Intelligent Narrative Technologies: Papers from the 2013 AIIDE Workshop (WS-13-21)

CONTACT: A Multimodal Corpus for Studying Expressive Styles and Informing the Design of Individualized Virtual Narrators

Cheikh Mbengue, David Antonio Gómez Jáuregui, Jean-Claude Martin
LIMSI-CNRS, B.P. 133, 91403 Orsay, France
cheikhatm@gmail.com, gomez-jau@limsi.fr, martin@limsi.fr

Abstract

Virtual agents used in storytelling applications should display consistent and natural multimodal expressions of emotions. In this paper, we describe the method that we defined to endow virtual narrators with individual gesture profiles. We explain how we collected a corpus of gestural behaviors displayed by different actors telling the same story. Videos were annotated both manually and automatically. Preliminary analyses are presented.

Introduction

Research in the field of virtual characters aims at using multiple interaction modalities (e.g. speech, gestures, and postures) (Cassell et al. 2000), (Knapp and Hall 2005), (Scherer and Ellgring 2007). These virtual characters play an important role in communication between users and systems. One goal of researchers is to provide these virtual characters with social intelligence and natural multimodal behaviors to allow these characters to interact naturally and easily with the user (Kihlstrom and Cantor 2000).

Designing virtual characters requires detailed knowledge on how to convey emotions (Scherer and Ellgring 2007), (Mehrabian 1996), (Russel and Mehrabian 1976). Databases of facial expressions of actors during dyadic interactions have been collected (Busso et al. 2008) but to our knowledge there is no database of individual gestural performance of actors during a storytelling application. General knowledge is not enough and corpora enable to grasp individual gesture profile (Kipp 2004) which might be needed for an entertaining storytelling system.

The goal of our work is to inform the design of virtual narrators (e.g. how their gestures are selected and realized). We propose a multimodal corpus-type approach based on manual annotations and automatic video analysis. Manual annotations are made for gestures and emotions while automatic video analysis is made to estimate quantity of motion. We explain how our method enables to provide knowledge about how several narrators' gesture. We also illustrate how our method enables to grab individual gesturing and expressive styles of each different narrator.

Method

The goals of this study are threefold: 1) analyze the most frequent gestures and emotions in a storyteller task for each individual participant, 2) specify (or identify) the relationship between gestures and emotional states, and 3) study the variations in terms of quantity of motion (since previous studies have observed this feature to be related to the expression of emotions).

We collected a video corpus that we called "ContAct": Conte Acté ("Acted Story" in French). This video corpus contains multimodal expressions of emotions displayed by six actors telling the same story. The story, *"Three small pieces of night"* was selected due to its wide range of emotional content (positive and negative emotions). Actors were asked to tell the story using their whole body. Two cameras were used to provide different viewing angles (front and profile) over the gestures since the three dimensions are known to be of importance for gestural and postural expressions of emotions (figure 1).

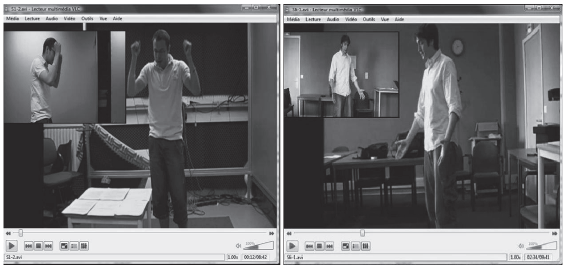
S1-2.avi - Lecteur multimédia VLC
Média Lecture Audio Vidéo Outils Vue Aide
S1-2.avi
1.00x
00:12/08:42
S6-1.avi - Lecteur multimédia VLC
Média Lecture Audio Vidéo Outils Vue Aide
S6-1.avi
1.00x
02:34/09:41

Figure 1: Frames of the ContAct corpus

In order to annotate gestures, we used the following steps (Kipp 2004):

1. **Gesture segmentation.** Each video was segmented in different time intervals which contain a specific gesture (gesture phase).
2. **Handedness.** The use of hands was annotated for each gesture (left hand, right hand or both hands).
3. **Gesture category.** Each gesture was identified as one of the five gestural categories (Kipp 2004).

The emotional categories were annotated by segmenting each video with respect to the different sentences spoken by the actors. For each sentence, the emotion was annotated by giving a value from High, Medium and Low to each of the three main dimensions of emotions (pleasure, activation and dominance). By combining these three variables, we obtain a category of emotion selected from the following list (Mehrabian 1996), (Kipp and Martin 2009), for example: Exuberant, admired, bold, excited (+P +A +D).

The validation of annotations was made with the calculation of Kappa (Carletta 1996). Two observers validated the manual annotation of use of hands and gestures. The inter-rater reliability for the use of hands was found to be Kappa = 0.83 and for the gestures was Kappa = 0.59. For the moment, the emotional categories have been validated by one observer, as future perspective the annotations will be made by several observers and a measure of inter-coder agreement.

The quantity of motion for each frame was automatically extracted using computer vision algorithms. In order to obtain the quantity of motion, we compute motion history images (MHI) (Bobick and Davis 2001). This method has been proven to be very robust in detecting motion and is widely employed by various research groups for action recognition, motion analysis and other applications (Ahad et al. 2010). These MHI provide, for each frame, an image template where pixel intensity is a function of the recency of motion in a sequence. Thus, for each frame of a video sequence, the quantity of motion is the total number of pixels in the image where some motion has been detected. The quantity of motion was normalized to be ranged from 0 to 1. As the value gets closer to 1, the quantity of motion will be higher in our automatic annotation.

Results

The use of both hands at the same time was the most frequent over all video sequences with a percentage of 74.30%. The right and left hand obtained percentages of 11.94% and 13.61% respectively.

The results of our study show that for the entire video corpus, the most frequent gestural category is *Emblem* (49.16%), followed by *Metaphoric* (17.08%), *Deictic* (14.16%), *Beats* (10.14%) and *Iconic* (7.36%)

The frequency of emotional categories was analyzed for the entire video corpus. According to results, the most frequent emotional category presented for the task is +P+A-D (55.12%), followed by -P-A+D (18.65%), -P+A+D (12.10%), -P+A-D (10.92%), -P-A-D (1.34%), +P+A+D (1.18%), +P-A+D (0.33%) and +P-A-D (0%).

We analyzed the frequency of each gesture regarding the most frequent emotion (+P +A-D). *Emblem* gestures represent almost half (47.9%) of the total of gestures for this emotion (+P +A -D), followed by *Beat* (20.5%), *Metaphoric* (16.4%), *Deictic* (8.89%) and *Iconic* (7.3%).

The average of the quantity of motion was calculated for each video sequence (storyteller) with respect to handedness, gesture category, and emotion. We computed a two-way ANOVA analysis design for each group of two factors (use of Hands x Storytellers, Gestural Categories x Storytellers, Emotion x Storytellers). The response measure was the average of the quantity of motion.

The quantity of motion measure was entered into a two-way factor (*Use of hands* x *Storyteller*) ANOVA analysis. For the *use of hands* factor, the independent variables were: *Two Hands* (2H), *Right Hand* (RH) and *Left Hand* (LH). For the *Storyteller* factor, the independent variables were: *Storyteller 1* (S1), *Storyteller 2* (S2), *Storyteller 3* (S3), *Storyteller 4* (S4), *Storyteller 5* (S5), *and Storyteller 6* (S6). The dependent variable was the *quantity of motion*.

No significant main effects were found when examining the *use of hands* factor ($F(2,18)$ = 0.32; p = 0.728). However, results showed significant main effects for *Storyteller* factor ($F(5,18)$ = 15.21; $p < 0.0001$). Post-hoc comparisons revealed a significant main effect between S1 and S3 (T = -7.76; $p < 0.001$), S1 and S4 (T = -5.29; $p < 0.01$), S1 and S5 (T = -6.77; $p < 0.01$), S3 and S6 (T = 4.46; $p < 0.05$) and finally between S5 and S6 (T = 3.47; $p < 0.05$). Higher quantity of motion means were found for *Storyteller 1* (S1) and *Storyteller 6* (S6) while lower quantity of motion means for *Storyteller 3* (S3), *Storyteller 4* (S4) and *Storyteller 5* (S5). No interaction was found between both factors.

The quantity of motion measure was entered into a two way factor (*Gestural categories* x *Storyteller*) ANOVA analysis. For the Gestural categories factor, the independent variables were: *Emblem*, *Metaphoric*, *Deictic*, *Iconic* and *Beat*. For the Storyteller factor, we have the independent variables: *S1*, *S2*, *S3*, *S4*, *S5* and *S6*.

No significant main effects of quantity of motion were found for the *Gestural categories* factor ($F(4,30)$ = 1.23; p = 0.326). Regarding the *Storyteller* factor, a main effect was found ($F(5,30)$ = 8.15; $p < 0.0001$). Post-hoc comparisons showed a significant main effect between S1 and S3 (T = -5.51; $p < 0.001$), S1 and S4 (T = -4.83; $p < 0.01$), and finally between S1 and S5 (T = -4.50; $p < 0.01$). S1 and S6 showed higher quantity of motion while S3, S4 and S5 showed lower quantity of motion (Similarly to the

Use of hands x *Storyteller* analysis). No interaction was found between both factors.

The quantity of motion measure was entered into a two way factor (*Emotions* x *Storyteller*) ANOVA analysis. For the *Emotions* factor, the independent variables were: *+P+A+D, -P-A-D, +P+A-D, -P-A+D, +P-A+D, -P+A-D, +P-A-D, -P+A+D*. For the Storyteller factor, we have the independent variables: *S1*, *S2*, *S3*, *S4*, *S5* and *S6*.

Results showed no significant main effect for *Storyteller* factor (*F(5,48)* = 2.44; *p* = 0.05). However, the *Emotions* factor showed significant main effects found (*F(7,48)* = 3.97; *p* < 0.01). Post-hoc comparisons showed a significant main effect between emotions *-P-A-D* and *+P+A+D* (*T* = 3.57, *p* < 0.05). Another significant main effect was showed between emotions *+P-A-D* and *+P+A+D* (*T* = 4.38, *p* < 0.01). The highest quantity of motion was observed during the expression of the emotion *+P+A+D* while the lowest quantities of motion were showed during emotions *+P-A-D* and *-P-A-D*. No interaction was found between factors.

Discussion

We identified the most frequently used gestures and emotions. We observed significant differences in motion activity between each gesture, emotion and storyteller.

With respect to the use of hands, the results showed that the use of both hands at the same time is the most frequent. This may be explained by the need for actor to provide a high expressivity and clear explanations about the story. *Emblem* was the most frequently used gesture. The use of the gestures in this task seems to reinforce the verbal discourse by adding new semantic meaning. The most common expressed emotion by the storytellers was +P+A-D (Dependent, surprised, happy, friendly). As expected, this emotion is consistent with the story. Results showed that the frequency of gestures is determined by the emotion expressed. This suggests that it might be possible to identify the emotion expressed by an actor by identifying his most frequent gestures. Symmetrically, it might be possible to predict the most probable gesture that can be used to express a given emotion.

The *Use of hands* and *Gestural categories* have no influence on the motion activity of the storyteller. However, significant differences in the quantity of motion between storytellers suggest that motion activity could be related with their personality or their narrative style. Finally, we found that the motion activity is associated with the emotion expressed by the storyteller.

The results suggest that two handed gestures should be selected if we want to design a virtual storyteller. While some gestures can be performed with one or two hands, those executed with two hands may be more expressive. The *Emblem* gesture must also be correctly simulated and adapted to the semantic of the speech. The frequency of specific gestures and the motion activity must be directly related with the emotion simulated. Finally, the intensity of the motion showed by the virtual storyteller might be related to its personality or narrative style.

Conclusions

The results found in this study can be used to inform the design of individual virtual storytellers. For future work, other modalities (e.g. posture, facial expressions) will be integrated to our approach and their relations will be studied. The method will be extended to consider personality and narrative style.

References

Ahad, M.A.R.; Tan, J.K.; Kim, H.; and Ishikawa, S. 2010. Motion history image: its variants and applications. *Machine Vision and Applications (MVA)* 23(2):255-281.

Bobick, A.F.; and Davis, J. 2001. The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(3): 257-267.

Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; and Narayanan, S.S. 2008 IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42(4): 335-359.

Carletta, J. C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2): 249-254.

Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E. 2000. Embodied Conversational Agents. MA: MIT Press.

Kihlstrom, J.F.; and Cantor, N. 2000. Social Intelligence. *Handbook of Intelligence*. Cambridge University Press. 359-379.

Kipp, M. 2004. Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Ph.D. diss., Saarland University, Boca Raton, Florida.

Kipp, M.; and Martin, J.-C. 2009. Gesture and Emotion: Can basic gestural form features discriminate emotions?, In *Proceedings of the International Conference on Affective Computing and Intelligent Interactions (ACII 2009)*, 1-8. Amsterdam, The Netherlands: IEEE Press.

Knapp, M. L., and Hall, J. A. eds. 2005. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing Company.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14(4):261–292.

Russel, J.A.; and Mehrabian, A. 1976. Environmental variables in consumer research. *Journal of Consumer Research*. 3(1):62-63.

Scherer, K. R.; and Ellgring, H. 2007. Multimodal expression of emotion: affect programs or componential appraisal patterns? *Emotion* 7(1): 158–171.