2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization

Natural interaction multimodal analysis: Expressivity analysis towards adaptive and personalized interfaces

Stylianos Asteriadis, George Caridakis, Lori Malatesta, Kostas Karpouzis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
Athens, Greece
Email: stiast, gcari, lori, kkarpou@image.ntua.gr

***Abstract*—Intelligent personalized systems often ignore the affective aspect of human behavior and focus more on tactile cues of the user activity. A complete user modelling, though, should also incorporate cues such as facial expressions, speech prosody and gesture or body posture expressivity features, in order to dynamically profile the user, fusing all available modalities since these qualitative affective cues contain significant information about the user's non verbal behavior and communication. Towards this direction, this work focuses on automatic extraction of gestural and head expressivity features and related statistical processing. The perspective of adopting a common formalization of using expressivity features for a multitude of visual, emotional modalities is explored and grounded through an overview of experiments on appropriate corpora and the corresponding analysis.**

***Keywords*-Emotion Estimation; Gesture Recognition; Expressivity Features; Activity Recognition;**

I. Introduction

User modelling and behavior recognition during Human-Machine interactions is attracting more and more attention in bibliography. There are a lot of aspects where the issue can be approached from: Keeping track of user performance, history, profile details are among the simplest and most classical ways of fine-tuning an interaction scenario (e.g. a game) in order to adapt itself to current user needs. However, recent advances in technology have given rise to new add-ons regarding ways to create content according to user models and needs. The use of physiological signals, gaze recognition and motion expressivity understanding [1], [2], [3] constitute valuable modalities for understanding user emotions and cognitive states. In this way, intentions, needs and capacities can be modelled in relation to performance and skills, creating user clusters and control methodologies for guiding adaptation. Recent bibliography has showcased that the expected outcome (eg. learning effect, play performance) can be maximized [4] by following an affect-dependent content generation strategy.

Neuroscientific and psychological studies have revealed that body movement and its expressivity is an important modality of emotion communication [5] and [6].Within this view, postprocessing of affect-dependent non-verbal signals can be translated into valuable information conveying affective messages. Head movements and hand gestures are sometimes even more informative than facial expressions [7] and the extraction of features of the corresponding movements (e.g. spatial movements, speed, fluidity) carry necessary messages for non-verbal behavior recognition.

An abundance of research works within the fields of psychology and cognitive science related with the non verbal behavior and communication stress the importance of qualitative expressive characteristics and cues of body motion, posture, gestures and, in general, human action during an interaction session [8]. Nevertheless, it is hard to identify specific characteristics of non verbal behavior that could help us assess a user's emotional state. 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 in the IEEE Transactions on Affective Computing (TAC) as well as in the two books that have been recently published ([9], [10]) and deal with the entire spectrum of research related to Affective Computing. Investigating, though, Natural Interaction, especially in three dimensions, and performing comparative studies regarding gesture, head pose and full body expressivity formalization, remains a scarcely studied domain.

Some research work has been performed recently on the actor portrayals corpus [11]. The authors extract video-based nonverbal gesture features on human upper-body movements performed within the GEMEP (Geneva Multimodal Emotion Portrayals) corpus. They extend the EyesWeb XMI Expressive Gesture Processing Library in order to calculate static and dynamic expressivity features. Although they formulate three, common with current state of the art, expressivity features (energy, spatial extent and smoothness), their approach is based on monocular vision from frontal and side views. A similar approach was adopted in [12] in order to investigate emotional expression in music performance. Additionally, [13] focuses more on motion segmentation into motion primitives, based on energy monitoring. Joint representation of stylized motions, of a motion capture corpus, is processed in order to derive features and classify motions into four basic emotions. The interesting point of this work is the

978-0-7695-4887-6/12 $26.00 © 2012 IEEE
DOI 10.1109/SMAP.2012.11

incorporation of personal movement bias. Posture features, based on relative distances and orientations, are correlated with affective dimensions in [14]. Finally, such information has been fused with modalities used widely in Affective Computing such as facial expressions and speech prosody [15], [16], [17]. Nevertheless, very few, when compared to other modalities, contributions are available on the analysis of the dynamics of body movement to extract expressive and affective information. In this paper, we explore the applicability of utilizing expressivity features in the context of head, body and hand gestural movements. We present initial findings on the perspective of grounding a common framework for these cues, using expressivity features as the unique tool for mapping motion to emotional/behavioral states. The presented overview encourages a common formalization for fusing different modalities of expressiveness, under the immediacy of expressivity features that can summarize expressiveness in a compact and direct manner.

The structure of the rest of the paper is the following: Section II gives an overview of expressivity parameters and detection methodologies, while Section III presents a series of application-oriented examples, highlighting the relation of different expressivity features with affective dimensions. In Section IV we explore different options and contexts of utilizing expressivity for enhancing interaction, while Section V concludes the paper.

II. Gesture expressivity features

Features and cues of non verbal behavior are an integral part of the communication process since they provide information on the current emotional state and the personality of the interlocutor [18]. Common classification schemes include binary categories such as slow/fast, restricted/wide, weak/strong, etc. Our head and hand gesture expressivity modelling is close to these schemes in the sense that they provide a formulation and a quantitative measurement of the respective aspects of the gesture. Adopting a subset of the gesture synthesis expressivity modelling parameters (features) [19], we define five expressivity features: Overall Activation, Spatial Extent, Temporal, Fluidity and Power.

Overall Activation is considered as the quantity of movement during a dialogic discourse and is formally defined as the sum of instantaneous quantities of motion. Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). In order to provide a strict definition of this expressivity feature, Spatial Extent is considered as the maximum value of the instantaneous spatial extent during a gesture. The Temporal expressivity parameter denotes the speed of movement during a gesture and dissociates fast from slow gestures. The Power expressivity parameter refers to the movement during the stroke phase of the gesture. Detecting the stroke phase of the gesture is far from trivial and thus we opted to associate this parameter qualitatively with the acceleration of hands during a gesture. Fluidity differentiates smooth / elegant from sudden / abrupt gestures. This concept attempts to denote the continuity between movements and is usually suitable for modelling modifications in the acceleration of the upper limbs. Under this prism, we formally define as the gesture's Fluidity the variation of Power. According to the latter formalization, Fluidity expressivity parameter corresponds a quantity that is reversely proportional to the notion of fluidity. A computational formulation of the parameters described above can be found in [20] while real time computation of these parameters is illustrated in Fig. 1.

Figure 1. Real time computation of gesture expressivity parameters

III. Analysis

Current section discusses the application of the expressivity features, presented in Section II, into three modalities, namely, hand gestures, head motion and full body motion. The expressivity features computational formalization for the aforementioned modalities is adapted according to the interaction characteristics and expressiveness perception. Additionally, a comparative study on optimal, per feature formalization is performed for 3D full body motion. Finally, three corpora are incorporated ([21], [20] and [22]) each focusing on the modality investigated.

A. Using hands expressivity features as predictors of emotion

Regarding the hand detection and tracking step for extracting expressivity features from a gesture, we adopted a video-based, non obtrusive approach which focuses on low computational cost and robustness. The overall process for hand detection and tracking, described in detail in [20], includes creation of moving skin masks and tracking the centroid of these skin masks among the subsequent frames of the video depicting a hand gesture. Real time color models of the human skin are constructed by sampling the upper area of the box containing the head which corresponds to the forehead of the user, thus tackling illumination issues which often impede natural interaction processing. Object correspondence between two frames is performed by a heuristic algorithm and the fusion of color and motion information eliminates any background noise or artifacts, thus, reinforcing robustness. The overall process is depicted in Fig. 2.

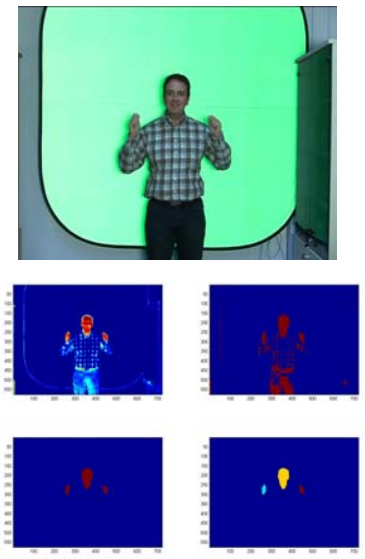

Figure 2. Image processing intermediate steps and final result for hand detection

Results on acted gestures in the dataset described in [20] are promising with regards to the construction of non-linear functions mapping expressivity parameters to the affective dimensions of Activation (arousal) and Evaluation (pleasure).

In order to evaluate the appropriateness of each of the expressivity geatures in estimating dimensions, we used Fisher's exact test [23]. To this aim, we quantized the values of Activation and Evaluation of typical videos of the dataset[1] to the closest integers (0 and 1), thus splitting the dataset in two groups for each dimension. A 3-bin histogram of low, medium and high values for each of the expressivity parameters was calculated for each of the two groups, one for low-high activation and one for low-high evaluation. The resulting distributions for the low and high values of each dimension separately, were compared against each other.

Fisher's exact test for histograms comparison was preferred over other methods (such as the chi-square method), because it is ideal for small scale data. Indeed in the current data set it is often the case that there are only a few instances with low or high values at the correspondent histogram bins (for example, the temporal parameter did not have a lot of instances in the third bin in the case of high activation judgments). Fisher's exact test is ideal in depicting such differentiations in cases of small samples.

The statistical test indicates the rejection of an expressivity parameter if its histogram values, for each dimension, are not significantly different ($p > 0.05$). In our case, we were led to rejecting the Overall Activation parameter, as a non-useful parameter at estimating the Activation dimension. This is qualitatively explained if one takes into account the fact that, by definition, Overall Activation is especially sensitive during the whole video process. Thus, while raters stress out the depicted gesture itself, the automatic parameter extraction takes into account the total number of the frames in a gesture, considering information not related to the gesture under consideration (apex and offset phases of the gesture).

Similarly, in the case of the Evaluation dimension, the Power parameter was discarded. The qualitative explanation for this is the fact that the same "amount" of Power may express either pleasure or displeasure in a gesture. Fig. 3 shows typical examples of features distributions for both dimensions.

[1]the values of Activation and Evaluation of each video sequence were calculated by annotations, in an online survey

B. Head Motion Expressivity during Gameplay

Experiments on the Mario dataset, described in [21], have shown that estimating expressivity parameters is a rich source of information regarding issues related to gameplay behavior and performance. Furthermore, analysis has shown that there do exist significant correlations between head expressivity and person-dependent characteristics (demographics, user profile, etc.). These findings are aligned with the need to deliver personalized and adaptive games, aiming at maximizing the notion of flow [24] during game play. More details regarding head detection and tracking can be found in [21].

Calculating head motion on the participants of the Mario dataset has shown that there exists correlation between head movement and the amount of time a person dedicates to game-playing on a weekly basis. Our results have shown that head movement is correlated (p=0.03) and decreases with the amount of hours a person spends on gameplay, while, similar, statistically significant results were found for the parameter of head movement spatial extent: Experienced players do not tend to make large movements while they play, in contrast to less experienced ones (p=0.03). One more factor that appears to affect Head Motion is age, where, the older a person is, the more chances they have to adopt an intense visual behavior (two groups of players were considered: those from 30 years old and older and players in their twenties). Although younger players would adopt higher levels of power in their motion than older ones (p=0.04), they exhibit lower levels of fluidity (as defined above) (see Fig. 4(a)), while older ones make larger movements (Spatial Extent) (p=0.044). Similar, gender seems to be a statistically important indicate of Overall Activation (women tend to be more expressive than men) and Spatial Extent (p=0.047), while measurements on ethnicity show that it also appears to play a role as well; comparisons among Greek and Danish players showed that Greeks have higher levels of overall activation than Danish people (p=0.0092) (see Fig. 4(b)),

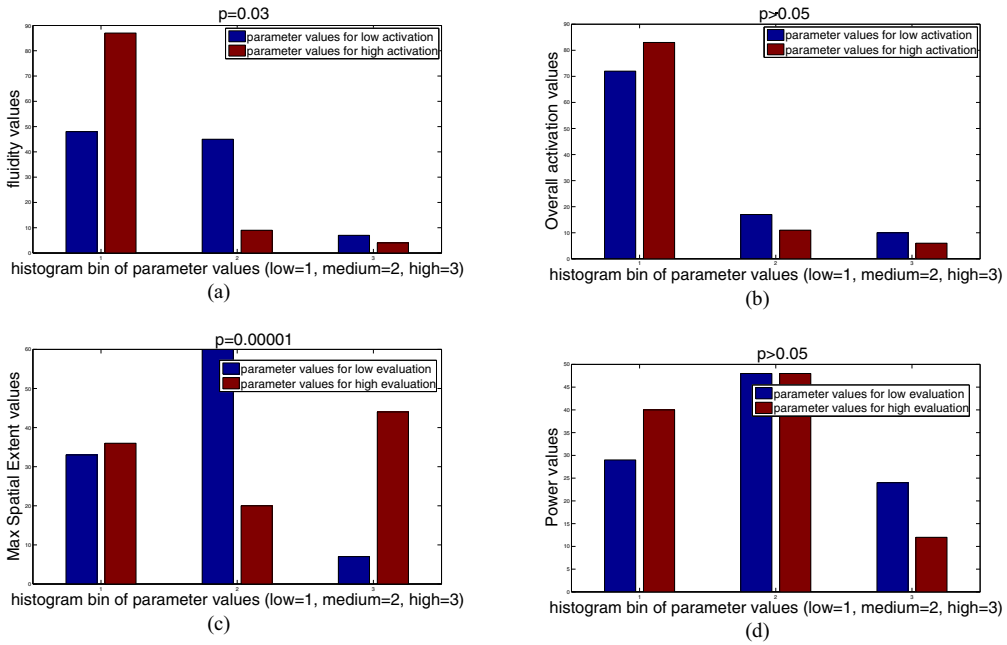

Figure 3. Examples of expressivity parameter distributions corresponding to high or low values of activation (a-b) and evaluation (c-d)

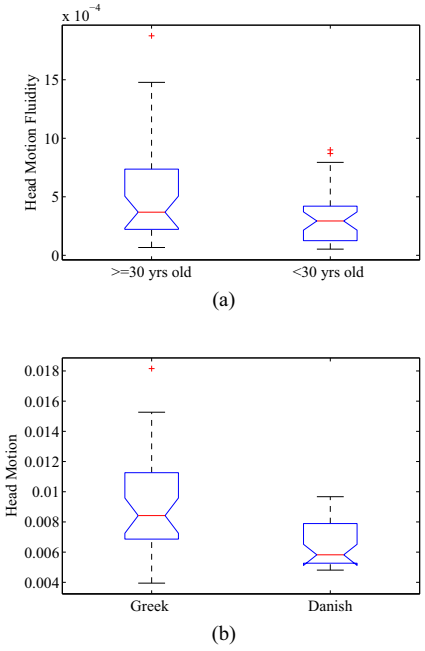

Figure 4. a) Fluidity and Age , b) Overall activation comparison between players of two different nations

while similar is the case for moments when a critical action is about to be taken (p=0.03).

C. *Full body 3D expressivity*

Attempting to extend the gesture expressivity to full body in a 3D expressivity computational formalization, a dataset was constructed by recording four users while performing variants of movements using Microsoft's Kinect [22]. Since this study's aim was to investigate the optimal approach to computationally formulate body expressivity, the dataset was constructed based on acted, extreme expressions and not natural or naturalistic expressions. During recordings, the subjects were asked to perform two body movements per expressivity feature corresponding to their interpretation of maximum and minimum value.

Figure 5. Original, depth and skeleton images of the dataset

Silhouette binary image, depth image map and skeleton joint rotations were calculated as shown in Fig. 5. This input is used to formulate each full body expressivity feature using one of the following approaches:

1) silhouette
2) limbs
3) joints

Although silhouette is usually used in full body expressivity analysis, limb-based expressivity formalization presents

Table I
PEARSON CORRELATION FOR ACTIVE AND PASSIVE MOVEMENTS

Silhouette	0.0192
Joint	-0.0435
Limb	0.22567

interest since it has been used before in half-body, desktop interaction context. One could argue that limb-based analysis is a subcase of the silhouette-based one but, on the other hand, extracting features or points/regions of interest using computer vision and image processing techniques is an entirely 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 a wider range of applications and interaction contexts. Finally, joint expressivity formalization is quite innovative, since, robustly extracting relative features, is an extremely challenging task and researchers opted to simpler and more robust approaches.

For each approach, only the Overall Activation expressivity feature is discussed since it is indicative of the formalization approach. [22] describes in detail all the expressivity features for all the approaches. Silhouette based formalization is based on the notion of fading silhouette motion volumes and is defined as:

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

Limb-based Overall Activation is defined similarly to the 2D gesture counterpart but does not include only hands, and the limbs are positioned in 3D space. Finally, joint-based Overall Activation is defined as a weighted sum of joints rotations derivative. Overall Activation mean values for active and passive movements for the three approaches are depicted in Fig. 6. Table I illustrates the Pearson correlation for different, with respect to the activation level, body movements and different expressivity modelling approaches.

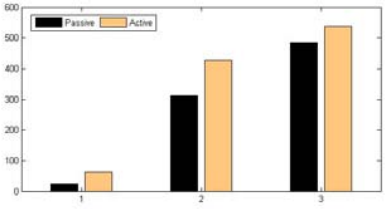

Figure 6. Mean values for 1: joint, 2: limb and 3: silhouette-based Overall Activation formalization for active and passive movements

IV. DISCUSSION

Motion expressivity parameters can play an important role in various contexts. For instance, Human-Machine interactions are currently using simplistic and heuristic means of having the user declare his presence, either by pressing a button or by applying face detectors for verifying the existence of a person in front of the machine. However, social and behavioral management of interaction is an important feature, missing from interaction systems. Being able to map user reactions to an avatar's behavior will help the interaction adopt a more naturalistic aspect. Moreover, exploring affective dimensions in game-play has already been shown to play an important role at estimating user preferences and states [25]. Such cues can be fused with game content features and can lead to personalized models and strategies for maximizing flow [24].

Regarding future directions of the research work presented here, these include: a) Investigating the applicability and effectiveness of the analysis discussed earlier on real life gaming or other interaction contexts and b) designing an integrated architecture for non verbal interaction analysis and adaptation mechanism. Concerning the former, we aim to construct a dataset consisting of real life interactions and test the affective analysis on this corpus. The recording of such a corpus is feasible due to the completely unobtrusive recording methods used. On the other hand, challenges related to real life interaction, such as uncontrolled behavior or subjective annotation, are issues that require tackling. Modelling and incorporating interaction or interaction context in the analysis is also going to be an extra flow of information, valuable for the analysis. Finally, appropriate ways (and, hopefully, an integrated architecture) to incorporate extracted expressivity features in interaction scenarios or agent behavior adaptation will constitute a challenging future research direction: Correlating and exploring dependencies among expressivity parameters (affective behavior), interaction performance (cognitive state) and user characteristics (personalization) is expected to utilize machine learning architectures that can constitute the link between observed behavior and adaptation mechanisms, capable to maximize user engagement and performance.

V. CONCLUSIONS

Being able to transfer affective information within interaction scenarios is of high importance for tailoring the content of interaction to the user's preferences and/or needs. The need for unobtrusive and robust mechanisms, capturing non-verbal behavior can support such applications. Modelling expressivity in motion has been shown to be effective at modelling user's profile characteristics and estimating human's emotional and cognitive state. Future directions of this research will boost adaptation mechanisms in various fields, so that affect-based procedural content generation experiments ground the validity of our results.

ACKNOWLEDGMENTS

This work was funded by the European ICT-Project 'Siren', (under contract FP7-ICT-258453)

REFERENCES

[1] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proc. International Conference on Multimodal Interfaces (ICMI-MLMI 2009)*. New York, NY, USA: ACM, 2009, pp. 119–126.

[2] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration." *International Journal of Man-Machine Studies*, pp. 724–736, 2007.

[3] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Estimation of behavioral user state based on eye gaze and head pose - application in an e-learning environment," *Multimedia Tools and Applications, Springer*, vol. 41, no. 3, pp. 469 – 493, 2009.

[4] T. Saari, M. Turpeinen, K. Kuikkaniemi, I. Kosunen, and N. Ravaja, "Emotionally adapted games - an example of a first person shooter," in *Proc. 13th International Conference on Human-Computer Interaction (HCI 2009). Part IV: Interacting in Various Application Domains*, 2009, pp. 406–415.

[5] J. Tracy and R. Robins, "The prototypical pride expression: Development of a nonverbal behavior coding system." *Emotion*, vol. 7, no. 4, p. 789, 2007.

[6] B. De Gelder, "Towards the neurobiology of emotional body language," *Nature Reviews Neuroscience*, vol. 7, no. 3, pp. 242–249, 2006.

[7] S. Langton and V. Bruce, "You must see the point: Automatic processing of cues to the direction of social attention," *Journal of Experimental Psychology Human Perception and Performance*, vol. 26, no. 2, pp. 747–757, 2000.

[8] S. Trenholm and A. Jensen, *Interpersonal communication*. Oxford University Press, USA, 2007.

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

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

[11] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Towards a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106–118, 2011.

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

[13] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Proc. of Affective Computing and Intelligent Interaction (ACII 2007)*. Springer, 2007, pp. 59–70.

[14] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proc. of Affective Computing and Intelligent Interaction (ACII 2007)*. Springer, 2007, pp. 48–58.

[15] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 1, pp. 64–84, 2009.

[16] 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*, pp. 375–388, 2007.

[17] M. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. 9th International Conference on Multimodal Interfaces*. ACM New York, NY, USA, 2007, pp. 38–45.

[18] A. Mehrabian, *Nonverbal communication*. Aldine, 2007.

[19] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents," in *Proc. 4th International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM, 2005, pp. 1095–1096.

[20] 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*, vol. 41, no. 3, pp. 367–388, 2007.

[21] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "A game-based corpus for analysing the interplay between game context and player experience," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII2011), EmoGames workshop*, 2011.

[22] G. Caridakis and K. Karpouzis, "Full body expressivity analysis in 3d natural interaction: a comparative study," in *Proc. of International Conference on Multimodal Interaction (ICMI 2011), Affective Interaction in Natural Environments workshop*, Nov. 2011.

[23] R. Fisher, *Statistical methods for research workers*. Oliver and Boyd, 1954.

[24] M. Csikszentmihalyi, *Beyond boredom and anxiety: the experience of play in work and games*. San Francisco: Jossey-Bass, 1975.

[25] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.