Full body expressivity analysis in 3D Natural Interaction: a comparative study*

George Caridakis †
Intelligent Systems, Content and Interaction Lab
National Technical University of Athens, Greece
Iroon Polytexneiou 9, 15780 Zografou, Greece
gcari@image.ntua.gr

Kostas Karpouzis
Intelligent Systems, Content and Interaction Lab
National Technical University of Athens, Greece
Iroon Polytexneiou 9, 15780 Zografou, Greece
kkarpou@image.ntua.gr

ABSTRACT
Current article presents preliminary research work on defining and extracting full body expressivity features within the framework of using natural interaction in games and game based learning. Behavior expressiveness is an integral part of the communication process since it can provide information on the current emotional state, the personality of the interlocutor and his performance when the aim of the interaction is measurable. Many researchers have studied characteristics of human movement and coded them in binary categories such as slow/fast, restricted/wide, weak/strong, small/big, unpleasant/pleasant in order to properly model expressivity. Expressivity dimensions are selected as the most complete approach to body expressivity modeling, since they cover the entire spectrum of expressivity parameters related to emotion and affect. Derived from the field of expressivity synthesis five parameters have been computationally defined following different approaches and comparison of these approaches aims to investigate the most suitable for representing each expressivity feature.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms
Affective Computing

Keywords
Natural Interaction, Body expressivity

1. INTRODUCTION
An abundance of research within the fields of psychology and cognitive science related with the non verbal behavior and communication stress out the importance of qualitative expressive characteristics of body motion, posture, gestures and in general human action during an interaction session [15]. Although such research work study primarily and mainly context of human to human interaction such approach can be extended to human computer interaction. Some work has incorporated gesture expressivity in HCI context but the vast majority concentrates on the expressively enhanced synthesis of gestures by virtual agents and ECAs [11]. Currently, research on the automatic analysis of gesture expressivity is still immature and this fold of human action analysis is asymmetrically studied with reference to the synthesis counterpart.

Human Computer Interaction continuously introduces new means of communication and interaction with systems [6]. Alternative to conventional means of interaction, Natural Interaction is increasingly attracting the attention of researchers in related research areas. Within Natural Interaction context body actions, movement and postures, either intentional or not, convey important emotional content, enhanced with qualitative expressive cues. Body motion or posture qualitative aspects (formulated using different approaches) communicate affective and emotional content and are embodied in the direct and natural emotional expression of body movement [7].

Non verbal behavioral cues are by definition connected to alternative means of interaction such as NI. An abundance of research within the fields of psychology and cognitive science related with non verbal behavior and communication stress out the importance of qualitative expressive characteristics of body motion, posture, gestures and in general human action during an interaction session. Adaptation of interfaces and content according to the user's affective state is a requirement for successful and friendly interaction [12] .

Expressivity of body movement is a qualitative cue that is, or at least should be, incorporated in the design process of such applications. Alex Pentland [16] wrote in the Scientific American: "The problem, in my opinion, is that our current computers are both deaf and blind: they experience the world only by way of a keyboard and a mouse. ...I believe computers must be able to see and hear what we do before they can prove truly helpful". Moving a step further, we might add, that they should also interpret appropriately what they see and hear. Behavior expressiveness is an integral part of the communication process since it can provide information on the current emotional state, the personality of the interlocutor and his performance when the aim of the interaction is measurable. Many researchers have studied characteristics of human movement and coded them in binary categories such as slow/fast, restricted/wide, weak/strong, small/big, unpleasant/pleasant in order to properly model expressivity. Expressivity dimensions are selected as the most complete approach to body expressivity modeling, since they cover the entire spectrum of expressivity param-

†Corresponding author

eters related to emotion and affect [13]. Current article presents preliminary research work on defining and extracting full body expressivity features within the framework of non verbal behavioral cues in NI.

2. RELATED WORK

Within the wider research area of Affective computing, research has been performed towards gesture or body interaction analysis and related articles can be found both on the IEEE Transactions on Affective Computing (TAC) as well as in the two books that have been recently published [17] and [18] and deal with the entire spectrum of research related to Affective Computing. Investigating though Natural Interaction in three dimensions and performing comparative studies regarding full body expressivity formalisation, remains a scarcely studied domain, although some research work has been performed recently on actor portrayals corpus [9].

Affective analysis, aiming to classify interaction segments into emotions based on gestures or body information, has been proposed [1], [14] and [8]. Additionally, such information has been fused with modalities used widely in Affective computing such as facial expressions and speech prosody [10], [3] and [19].

Similarly, as discussed in the introductory section, mimicry of human behavior by or behavior adaptation of Virtual Agents or Embodied Conversational Agents has also been significantly studied [4], [2] aiming to enhance interaction and design believable agents based on gestural or bodily qualitative cues.

3. 3D FULL BODY EXPRESSIVITY MODELLING APPROACHES

To model expressivity, in our work, we use the six dimensions of behavior [5], as a more accomplished way to describe the expressivity, since it tackles all the parameters of expression of emotion [20]. Five parameters modeling behavior expressivity have been defined at the analysis level, as a subset of features derived from the field of expressivity synthesis:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power

The ultimate goal is to formulate each full body expressivity feature using one of the approaches described below. Initially a body pose P is formally defined as a sequence, of T frames $i \in [1, T]$, consisting of

$$P = [\vec{l}, \vec{r}, S, D, F, J]$$

:

- 3D coordinates of the left and right hand

$$\vec{l} = (x_l, y_l, z_l)$$

$$\vec{r} = (x_r, y_r, z_r)$$

- S silhouette binary image
- D depth image map
- F face information

$$F = [p, d, z]$$

, $p\ position, d\ diagonal\ size, z\ depth$

- J skeleton joints for left/right arm J_l/J_r
  - shoulder
  - elbow
  - hip
  - knee

Given the above definition of pose, expressivity features are formulated using different approaches, namely based on:

(a) silhouette

(b) limbs

(c) joints

Although silhouette is usually used in full body expressivity analysis, as discussed in section 2, limb based expressivity formalisation presents interest since it has been used before in half-body, desktop interaction context. One could argue that limb based analysis is a subcase of silhouette based one but on the other hand extracting features or points/regions of interest using computer vision and image processing techniques is entirely a different issue. Silhouette extraction is a trivial task for fixed background and feasible when depth information is available. Limb, actually limb's end effectors, detection and tracking, especially for the case of skin colored hands could be applied to wider range of applications and interaction contexts. Finally, joint expressivity formalisation is quite innovative since robustly extracting relative features is an extremely challenging task and researchers opted to simpler and more robust approaches.

Overall activation is considered as the quantity of movement during a dialogic discourse.

(a) For a given time window of w frames define fading silhouette motion volumes $FSMV$ adding a degrading weight depending on time and volume:

$$FSMV_t = ((\sum_{i=1}^{w} \frac{w-i}{w} S_{t-i}) - S_t)(|D_t - D_{t-w}|)$$

The general equation of silhouette based overall activation would be:

$$OA = \frac{volume\ of\ motion}{volume\ of\ silhouette}$$

or better defined as:

$$\frac{FSMV}{SV}$$

$$FSMV = \sum_{t=1}^{T} FSMV_t$$

$$SV = \sum_{i=1}^{T} S_i D_i)$$

SV being a normalization factor for distance and size invariant results.

(b) limb based OA defined as:

$$OA = \sum_{i=1}^{T} \left|\vec{rh_i} - \vec{rh_{i-1}}\right| + \left|\vec{lh_i} - \vec{lh_{i-1}}\right| + \left|\vec{rf_i} - \vec{rf_{i-1}}\right| + \left|\vec{lf_i} - \vec{lf_{i-1}}\right|$$

(c) weighted sum of joints rotations derivative:

$$OA = W_1(J'_{ls} + J'_{lh} + J'_{rs} + J'_{rh}) + W_2(J'_{le} + J'_{lk} + J'_{re} + J'_{rk})$$

$$s = shoulder\ e = elbow\ h = hip\ k = knee$$

Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). Let SE_0 be the spatial extent (according to each definition) of the neutral/calibration position.

(a) 2D silhouette based:

  (a) max and median of area of polygon consisting of left hand, head, right hand, right foot, left foot normalised by SE_0

  (b) max and median of sum of diagonals of Quadrilateral consisting of right hand/left foot and left hand/right foot normalised by SE_0

(b) limb based is already included in silhouette based

(c) joint based does not make sense since rotation is independent of spatial extent

Fluidity differentiates smooth / elegant from the sudden / abrupt gestures. This concept attempts to denote the continuity between movements. It is formally defined as the variance of the OA as described previously:

$$FL = Var(\frac{FSMV}{SV_0})$$

Please note that the quantity FL corresponds to is reversely proportional to the notion of fluidity. Thus, a motion with high value of FL expressive parameter demonstrates low fluidity and consequently is categorized as a sudden/abrupt movement. Inverting the definition of fluidity is not a trivial process since the upper and lower bound of the measure are not a priori known.

Temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures:

$$TE = \frac{mean(FSMV)}{SV_0}$$

$$SV_0 = S_0 D_0$$

again SV_0 as SE_0 is a normalizing factor

Power is associated qualitatively with the first derivative of which refers to acceleration:

$$PO = FSMV^{'}$$

4. EXPERIMENTAL VALIDATION

Initially, a preliminary dataset was constructed by recording four users while performing variants of movements using Microsoft's Kinect. Since this study aims to investigate the optimal approach to computationally formulate body expressivity, the dataset was constructed based on acted extreme expressions. Once validated it is only natural to extend the dataset onto natural or naturalistic expressions during gaming or other interaction contexts that include whole body movement, as will be discussed in section 5. During recordings the subjects were asked to perform two body movements per expressivity feature corresponding to their interpretation of maximum and minimum value. Prior to the recordings the subjects were introduced to the adopted classification scheme. Instances of the recording process can be shown below. Expressivity features have been extracted using the above definitions and for each feature the optimal formalisation will be selected based on its dissociation capabilities.

Figure 1: Different instances of user during recordings

Initially, S silhouette binary image and D depth image map were calculated as as described in Section 3 and shown in Figure 2. This input is used for silhouette based full body expressivity formalisation.

Figure 2: Depth and silhouette images provided by Kinect

Additionally, J_l/J_r skeleton joints were calculated for left/right shoulder, elbow, hip and knee. The rotations formed J described in Section 3 and used to model expressivity using joint rotations and skeleton and fused representations are shown in Figures 3 and 4 respectively.

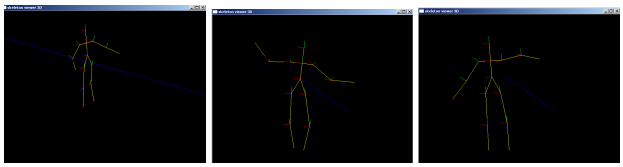

Figure 3: Skeleton (mirrored) representation using calculated joint rotations

Limb based expressivity formalisation is more straightforward and relies only on the end effectors of the kinematic chains of the upper and lower limbs.

5. CONCLUSIONS AND FUTURE WORK

Current article presented preliminary research work on defining and extracting full body expressivity features within the framework

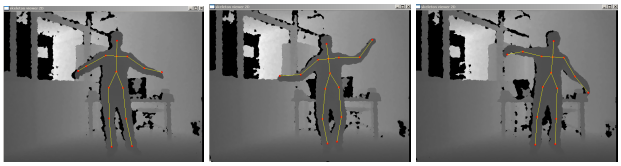

Figure 4: Fused depth and skeleton information images

of using natural interaction. Behavior expressiveness is an integral part of the communication process since it can provide information on the current emotional state, the personality of the interlocutor and his performance when the aim of the interaction is measurable. Regarding ingoing and future work, we are working on, initially experimentally, investigating the validity of each approach on an acted dataset of extreme and isolated body expressions. This research direction will be further validated on naturalistic user behaviour both during different interaction context. Finally, appropriate ways, and hopefully an integrated architecture, to incorporate extracted expressivity features in game scenario or agent behavior adaptation will constitute a challenging future research direction.

6. ACKNOWLEDGEMENTS

This work has been partially funded by the European Commission, Information Society and Media Directorate General, project SIREN - Social games for conflIct REsolution based on natural iNteraction, 258453 - FP7-ICT-2009-5, ICT-2009.4.2 Technology Enhanced Learning.

7. REFERENCES

[1] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. *Affective Computing and Intelligent Interaction*, pages 59–70, 2007.

[2] B. Brandherm, H. Prendinger, and M. Ishizuka. Dynamic bayesian network based interest estimation for visual attentive presentation agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 191–198. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[3] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. *Artificial Intelligence and Innovations 2007: from Theory to Applications*, pages 375–388, 2007.

[4] G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation*, 41(3):367–388, 2007.

[5] G. Caridakis, A. Raouzaiou, K. Karpouzis, and S. Kollias. Synthesizing gesture expressivity based on real sequences. In *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 Conference, Genoa, Italy*, pages 24–26. Citeseer, 2006.

[6] G. Castellano, G. Caridakis, A. Camurri, K. Karpouzis, G. Volpe, and S. Kollias. *A Blueprint for Affective Computing*, chapter Body gesture and facial expression analysis for automatic affect recognition. Oxford University Press, 2010.

[7] G. Castellano, M. Mancini, C. Peters, and P. McOwan. Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man and Cybernetics, Part A - Systems and Humans*, 2011.

[8] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, pages 103–119, 2008.

[9] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Towards a minimal representation of affective gestures. *Affective Computing, IEEE Transactions on*, PP(99):1, 2011.

[10] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):64–84, 2009.

[11] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 1095–1096. ACM, 2005.

[12] E. Hudlicka and M. Mcneese. Assessment of user affective and belief states for interface adaptation: Application to an air force pilot task. *User Modeling and User-Adapted Interaction*, 12(1):1–47, 2002.

[13] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. *Artifical Intelligence for Human Computing*, pages 91–112, 2007.

[14] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. *Affective Computing and Intelligent Interaction*, pages 48–58, 2007.

[15] M. Knapp and J. Hall. *Nonverbal communication in human interaction*. Wadsworth Pub Co, 2009.

[16] A. Pentland. Smart rooms. *Scientific American*, 274(4):54–62, 1996.

[17] P. Petta, C. Pelachaud, and R. Cowie, editors. *Emotion-Oriented Systems, The Humaine Handbook*. Springer, Series: Cognitive Technologies, February, 2011.

[18] K. R. Scherer, T. Banziger, and E. Roesch, editors. *A Blueprint for Affective Computing, A sourcebook and manual*. Oxford University Press, November, 2010.

[19] M. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

[20] H. Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.