connecting the centre of the elbow joint to the origin, X is the unitary vector normal to the plane containing the origin, the epicondyle and the epitrochlea, Z is the vector result of axes X and Y. For the torso, the origin of the coordinate system is the centre of the cervical joint, Y is the unitary vector connecting the lumbar joint to the origin, Z is the unitary vector normal to the plane connecting the origin, the lumbar joint and the suprasternal space, X is the vector result of axes Y and Z.

The coordinate system of each joint is defined through sets of adjacent segment coordinates, allowing the description of the three-dimensional articulation of the shoulder, the elbow, and the wrist at every moment of the gesture. To establish the kinematics of the joints, we used a sequence of successive rotations around the mobile axes, using Euler angles [55]. The dynamic sequence of rotations enables the definition of joint coordinates through the axes of two adjacent segments: one axis for the proximal segment and another for the distal segment; and a floating axis, perpendicular to the other two.

The various joint movements of the wrist, elbow and shoulder are calculated thanks to this biomechanical model, as are the palmar/dorsal flexion and the adduction/abduction of the wrist. These correspond to the flexion/extension and adduction/abduction of the hand as described in the action schemas (Sec. 4). The extension/flexion and supination/pronation of the elbow correspond to the extension/flexion of the forearm and supination/pronation of the hand respectively for the action schemas (see 4.2). Finally, shoulder motion is measured in retropulsion/forward flexion, abduction/adduction and internal/external rotation. These correspond, respectively, to the extension/flexion, abduction/adduction and external/internal rotation of the arm for action schemas.

4. Linguistic redefinition in light of human–avatar interaction

The recorded coverbal gestures match emblematic quote gestures [27, 42], i.e. semantically autonomous gestures, whose significance is independent of the surrounding discourse. The 91 gestures can be

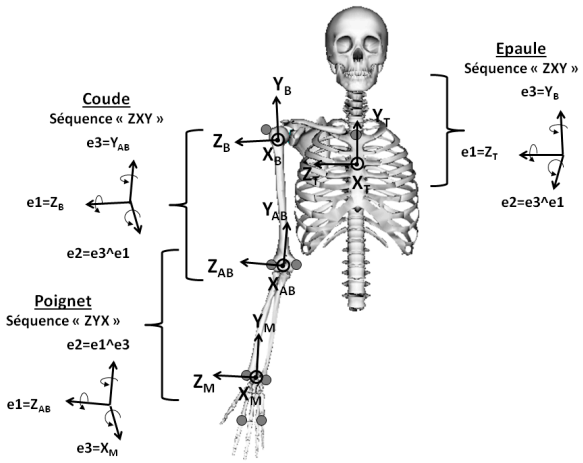

Figure 2. The adopted biomechanical model.

divided into a dozen GUs, with the following senses: reject, refuse, despise, discredit, pass, accept, consider something, consider someone, offer, not care, commit, revere. These semantic labels have been tested and validated with a French-speaking population in a previous study [6].

Each GU corresponds to a particular action schema implementing some (or all) of the segments of the upper limb. Action schemas are based on the motion of various DOF of the segments of the upper limb in a specific order. This order emerges from the difference in the range of motion of each DOF involved in the schema according to its range of motion. Motion is transferred through moments of inertia attached to each DOF and as a function of (involuntary) conjoint movement of the longitudinal axis (exterior/interior rotation or pronation/supination) associated with any joint with two DOF [9, 10, 35]. Thus, for the GU "refuse", for example (Fig. 3), the action diagram shows the hand motion towards the forearm.

4.1. Flow of motion propagation

In the action diagram, the position of the pole of adduction (motion towards the joint on the plane of the palm) determines the direction of motion propagation. If a movement of adduction is in first or second position, the flow of motion is distal-proximal, going from the hand towards the forearm. If adduction is in third position, then it is the result of the first two, and so does not present significant motion.

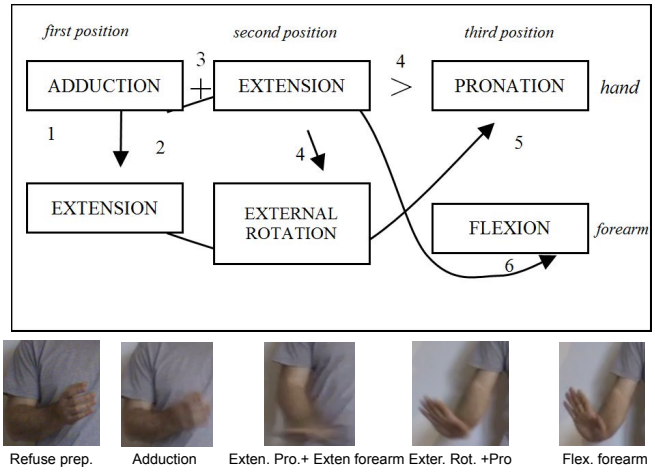

Figure 3. Action schema for the GU "refuse" . This gesture begins with a movement of the hand (Adduction, 1). The order of the motion is numbered. The photograms illustrate the execution of the gesture at different moments of what is gathered in the action schema. The first photogram capture the preparation phase before the stroke.

Therefore, the gesture is initiated on the forearm and spreads towards the hand in a distal-proximal flux. We thus define two types of GUs. The first 8 GUs in the list above are built on the hand while the last 4 (offer, not care, protect, revere) are built on the arm.

4.2. Action schema of hand motions

The sequence of hand motions is based on a structure such that the motion or position of the first two DOF cause involuntary motion of the third DOF. This third motion is either the result of a biomechanical constraint related to motion around the longitudinal axis (pronation/supination), or to a sequence based on the moment of inertia. In both cases, the poles of motion in third position are completely determinable and follow the first two movements such that their sequence affects the pole of the third motion. So, the sequence ADD.EXTEN leads to involuntary PRONATION, while the reverse order, EXTEN.ADD leads to SUPINATION [5, 6].

4.3. Grouping GUs by direction

Tracking the order of the poles in motion is affected by the range of motion, the temporal sequence of the emergence of motion, the initial position and the acceleration, but these criteria, which vary even among themselves, are difficult to hierarchise. On the other hand, it is possible to classify GUs on a formal basis by semantic field (Fig. 4).

Initially, it is necessary to determine the spread of motion; either the gesture starts from the hand and motion goes up the forearm, or it starts in the arm and spreads towards the hand (hand and arm in the diagram). For the hand (Fig. 4, left), the initial prono-supination of the gesture may be marked or unmarked. At the next level, we examine prono-supination with respect to the initial position. This gives us 8 manual action schemas. For the arm (Fig. 4, right), we examine the ADD/ABD position or motion of the arm. Subsequently, the 4 GUs of the arm can be distinguished through prono-supination.

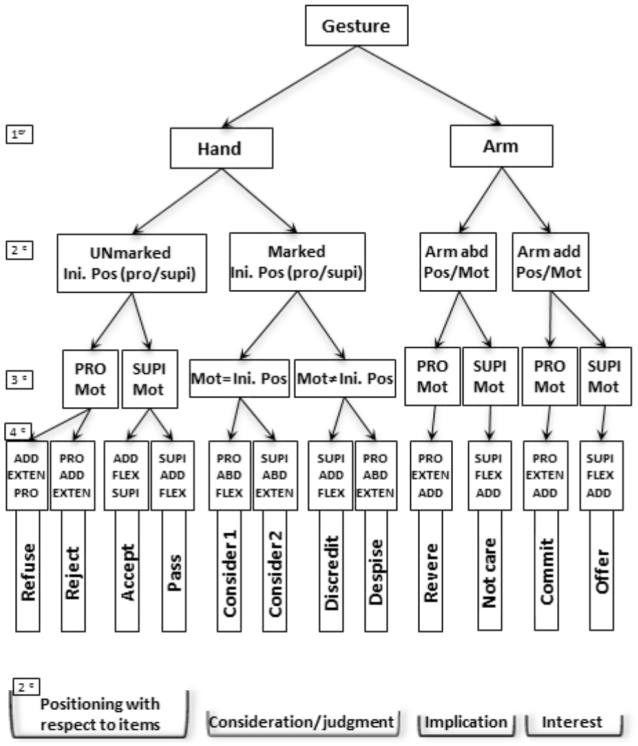

Figure 4. Diagram presenting the formal presentation of gestures according to semantic characterization.

Each of these GUs has a semantic label. A first level of hyperonymic grouping includes 4 semantic sets: i) positioning with respect to items, ii) consideration or judgment, iii) implication, and iv) interest. This semantic level corresponds to the 2nd level of formal disjunction in the diagram. Another possibility is a two part semantic grouping corresponding to the first

formal disjunction (hand or arm): positioning in the world *versus* relational positioning. Therefore, various levels of formal differentiation correspond to specific semantic labeling.

5. Segmentation of gestural signals

Each gestural signal in our database is composed of a sequence: T-pose, gesture, T-pose. Extraction of the gesture requires automatic segmentation. The generally accepted sequence includes 4 phases [11, 36]: 1. resting position; 2. preparation (pre-stroke); 3. core (stroke); 4. retraction (post-stroke). This sequence describes the structure of a gesture. However, it is impossible to find an automatic, objective criterion to extract the stroke, the semantically significant part. This is a complex operation even for a human, and remains uncertain [47].

In our case, segmentation is carried out on the basis of morpho-kinetic properties (as defined by Kendon [29]). Indeed, the preparation of motion consists of a ballistic motion that brings the arm(s) to the core of the motion [8]. This ballistic motion involves acceleration followed by deceleration as the final position is approached, then symmetrical acceleration and deceleration to the first set, and a return to the resting position. T-poses are also characterized by acceleration and deceleration of movement.

In order to extract the stroke of each gesture, we consider the absolute value of the derivative of the Y index positions (seen in all cases as the body part that moves the most): the minimum of this signal represents the transition between acceleration and deceleration. So for automatic segmentation the stroke considered is the part between the minimal phase that precedes the second maximum (property of the beginning of a stroke) and the minimal phase that follows the penultimate maximum (end of a stroke) (Fig. 5). A threshold is set up to avoid minimal and maximal phases due to noise (small adjustment or preparatory motions) from being considered.

5.1. Segmentation evaluation

To evaluate automatic segmentation methods, it is necessary to compare an automatic segmenter's performance against the segmentations produced by human judges (coders). In general, methods for performing this comparison designate as comparison reference only the segmentation of a single coder [44]. However, this approach assumes that the only coder is unbiased and able to provide a perfect segmentation. Indeed, previous works, e.g. [22], showed that inter-annotator agreement between human coders can be rather poor. Thus, an automatic segmenter should be compared directly against different coders [19] to ensure that it does not over-fit to the preference and bias of one particular coder.

Given our dual aim, to evaluate inter-annotator agreement on the one hand and automatic segmentation on the other, we decided to adopt two methods: Accurate Temporal Segmentation Rate (ATSR) [45] and F-score [49]. ATSR is a time-based metric that measures performance in terms of accurately detecting the beginning and end of the stroke for each gesture signal. F-score provides more information than accuracy and enables individuated errors typologies. Three different cases are evaluated with both methods:

1. automatic segmentation is compared with the annotator considered as ground truth (case 1);

2. automatic segmentation is compared to a second annotator (the ground truth) (case 2);

3. the two annotators are compared against one another (case 3).

For each considered gesture, the ATSR was computed as follows: the Absolute Temporal Segmentation Error (ATSE) is evaluated by summing the absolute temporal error between the ground truth and the result of the algorithm for the start and stop event and dividing this sum by the total length of the gesture occurrence measured from the ground truth as formalized in Equation 1. Once the ATSE are calculated, ATSR metrics are computed by subtracting the average ATSE to 1 in order to obtain the accuracy rate as shown in Equation 2. A perfectly accurate segmentation produces an ATSR of 1.

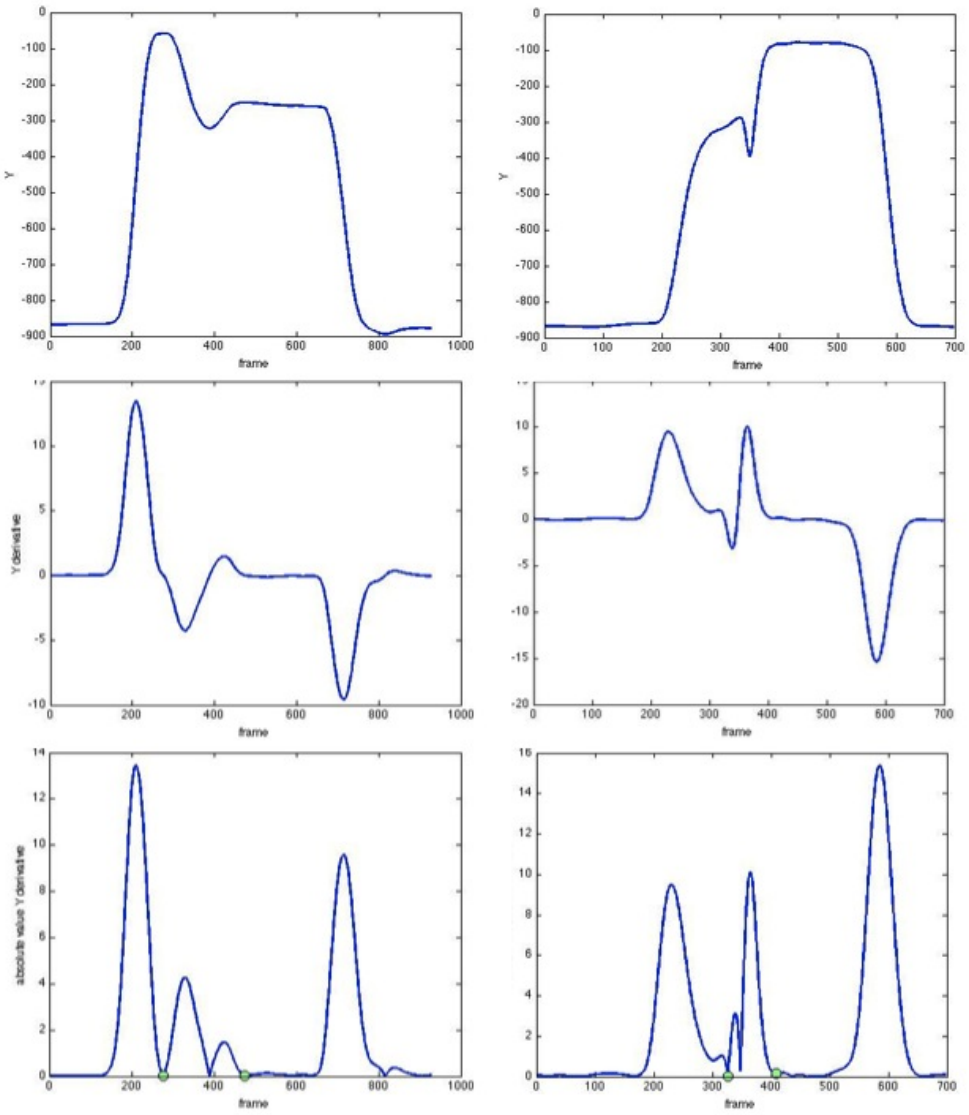

Figure 5. On the left, the signals of the gesture "revere". On the right, the signals for "accept". From top to bottom: positions of the index for Y, speed (derivative) and absolute value of the speed with automatic individuation of the stroke between two green points.

$$ATSE = \frac{|\, Start_{GT} - Start_{Alg} \,| + |\, Stop_{GT} - Stop_{Alg} \,|}{|\, Stop_{GT} - Start_{GT} \,|} \tag{1}$$

$$ATSR = 1 - \frac{1}{n} \sum_{i=1}^{n} ATSE(i)\,. \tag{2}$$

Equation 1 counts differences that occur frame by frame, so an error is taken into account even when annotations differ for just a few frames. To limit this effect and so to avoid small ground truth timing errors producing irrelevant penalties during the computation of the ATSE [45], it is possible to fix a toleration value α so that

$$\text{if } ATSE(i) < \alpha \text{ then } ATSE(i) = 0\,. \tag{3}$$

As a stroke is in general of about 100 frames, we took $\alpha = 0.2$. This corresponds to a global difference of $\alpha * 100 = 20$ frames (around $0.17s$, given the acquisition rate is $120f/s$) compared to the duration of the ground truth, which is an adequate choice considering that, on average, it is easy to have 10-frame-errors for each start or stop. In case 1 we obtain $ATSR = 0.6038$, in case 2 $ATSR = 0.5857$ and in case 3 $ATSR = 0.8707$.

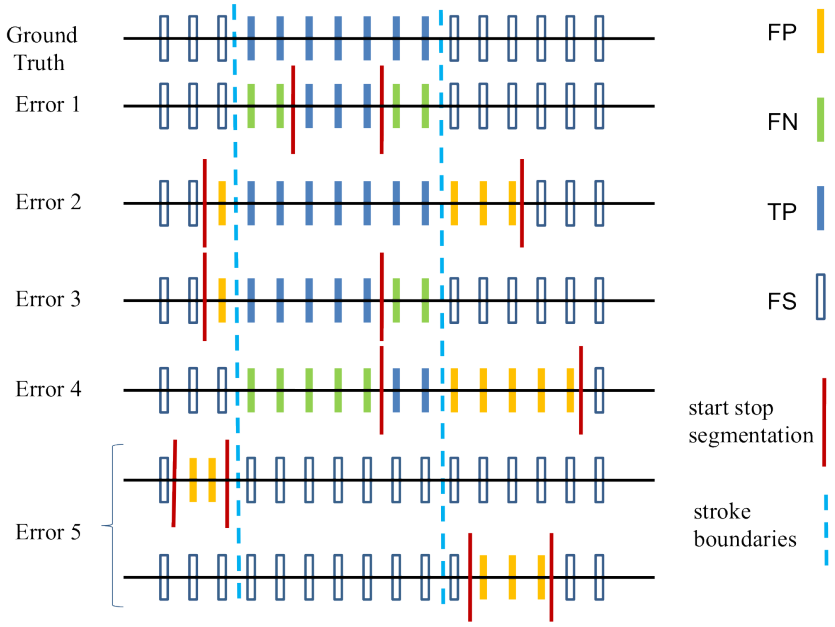

Figure 6. Different possible segmentation errors. Here, FP = False Positive, FN= False Negative, TP= True Positive and FS=outside segmentation.

This kind of method lacks completeness as it does not categorize the errors [52]. In fact, the errors can be categorized into 5 types, as shown in Figure 6.

It is, of course, quite important to know whether the automatic segmentation is wrong but the stroke is preserved (Error 2 in Figure 6) or cut out (all other cases in the figure). In order to assess the quality of our segmentation and of inter-annotator agreement, let us consider precision (p) and recall (r) [16, 41]: precision is the fraction of detections that are true positives rather than false positives (Equation 4), while recall is the fraction of true positives that are detected rather than missed (Equation 5). In probabilistic terms, precision is the probability that detection is valid, and recall is the probability that ground truth data was detected:

$$p = \frac{TP}{TP + FP} \quad (4) \qquad r = \frac{TP}{TP + FN}\,. \quad (5)$$

Precision and Recall can be combined in the F-score as follows:

$$F_\beta = (1 + \beta^2) * \frac{p * r}{\beta^2 p + r}\,. \quad (6)$$

When the parameter $\beta = 1$, F-score is said to be balanced and written as F_1:

$$F_1 = 2 * \frac{p * r}{p + r}\,. \quad (7)$$

The F_1 score can be seen as a weighted average of precision and recall; F_1 score reaches its best value at 1 and worst one at 0. The obtained results are summed up in Table 1. In general, high F_1 values are obtained; r is higher than p in the comparison with automatic segmentation meaning that the algorithm returned most of the relevant results, while p is higher for the inter-annotator agreement: they obtained more relevant than irrelevant agreement. Results concerning the kinds of errors are presented in Table 2.

It is worth highlighting that in Cases 1 and 2, Error 2 is the most frequent. This means that the segmentation method preserves the strokes despite the error. Moreover, the lower error is the cut stroke (Error 1). We can therefore assume that the presented segmentation method is robust to analyze the presented gestures model. For the inter-annotator agreement, we note that most errors stem from one annotator cutting the stroke or because one anticipated the other (Error 3).

Table 1. Results obtained for the three cases of study.

	p	r	F_1
Case 1	0.7430	0.92162	0.8227
Case 2	0.7358	0.9151	0.8157
Case 3	0.9077	0.9053	0.9065

Table 2. Errors occurred in the cases of study.

	Case 1	Case 2	Case 3
Error 1	4	5	17
Error 2	53	49	15
Error 3	24	16	54
Error 4	10	21	5
Error 5	0	0	0

6. Action schema component properties

To characterize the action schemas for each of the coverbal gestures recorded, the segmented signals are transformed into kinematic data in accordance with the biomechanical model (Sec. 3.1). The motion of different degrees of freedom for each joint in the right upper limb (shoulder, elbow, and wrist) is taken into account and normalized temporally on 101 points [14]. Our starting point for this characterization is the assumption that in human-human communication, the motion of DOF is most visible in prono-supination regardless of the type of gesture (performed on the entire upper limb or only one of its segments). We therefore decided to focus the analysis, initially, on the signals that were temporally aligned with the prono-supination zone, which contains the widest variation. Biomechanical parameters, such as initial and final positions of each DOF and their maximal range, were taken into account (Fig.7).

We analyzed gestures involving the motion of all segments of the upper limb (33 of the 91 gestures captured). The stages considered in the decision tree (diagram in Fig. 4) move from i) the 1st node (manual or brachial flow) to ii) the 4th node (separation of gestures by prono-supination).

The first stage involves determination of the flow of motion from the arm (proximal-distal) or from the hand (distal-proximal). We therefore calculated: 1) for the initial position, the moment in which the min. and max. of each DOF appear within the automatically cut stroke; 2) the temporal difference between the min. and the max. of the DOF from one segment to another (arm [offer, not care, commit and revere], forearm and hand [for all the other gestures]). Within the latter calculation, the choice of the min. or max. value for one DOF or another correspond to the initial position, therefore a priori opposed to the pole seen in motion during the stroke. If hand motion is EXTEN (positive value), then the initial position corresponds to a minimum (flexion, negative value). Thus, for example, for the top line of diagram in Fig. 3, which illustrates the poles in motion in the “refuse” gesture, ADD.EXTEN >PRO, the initial position chosen was the max. value of the ADD/ABD, the min. value of the FLEX/EXTEN and the min. value of the SUPI/PRO. We set a minimal threshold of 10 frames, corresponding to 2 running video frames at 25f/s, for a temporal difference that enables the determination of the flow.

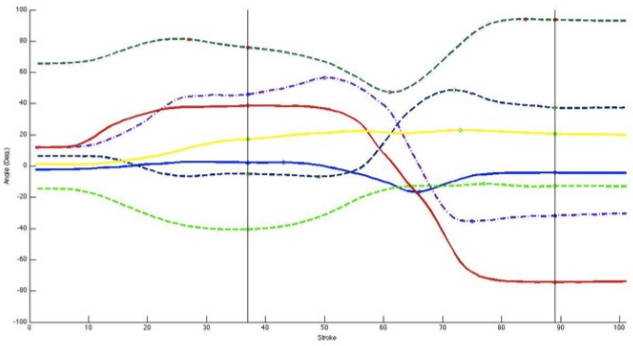

Figure 7. The signals for different segments of the upper limb. For the hand, supination/pronation in red, abduction/adduction in sky blue and dotted purple for flexion/extension. For the forearm, extension/flexion in dotted dark green. For the arm, internal/external rotation in dotted dark blue, abduction/adduction in dotted green and extension/flexion in yellow. Vertical lines indicate the highest variation in prono-supination.

6.1. Discussion

Determination of flow using this method (for the 33 gestures tested, covering the 12 GUs presented in Sec. 4; we underline that each GU occurred between 2 and 3 times), is conformed to expectations in 87.88% of cases. Three of the four cases not validated were below the 10-frame threshold and therefore meet no determinable flow; a further case (a realization of “revere") shows a flow reverse to expectations.

The fourth stage of characterization – wherein poles in motion are used to determine the action schema – was conducted with two types of data (a. and b. below).

Initial calculations concern the average of two or three realizations by GUs (33 gestures in total), thus covering the 12 averaged GUs for which moving poles and semantic labels are already known. We then determine the maximum range for each DOF, either a. within the boundary of prono-supination as shown in Fig. 7, or b. a wider range, starting from the stroke and calculating the difference between the final and initial position of each DOF. We thus obtained the motion poles for all DOF that characterize the GUs, that is 60 DOF for 12 GUs. Results for the first type of data (a. in the demarcation of prono-supination) show a recognition rate of 76.67%. The other option (b. in the stroke with the difference in initial and final position) shows much better characterization ratios 90%. Out of 60 DOF, the opposite pole appears only for 6 expected poles. In both options, an average of 6 DOF measured per GU, the pole that was most prone to error was the hand in ABD/ADD (36% error in a., 67% error in b.). This pole also shows the smallest range (25° and 35°).

Intermediary stages of the characterization (2 and 3 diagram Fig. 4) involve - marking of the initial positions of prono-supination and ABD vs. ADD motion of the arm (stage 2); - determination of the initial position and the identical vs. opposed motion of prono-supination and the motion pole between PRO and SUPI (stage 3).

The characterization of ABD vs ADD of the arm and PRO vs SUPI is unproblematic. In contrast, marking the initial positions of prono-supination (stage 2) does not give the expected results. In this case, only the range difference of the interior/exterior rotation between the beginning and end of the stroke is significant. For a confidence interval of 95%, there is no overlap between "reject/refuse" on the one hand and "despise" on the other. For the trio "pass/accept/discredit" non-overlap was also checked. Thus, the DOF marking criterion (PRO/SUPI) of stage 2 should be modified into a differential of the range of rotation in relatively marked EXT/INT. For step 3, the identity or opposition between the initial position and prono-supination is a good criterion, since, with a confidence interval of 95%, there is no overlap between "reject/refuse/despise" on the one hand, and "consider something" on the other. The same is true between "pass/accept/discredit" on the one hand, and "consider someone" on the other.

All in all, the only phases which are not fully satisfactory are phases 1 (with a single case of inversion for "revere") and 4 (with 90% of expected poles). Intermediary steps are 100% reliable.

7. Conclusion

In this study, we have presented a method for the characterization of 12 gestural units involving the upper limb. A motion-capture system was used to build a reliable gesture database to prove that the meaning of different Gestural Units can be defined on the basis of forms along the upper limb, using multiple reference points, that are not limited to body-orientation, but oriented via each of the segments (hand, forearm, arm). For this purpose an automatic segmentation was exploited. Tests of the segmentation protocol demonstrate its robustness in the individuation of the stroke necessary for the characterization of gestures.

Simple characterization methods fulfill the requirement to associate each stage with formal semantic tagging. This is so since GUs that share the same poles differ only in the sequence in which these poles appear in the action schema. However, we have yet to characterize 4 of these. In this study, we cannot separate 'refuse' from 'reject' and 'accept' from 'pass'. Still both groups can be labeled: on the one hand, negative positioning with respect to things, and on the other, the same type of positioning, only positive. Consequently, all gestures are semantically associated with a variable granularity.

As this characterization has been conducted in light of a human-avatar interaction with an off-line system, the next step is to test the presented method in an on-line set-up using a simpler capture system, namely a kinect, on different subjects reproducing the presented kind of gesture also in other form as, for example, involving only a single moving segment (e.g., the hand).

References

[1] Abdelkader, M.F., Abd-Almageed, W., Srivastava, A. and Chellappa, R. (2011) Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding* **115**(3): 439–455.

[2] Aleotti, J., Cionini, A., Fontanili, L. and Caselli, S. (2013) Arm gesture recognition and humanoid imitation using functional principal component analysis. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on* (IEEE): 3752–3758.

[3] Alon, J., Athitsos, V., Yuan, Q. and Sclaroff, S. (2009) A unified framework for gesture recognition and spatiotemporal gesture segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **31**(9): 1685–1699.

[4] Bobick, A.F. and Ivanov, Y.A. (1998) Action recognition using probabilistic parsing. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on* (IEEE): 196–202.

[5] Boutet, D. (2008) Une morphologie de la gestualité: structuration articulaire. *Cahiers de linguistique analogique* (5): 81–115.

[6] Boutet, D. (2010) Structuration physiologique de la gestuelle: modèle et tests. *Lidil. Revue de linguistique et de didactique des langues* (42): 77–96.

[7] Brand, M., Oliver, N. and Pentland, A. (1997) Coupled hidden markov models for complex action recognition. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (IEEE): 994–999.

[8] Chellali, R. and Renna, I. (2012) Emblematic gestures recognition. In *ASME 2012 11th Biennial Conference on Engineering Systems Design and Analysis* (American Society of Mechanical Engineers): 755–761.

[9] Cheng, P.L. (2006) Simulation of codman's paradox reveals a general law of motion. *Journal of biomechanics* **39**(7): 1201–1207.

[10] Codman, E.A. (1984) *The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa* (RE Kreiger).

[11] Corradini, A. et Cohen, P.R. (2002) Speech-gesture interface for handfree painting on a virtual paper suing partial neural networks as gesture recognizer. In *Proceedings IJCNN'02*: 2293–2298.

[12] Cutler, R. and Turk, M. (1998) View-based interpretation of real-time optical flow for gesture recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE Computer Society): 416–416.

[13] Darrell, T. and Pentland, A. (1993) Space-time gestures. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on* (IEEE): 335–340.

[14] Desroches, G., Dumas, R., Pradon, D., Vaslin, P., Lepoutre, F.X. and Chèze, L. (2010) Upper limb joint dynamics during manual wheelchair propulsion. *Clinical Biomechanics* **25**(4): 299–306.

[15] Dumas, R., Cheze, L. and Verriest, J.P. (2007) Adjustments to mcconville et al. and young et al. body segment inertial parameters. *Journal of biomechanics* **40**(3): 543–553.

[16] Fawcett, T. (2006) An introduction to roc analysis. *Pattern recognition letters* **27**(8): 861–874.

[17] Feng, Z. and Cham, T.J. (2005) Video-based human action classification with ambiguous correspondences. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*: 82–82. doi:10.1109/CVPR.2005.549.

[18] Fod, A., Matarić, M. and Jenkins, O. (2002) Automated derivation of primitives for movement classification. *Autonomous Robots* **12**(1): 39–54. doi:10.1023/A:1013254724861, URL http://dx.doi.org/10.1023/A%3A1013254724861.

[19] Fournier, C. and Inkpen, D. (2012) Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Montréal, Canada: Association for Computational Linguistics): 152–161. URL http://www.aclweb.org/anthology/N/N12/N12-1016.

[20] Françoise, J. (2011) *Realtime Segmentation and Recognition of Gestures using Hierarchical Markov Models*. Master atiam, UPMC - Ircam, Paris. URL http://articles.ircam.fr/textes/Francoise11a/index.pdf.

[21] Gritai, A., Sheikh, Y. and Shah, M. (2004) On the use of anthropometry in the invariant analysis of human actions. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04 (Washington, DC, USA: IEEE Computer Society): 923–926. doi:10.1109/ICPR.2004.652, URL http://dx.doi.org/10.1109/ICPR.2004.652.

[22] Hearst, M.A. (1997) Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.* **23**(1): 33–64. URL http://dl.acm.org/citation.cfm?id=972684.972687.

[23] Jenkins, O.C. and Mataric, M.J. (2002) Deriving action and behavior primitives from human motion. In *In International Conference on Intelligent Robots and Systems*: 2551–2556.

[24] Kahol, K., Tripathi, P. and Panchanathan, S. (2004) Automated gesture segmentation from dance sequences. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*: 883–888. doi:10.1109/AFGR.2004.1301645.

[25] Kendon, A. (1972) Some relationships between body motion and speech. an analysis of an example. *Studies in dyadic communication* .

[26] Kendon, A. (1980) Gesticulation and speech: Two aspects of the process of utterance. *The Relation Between Verbal and Nonverbal Communication* .

[27] Kendon, A. (1988) How gestures can become like words. *Cross-Cultural Perspective in Nonverbal Communication* .

[28] Kendon, A. (1990) Gesticulation, quotable gestures, and signs. *Senri ethnological studies* .

[29] Kendon, A. (1996) An agenda for gesture studies. *Semiotic Review of Books* .

[30] Kendon, A. (2004) *Gesture: Visible Action as Utterance* (Cambridge University Press).

[31] Kita, S., Van Gijn, I. and Van der Hulst, H. (1998) *Gesture and Sign Language in Human-Computer Interaction* (Springer).

[32] Lim, B., Ra, S. and Park, F. (2005) Movement primitives, principal component analysis, and the efficient generation of natural motions. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*: 4630–4635. doi:10.1109/ROBOT.2005.1570834.

[33] Lv, F. and Nevatia, R. (2007) Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*: 1–8. doi:10.1109/CVPR.2007.383131.

[34] Lv, F. and Nevatia, R. (2006) Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV'06 (Berlin, Heidelberg: Springer-Verlag): 359–372. doi:10.1007/11744085_28, URL http://dx.doi.org/10.1007/11744085_28.

[35] MacConaill, M.A. (1948) The movements of bones and joints. *Journal of Bone & Joint Surgery* **30-B**(2): 322–326.

[36] McNeill, D. (1992) *Hand and Mind: What Gestures Reveal about Thought* (University of Chicago Press). URL http://books.google.fr/books?id=3ZZAfNumLvwC.

[37] McNeill, D. (2008) *Gesture and Thought*, Phoenix Poets Series (University of Chicago Press). URL http://books.google.fr/books?id=NOSmyU4TKRwC.

[38] Morency, L.P., Quattoni, A. and Darrell, T. (2007) Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR* (IEEE Computer Society). URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.html#MorencyQD07.

[39] Morguet, P. and Lang, M. (1998) Spotting dynamic hand gestures in video image sequences using hidden markov models. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*: 193–197 vol.3. doi:10.1109/ICIP.1998.999009.

[40] Ogale, A.S., Karapurkar, A., Guerra-filho, G. and Aloimonos, Y. (2004) View-invariant identification of pose sequences for action recognition. In *In VACE*.

[41] Olson, D.L. and Delen, D. (2008) *Advanced Data Mining Techniques* (Springer Publishing Company, Incorporated), 1st ed.

[42] Payrató, L. (1993) A pragmatic view on autonomous gestures: A first repertoire of catalan emblems. *Journal of Pragmatics* **20**(3): 193 – 216. doi:http://dx.doi.org/10.1016/0378-2166(93)90046-R, URL http://www.sciencedirect.com/science/article/pii/037821669390046R.

[43] Peursum, P., Bui, H., Venkatesh, S. and West, G. (2004) Human action segmentation via controlled use of missing data in hmms. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, **4**: 440–445 Vol.4. doi:10.1109/ICPR.2004.1333797.

[44] Preciado, M. (2012) *Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation*. URL http://books.google.fr/books?id=7DvyngEACAAJ.

[45] Ruffieux, S., Lalanne, D. and Mugellini, E. (2013) Chairgest: a challenge for multimodal mid-air gesture recognition for close HCI. In *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*: 483–488.

[46] Shi, Q., Wang, L., Cheng, L. and Smola, A. (2008) Discriminative human action segmentation and recognition using semi-markov model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*: 1–8. doi:10.1109/CVPR.2008.4587557.

[47] Sigalas, M., Baltzakis, H. and Trahanias, P. (2010) Gesture recognition based on arm tracking for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*: 5424–5429. doi:10.1109/IROS.2010.5648870.

[48] Sminchisescu, C., Kanaujia, A., Li, Z. and Metaxas, D. (2005) Conditional models for contextual human motion recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, **2**: 1808–1815 Vol. 2. doi:10.1109/ICCV.2005.59.

[49] Van Rijsbergen, C.J. (1979) Information retrieval. *Butterworth* .

[50] Vogler, C. and Metaxas, D.N. (1999) Parallel hidden markov models for american sign language recognition. In *ICCV*: 116–122. URL http://dblp.uni-trier.de/db/conf/iccv/iccv1999-1.html#VoglerM99.

[51] Wang, T., Shum, H.Y., Xu, Y.Q. and Zheng, N. (2001) Unsupervised analysis of human gestures. In Shum, H.Y., Liao, M. and Chang, S.F. [eds.] *IEEE Pacific Rim Conference on Multimedia* (Springer), *Lecture Notes in Computer Science* **2195**: 174–181. URL http://dblp.uni-trier.de/db/conf/pcm/pcm2001.html#WangSXZ01.

[52] Ward, J.A., Lukowicz, P. and Gellersen, H.W. (2011) Performance metrics for activity recognition. *ACM Trans. Intell. Syst. Technol.* **2**(1): 6:1–6:23. doi:10.1145/1889681.1889687, URL http://doi.acm.org/10.1145/1889681.1889687.

[53] Weinland, D., Ronfard, R. and Boyer, E. (2011) A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **115**(2): 224–241. doi:10.1016/j.cviu.2010.10.002, URL http://dx.doi.org/10.1016/j.cviu.2010.10.002.

[54] Wilson, A. and Bobick, A. (1999) Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **21**(9): 884–900. doi:10.1109/34.790429.

[55] Wu, G., van der Helm, F.C., Veeger, H.D., Makhsous, M., Roy, P.V., Anglin, C., Nagels, J. *et al.* (2005) Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part ii: shoulder, elbow, wrist and hand. *Journal of Biomechanics* **38**(5): 981 – 992. doi:10.1016/j.jbiomech.2004.05.042, URL http://www.sciencedirect.com/science/article/pii/S002192900400301X.

[56] Yacoob, Y. and Black, M.J. (1998) Parameterized modeling and recognition of activities. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98 (Washington, DC, USA: IEEE Computer Society): 120–. URL http://dl.acm.org/citation.cfm?id=938978.939136.

[57] Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K. and Yang, J. (2011) A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Transactions on Systems, Man, and Cybernetics, Part A* **41**(6): 1064–1076. URL http://dblp.uni-trier.de/db/journals/tsmc/tsmca41.html#ZhangCLLWY11.

[58] Zhao, L. (2001) *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures*. Master's thesis, University of Pennsylvania.

[59] Zhong, H., Shi, J. and Visontai, M. (2004) Detecting unusual activity in video. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04 (Washington, DC, USA: IEEE Computer Society): 819–826. URL http://dl.acm.org/citation.cfm?id=1896300.1896418.