Laban movement analysis for action recognition

A. Truong, H. Boujut, T. Zaharia

ARTEMIS Department, Institut Mines-Telecom, Telecom SudParis, CNRS UMR 8145 – MAP5, France.

{arthur.truong, hugo.boujut, titus.zaharia}@telecom-sudparis.eu

Abstract

In this paper, we introduce a new 3D expressive model of gesture descriptors based on the Laban Movement Analysis (LMA). The proposed model is tested and evaluated within an action recognition context. Experimental results, obtained on the Microsoft Research Cambridge-12 dataset, show that the approach yields very high recognition rates (more than 97%).

Introduction

The interpretation of gestures is a stake for numerous applications: e-health, video games, artistic creation, video surveillance... It involves computer vision, machine learning methodologies, and requires the elaboration and development of effective gesture descriptors. Throughout the last decade, several researchers investigated the field of gesture/posture/action recognition. Let us quote some approaches based upon global representation like body angles shaped by limbs positions [1] or spatio-temporal "bouding volumes" [2]. Other approaches are based upon local representations of motion by points saliency, which can be evaluated according to the information contained in spatiotemporal neighborhood [3] or computed in relationship with energy functions putting at stake color, intensity and orientation conspicuities [4].

These different works fail to propose effective descriptors models, and barely take into account the expressive character of body gestures in their analysis. In this context, a key issue concerns the modeling of the various features involved in gestures inter-subjective character, and the testing of these features ability to characterize different types of actions. After its validation, such a general gesture model could then be used for the analysis of other types of high-level contents, like emotions or affects. In this paper, we propose a set of body motion expressive descriptors, inspired from the LMA model [5] and test their relevance in action recognition context.

Proposed approach

The Laban Movement Analysis (LMA [5]) model, introduced by Rudolf Laban consists of describing body movement in terms of qualities relating to different characterizations of the way this movement is performed. More precisely, the LMA model is composed of the following five major qualities: *Body*, *Space*, *Effort*, *Shape*, and *Relationship*. For this study, we only considered the *Space*, *Effor*t and *Shape* components, for they relate to expressivity and communicational aspects (The *Body* quality deals with the structural aspect of the gesture and the *Relationship* quality is better suited for group analysis). The *Space* component refers to the place, direction and path of the movement. The *Effort* component depicts how the body concentrates its effort to perform the movement and deals with expressivity and style. The *Effort* further includes the following sub-components: *Space* (not to be confused with *Space* quality) that we did not retain in our work because of its redundancy with the generic *Space* component; *Time* which separates movements between sudden and sustained (or continuous) ones; *Flow* which describes movements as free or constrained; *Weight*, which aims at distinguishing between light and heavy movements. The *Shape* component includes 3 factors: *Shape Flow*, which relates to participant's concern about changing relationships between its different body parts; *Directional movement*, which describes the direction of the movement toward a particular point; *Shaping*, characterizing how the body changes its shape in a particular direction: rising/sinking, retreating/advancing and enclosing/spreading, which will be retained for our model for it is the most concrete *Shape* sub-quality.

Proceedings of Measuring Behavior 2014, (Wageningen, The Netherlands, August 27-29, 2014).
Editors: A.J. Spink, L.W.S. Loijens, M. Woloszynowska-Fraser & L.P.J.J. Noldus. www.measuringbehavior.org

Some first studies like [6] attempted to identify and classify cetain Laban qualities, based on visual gesture descriptors, with the help of supervised classification approaches, showing that a mid-level Laban representation can be obtained starting from visual descriptors. Our objective is slightly different: define a set of descriptors characterizing the individual Laban qualities that we considered relevant (i.e. *Space*, *Effort* and *Shape*), and exploit them directly for action recognition purposes, without explicitly determining the Laban qualities.

The proposed descriptors are associated to 3D body skeleton joints trajectories that can be recorded with a depth sensor (*e.g.*., Kinect camera) at a rate of 30 frames per second. The Kinect sensor provides a maximum number of 20 joints (Figure 1.a). For each body joint *i*, the trajectory will be represented as a sequence $(x_{i,t}, y_{i,t}, z_{i,t})_{t=0}^{N-1}$ where N is the total number of frames. For all the joints, $(x_{i,t}^{trans}, y_{i,t}^{trans}, z_{i,t}^{trans})_{t=0}^{N-1}$ will refer to a trajectory transform consisting in moving the body at every time t so that: a) the shoulders and the hip center belong to a same plane parallel to (yOz) plane; b) both shoulders are at the same height. Figure 1.b illustrates the result of this body alignment process, so that (xOy), (yOz) and (zOx) planes respectively correspond to sagittal, vertical and horizontal body planes. The approach leads to a total number of 81 features we are now introducing.

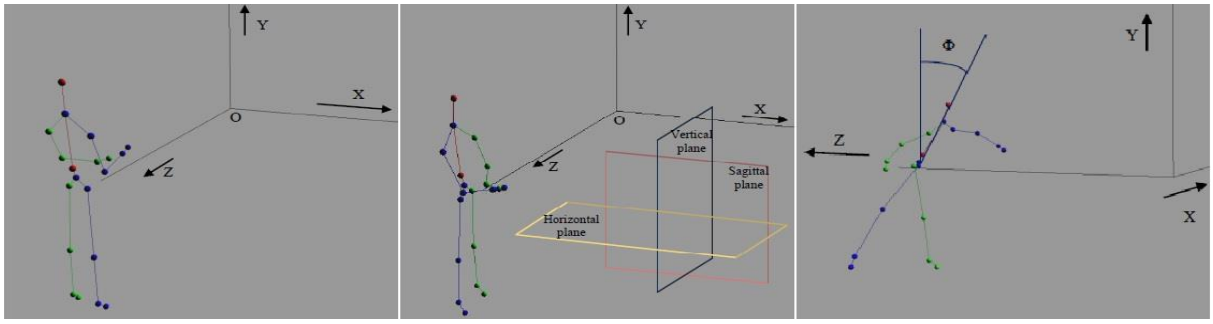

Figure 1 Body skeleton joints at a particular frame (a); skeleton joints new positions after transforms application and body planes representation (b); illustration of forward tilt angle (c).

For the *Space* quality (9 features), the first characteristic is the total length of the head trajectory. Then, we retain the number of zero crossings $(n_{ZC}(x'^{trans}_{Head,t})_{t=0}^{N-1})$ of the first derivative of the head's component in the direction parallel to the vertical plane (which measures the number of head's retreats/advances), the maximal amplitude of the head movement following this direction $(max_{t=0}^{N-1}(|x_{Head,t}^{trans}|) - min_{t=0}^{N-1}(|x_{Head,t}^{trans}|))$ and the relative temporal instant of maximum's reaching: t_{max}/N. Then, we consider the forward tilt angle Φ defined for each frame as the angle between the vertical direction *y* and the axis binding the center of the hip and the head, expressed in radians (Figure 1.c). The resulting tilt angle sequence is described by the following 5 parameters: mean, standard deviation, ratio between the global minimum and maximum values, number of local maxima and relative temporal instant of the global maximum.

For the *Time* subcomponent of the *Effort* quality (8 features), we firstly consider the total gesture duration (number of frames). Then, we compute features based on the kinetic energy sequence defined for head and left and right hands. The frames of the sequence with kinetic energy inferior to 1/10 of the maximal energy reached throughout the gesture are considered as pauses, the others as high/medium activity frames. We compute the percentage of low activity frames relatively to the whole sequence, as well as the mean, standard deviation and maximum value of the kinetic energy sub-sequences for both high/medium and low activity components.

The *Flow* subcomponent of the *Effort* quality (10 features) is described with the help of the third order derivative of the left and right hands trajectories, so-called jerk. The 2 jerk series are statistically described by the following entities: mean, standard deviation, ratio between the maximal and mean values, number of local maxima and the relative temporal instant of the global maximum.

For the *Weight* subcomponent of *Effort* quality (30 features), we consider the vertical components of the velocity and acceleration sequences (*i.e.*, $y'_{.,t}$ and $y''_{.,t}$ signals) associated to 3 joints: the center of the hip, the left and the right hand. The following 5 features are retained: mean, standard deviation, maximal amplitude, number of local

minima, and relative temporal instant of the global minimum value. Such an approach makes it possible to characterize the vertical motion of the gesture sequence.

Finally, for the *Shaping* sub-quality (24 features), a first part of the description aims at globally characterizing the spatial dissymmetry between the two hands over the whole sequence. For each frame, we associate a dissymmetry measure defined as: $Dys = \frac{d_{left,\,center}}{d_{left,\,center}+d_{right,\,center}}$, where $d_{left/right,\,center}$ denotes the distance between the left/right hand and the center of the shoulders. The Dys_t sequence is globally described by 6 parameters: mean, standard deviation, global maximum/mean ratio, global minimum/mean ratio, number of local extrema, and the relative temporal position of the global extremum. In addition, the same parameters are associated to 3 other sequences, respectively associated to global body amplitudes in the directions perpendicular to vertical, horizontal and sagittal planes (Figure 1.b), respectively denoted by A^x, A^y, and A^z and defined as $A_t^d = (max_i(\left|d_{i,t}^{trans}\right|) - min_i(\left|d_{i,t}^{trans}\right|))_{t=0}^{N-1}$, d ∈ {x; y; z} where *i* indexes the skeleton joints.

Table I Mean F1 scores (in %) obtained for the various classification method retained.

Gesture	SVC	Extra Trees
Iconic	98.8	99.5
Metaphoric	98.2	99.2
All gestures	97.2	98.7

Experimental results

We tested our expressive model on Microsoft Research Cambridge-12 (MSRC-12 Figure 2) dataset [7], which is publicly available and provides two different types of gesture captured by a Kinect camera: 6 gestures are *iconic* and represent actions/objects and 6 others are *metaphoric* (more related to abstract concepts). As proposed in [8], we have first performed the recognition separately on gestures of each given type (*iconic*, *metaphoric*). Then, we have considered globally all the 12 gesture categories, in order to test for scalability. For each of these 3 different recognition runs, two different methods were compared. The first one uses *Support Vector Classifiers* [9] with the *one versus one* strategy: there is one classifier for each pair of classes and at the testing stage, the class collecting the highest number of classifier votes, is retained. The second one is the *Extremely Randomized Trees* (Extra Trees [10]): at each tree node, a component of the feature vector and a threshold are randomly selected and split in left and right child nodes. When all leafs contain only one training sample the class is associated to and the process is stopped. At prediction time, the feature vector is processed by all the trees, and the class collecting the highest number of votes is retained. We use the F-score as performance measure: $F1\ score = 2\frac{precision\ .recall}{precision+recall}$. We have applied a 5-fold cross-validation scheme, *i.e*. with a training/testing ratio of 80%/20% and 5 cross-validation steps, by splitting the data into 5 blocks preserving the initial class distribution.

The classification results are summarized in Table I. They correspond to the average F-scores obtained on all gesture classes involved (iconic, metaphoric, all). The results correspond to the best obtained among different parameters combinations associated with each classification strategy (SVC or Extra Trees) and obtained through an optimization strategy. The mean F-measures obtained are in all cases superior 97%, whatever the classification strategy involved. The testing of the proposed model on MSRC-12 dataset gives better results than any other approach having privileged the common use of joints positions or velocities sequences and having put aside the semantic interpretation of gesture in the designing of its descriptors [8] [11] [12]. This demonstrates the capacity of our descriptors to efficiently capture relevant motion indices for action recognition purpose.

Figure 2 Examples of motions tracked for MSRC-12 datasets constitution.

Conclusion and perspectives

In this paper, we introduced a gesture description approach, based on a set of descriptors dedicated to various entities defined in the LMA model. Our perspectives of future work concern the research of other descriptors still inspired by considerations on expressivity, their dimensional reduction to 2D context in order to test them on numerous available 2D video datasets and their application to other types of gestural contents.

References

1. Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen, and Suh-Yin Lee, "Human action recognition using star skeleton," in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*.: ACM, 2006, pp. 171-178.

2. Vili Kellokumpu, Guoyin Zhao, and Matti Pietikäinen, "Human activity recognition using a dynamic texture based method," in *BMVC*., 2008, pp. 1-10.

3. Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 3, pp. 710-719, 2005.

4. Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*.: IEEE, 2009, pp. 1454-1461.

5. Rudolf Laban, *La Maîtrise du Mouvement*. Arles: Actes Sud, 1994.

6. Liwei Zhao and Norman I. Badler, "Acquiring and validating motion qualities from live limb gestures," *Graphical Models*, vol. 67, no. 1, pp. 1-16, 2005.

7. Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*.: ACM, 2012, pp. 1737-1746.

8. Yale Song, Louis-Philippe Morency, and Randall Davis, "Distribution-Sensitive Learning for Imbalanced Datasets," in *2013 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2013)*.: IEEE, 2013.

9. Johan A.K. Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293-300, 1999.

10. Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3-42, 2006.

11. Farhood Negin, Firat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil, "A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras," in *Image Analysis and Recognition*.: Springer, 2013, pp. 648-657.

12. Xinbo Jiang, Fan Zhong, Qunsheng Peng, and Xueying Qin, "Online robust action recognition based on a hierarchical model," *The Visual Computer*, pp. 1-13, 2014.