Technique for automatic emotion recognition by body gesture analysis.

Anonymous FG2008 submission

Abstract

This paper illustrates our recent work on the analysis of expressive gesture related to the motion of the upper body (the head and the hands) in the context of emotional portrayals performed by professional actors. An experiment is presented which is the result of a multidisciplinary joint work. The experiment aims at (i) developing models and algorithms for analysis of such expressive content (ii) individuating which motion cues are involved in conveying the actor's expressive intentions to portray four emotions (anger, joy, relief, sadness) via a scenario approach. The paper discusses the experiment in detail with reference to related conceptual issues, developed techniques, and the obtained results.

1. Introduction

The human ability to communicate through body movement and gestures has been studied with increasing interest [1]. Mainly an interest in the psychological field, attention to the behavioral expressions of emotions has recently been sustained by research in computer science. It is now acknowledged that a human computer interface (like a robot or an avatar) would greatly improve its interactivity by recognizing and interpreting users' emotional states and by communicating expressive-emotional information to them [2]. As a consequence, much of the research in this domain is wrapped up in designing automated video analysis algorithms aiming to extract, describe and classify information related to the emotional state of individuals.

This paper presents a study on a database of portrayed emotional expressions which contain more than 7000 audio-video emotion portrayals that represent 18 emotions, portrayed by 10 actors, the GEMEP Archive [3]. Acted emotion portrayals allow the recording of a defined group of foreseen senders with good picture quality, as well as maximal expressive variability. The uniform structure of acted portrayals makes them well suited for developing automatic recognition systems. Our analysis is based on a selection of 40 portrayals that have been systematically rated by experts and non-experts.

The research aims at investigating the motion cues that characterize the expressions of emotions by means of the video analysis of movement and gesture. We provide a novel approach to quantitatively observe and evaluate the users in an ecological environment, particularly the kinematical features of the upper body motions (the head and the hands) of professional actors portraying four emotions (anger, joy, relief, sadness). Motor cues are identified that are related to the qualitative aspects of the gesture (i.e. the way the movement is performed, e.g. impulsive or fluid).

The first two results concerning the velocity of the head and hands displacement and their space occupation are explained in detail. It is suggested that the dynamic of 3D points corresponding to the head and hands main body extremities can be sufficient to distinguish the four emotions.

This work is a preliminary step that will be extended to a more extensive and multimodal coreset of portrayals, including audio voice emotional cues. Results related to body expression will be put into relation with acoustic and facial expression cues.

2. Background

Recent psychological studies gave significant insights of human expressive movement by the use of impoverished displays (e.g. point light display) to understand the respective contribution of dynamic and form information to the recognition of emotional processes [4]. Other studies showed that emotions can be related to specific movement qualities [1,3,5,6]. In the last years, coding systems have been developed to reach a more complete and standardized description of body movements relevant to emotions. For example, Ekman et al. [7] realized the facial action coding system (FACS) in order to describe changes in the appearance of the face. Wallbott [1] focused his attention on specific body parts, especially the hands and head which are known to be responsible of the majority of movement frequency located at these body parts. He pointed out that different types of hand and arm postures and movements are specific for some emotions (e.g. arms stretched sideways for terror). He also found that external (lateral or frontal) hand or arm movements

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

are most frequent during 'active' emotions like hot anger, cold anger, elated joy and terror.

In human-computer interaction, there is increasing attention given to automated analysis techniques aiming to extract and describe information related to the emotional state of individuals. In particular, some attempts were made towards the development of systems able to analyze expressive body movements and automated emotion recognition.

Camurri, Lagerlöf and Volpe [8] classified expressive gestures in human full-body movement. In particular, they identified motion cues like overall duration of time, contraction index, quantity of motion and motion fluency. On the basis of these motion cues, they defined an automated classifier that was able to distinguish between four emotions (anger, fear, grief, and joy). Kapur et al. [9] used full-body skeletal movement data obtained with a technology based on the VICON motion capturing system able to distinguish between four basic emotional states. They showed that very simple statistical measures of motions' dynamics are sufficient for such classification. In a study on non-propositional movement (i.e., movement that does not convey an explicit meaning such as a raised hand to indicate stop), Castellano, Villalba and Camurri [10] found that velocity of the hand and the quantity of motion of the upper body played a major role in discriminating between different emotions (anger, joy, serene, sadness). The role of kinematical cues has been further established by the recent study of Bernhardt and Robinson [11]. Further developing the motion-captured knocking motion from Pollick [4], they developed a computational approach to extract affect-related dynamic features. The measures of the velocity, acceleration and jerk of each joint composing the skeletal structure of the arm proved successful in the automatic recognition of the expressed affect (neutral, anger, happy and sad).

3. Method

A layered approach [8] has been adopted to model human movement and gesture moving from low-level physical measures (e.g., position, speed, acceleration of body parts) toward descriptors of overall motion features (e.g., motion fluency, directness, impulsiveness). Such a high-level description of the gesture dynamics aims at revealing the qualitative aspects of movement (how it is performed, e.g., whether it is impulsive or smooth, hesitant or fluent) more rather than on what is achieved (task description). Our work builds on these dynamic approaches, gaining its main affective information from variables such as body extremities velocity. Further extraction and analysis were supported by statistical and computer engineering techniques. The EyesWeb XMI software open platform [12] has been chosen to perform video processing (segmentation and tracking of body-parts). Extraction of expressive features from human movement has been partly carried out using the EyesWeb Expressive Gesture Processing Library, the collection of software modules implemented in the same platform.

3.1. Material

For this study we used a database of portrayed emotional expressions (Geneva Multimodal Emotion Portrayals, GEMEP), recorded at the University of Geneva. The GEMEP corpus consists of more than 7000 audio-video emotion portrayals representing 18 emotions, portrayed by 10 actors. A systematized selection was made on the basis of expert and non-expert ratings resulting in a set of 150 portrayals of excellent visual and technical quality with the highest recognition of the actors' emotional intention. Our analysis is based on the selection of this set comprised of 40 portrayals. These portrayals have been selected because they present four emotions (anger, joy, relief and sadness), each belonging to one quadrant of two major affective dimensions, namely arousal and valence [13,14].

	Positive valence	*Negative valence*
High-arousal	Joy	Anger
Low-arousal	Relief	Sadness

Table 1. the emotions in the space valence-arousal

Actors were recorded in a flexible environment with regard to lighting and actor appearance (e.g. clothing variations). During the recording procedure, the actors were requested to express an emotion in interaction with a professional theatre director on the basis of three short scenarios with a definition of every intended emotion. A full description of the database and the rationale of the recording procedure can be found in Bänziger and Scherer [3].

Two digital cameras with constant shutter, manual gain and focus, at 25 fps with a 720x576 pixel resolution were used to record the body movements of the actors from both the frontal and profile view. With regard to body movement restrictions, the actors were only instructed not to move away from the focus of the camera.

3.2. Automated extraction of expressive cues

In order to perform a movement analysis of the actors, motion cues were automatically extracted from the video recordings. Our analysis was based on the position and the dynamics of three main body extremities (hands and head position). EyesWeb modules based on skin color tracking algorithm were created to extract the three blobs (see figure 1). A three-dimensional position of the blobs' barycentre (centroid) was achieved by using the two synchronized separate movies (frontal and lateral) for each gesture in the database. The velocity of the head and the

hand's movements were also calculated on the basis of the coordinates (x,y,z) of the blob's centroids.

The non uniform type and color of clothes worn by the actors made the blob extraction process delicate but successful. Different calibrations were required for each subject both for the position of the body extremities (e.g blob size increases when an actor wore a short-sleeved shirt) so we introduced in some cases truncation of the arm to reduce the blobs to hands only.

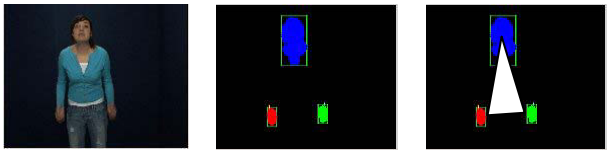

Figure 1: from left to right: the original image, the three extracted blobs based on color skin tracking algorithms and the bounding triangle used to extract contraction expansion and symmetry cues.

After the raw data was exported from the EyesWeb platform, we ran a custom built MATLAB program for importing, filtering the EyesWeb data and extracting features. A source of noise is related to the sensitivity of the tracking algorithm to luminance changes. To "smooth out" signals, we applied digital smoothing 3rd order polynomial filters (Savitzky-Golay) where the window's size was set to 11 (Savitzky-Golay smoothing filters performed better than standard averaging FIR filters, which tend to filter out a significant portion of the signal's high frequency content along with the noise). We had a particular interest in preserving the high frequency components of the signal because they corresponded to the short and frequent moves of the blobs (e.g. fast movement of the hands). We also decided to restrict our analysis to the trajectory of points and their velocity.

Following the layered approach proposed by Camurri (Camurri et. al 2003) our analysis begins with this low-level physical level (the 3-D coordinates of head and hands and their related velocity) to investigate which are the significant static and dynamic cues. The first motor cue we examined is the total quantity of displacement in all of three extremities. This first motor cue can be considered as the overall energy spent by the actor for each of his performance. This cue is related to the notion of movement activity developed by Wallbott [1]. Wallbott noticed that the activity factor accounts for some amount of variance of the differences between the emotions, in particular between active and passive emotions (e.g. anger vs. sad). This first cue also refers to the quantity of motion algorithm developed by Camurri et al. which is an approximation of the amount of detected movement based on the analysis of the silhouette variations [5.] Camurri et al. noticed that the quantity of motion is a relevant cue in recognizing emotion from the full-body movement of dancers. To calculate this first cue, we summed the modules of velocity of each extremity. The module of velocity was computed taking into account its horizontal, vertical and depth components.

We then considered a "bounding" triangle that relates the three blobs' centroids (see figure 1). By measuring its variation over time, we were able to approximate the space occupied by the head and the hands from the frontal view. The use of space in terms of judged expansiveness or spatial extension of movements have also been regarded by Wallbott as another relevant indicator for distinguishing between active and passive emotions, and in particular expansion/contraction and symmetry/ asymmetry with respect to the vertical axis.

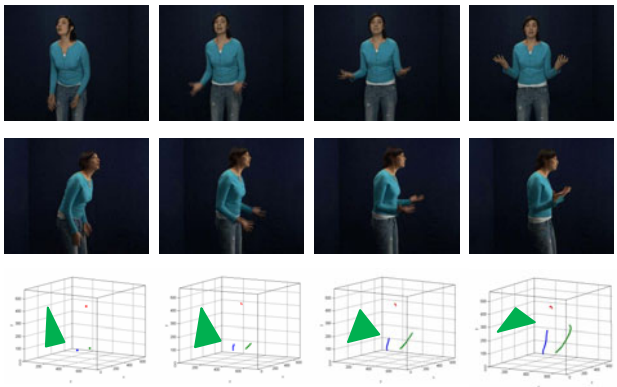

Figure 2: from left to right, four key-frames of a portrayal representing joy; from top to bottom, the frontal view, the profile view, the 3-D representation of the displacement of the head and the two hands in the space with the "bounding" triangle relating the three blobs.

A further processing step converted these time series into a fixed set of features conveying information about the dynamics of the cue [9]. We have explored in detail the following features (see figure 3):

- the attack and release parts of the cue, i.e. the slope of the line joining the first value and the first relative extremum and the slope of the line joining the last value and the last relative extremum.
- the number of local maxima of the cue, the ratio between the maximum and duration of the largest peak to approximate the overall impulsiveness of the movement (an impulsive movement is characterised by short peak duration with high absolute maximum, while sustained movement is characterised by longer peak duration with low absolute maximum).
- the local maxima preceding the absolute one to assess how the magnitude of the motion evolves over time
- the 2nd and 4th order statistical moments (standard deviation and kurtosis) have also been calculated for each motion cue to measure the distribution profile of values.

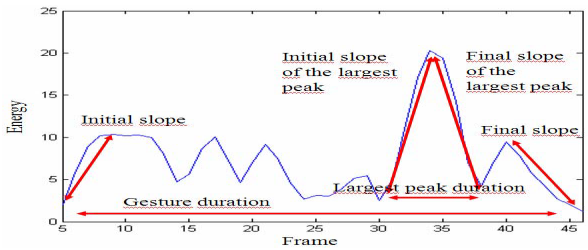

Figure 3: Selected dynamic features of the cue

Automatic extraction of the selected features was made using new software modules developed in EyesWeb. This process was used for each motion cue, so that each gesture is characterized by a set of 42 (2x21) features.

4. Preliminary results and discussion

In order to examine the effects of the different emotional conditions of the actor's performance, each of the 42 dependent variables described previously (2 motion cues x 21 features) was submitted to a repeated-measures analysis of variance (rm-ANOVA), using emotion (4 dimensions) as within subjects factor. The Greenhouse-Geisser correction was used when necessary to mitigate violations of the sphericity assumption in repeated measure designs [15]. To control the inflation of type I error probability due to multiple comparisons, the Bonferroni correction was applied to P-values (the levels of statistical significance).

The rm-ANOVA identified a significant effect of the different emotion conditions on a subset of the measured features. A larger effect of significant effects emerged for energy rather than for perimeter features. For these features, Bonferroni-corrected post-hoc analyses were performed, in order to assess the specific difference among the emotion conditions.

Motion cue "energy"		
Variable	Greenhouse-Geisser corrected F (dfs)	Post hoc test**
Gesture final release	8.426(1.585; 14.269)**	JOY,ANG>REL=SAD
Largest peak attack	8.754(1.867;16.801)**	ANG,JOY>REL=SAD
Largest peak release	8.792(1.692;15.226)**	ANG=JOY>REL=SAD
Maximum value	13.978 (3;27)*	ANG=JOY>REL=SAD
Mean value/ max value	11.855 (3;27)**	REL,ANG=JOY>SAD
Largest peak duration / gesture duration	7.931(1.548; 13.935)**	REL ,ANG=JOY>SAD
Max value / peak duration	12.783 (3;27)**	ANG,JOY>REL=SAD
Mean value of energy	18.943 (1.320;11.820)**	ANG=JOY>REL=SAD
Std value of energy	14.659(3;27)**	ANG=JOY>REL=SAD

Note: ANG = Anger; JOY = Joy; REL = Relief; SAD = Sadness
* Post Hoc Bonferroni test
** $p < 0.0083$ (Bonferroni-corrected alpha value)

Table 2: display of the most significant effects of emotion conditions on features related to the motion cue **energy**

Motion cue "perimeter"		
Variable	Greenhouse-Geisser corrected F (dfs)	Post hoc test**
Gesture final release	8.268 (1.845;16.609)**	JOY,ANG>REL=SAD
Max value	9.351(1.633;14.699) **	ANG=JOY>REL=SAD
Peak number	9.311(1.335;12.016) **	ANG=JOY>REL=SAD
Largest peak duration / gesture duration	5.579(3;27) **	REL,ANG=JOY>SAD

Note: ANG = Anger; JOY = Joy; REL = Relief; SAD = Sadness
* Post Hoc Bonferroni test
** $p < 0.0083$ (Bonferroni-corrected alpha value)

Table 3: display of the most significant effects of emotion conditions on features related to the motion cue **perimeter**

A significant effect of emotion conditions resulted from 9 features related to the energy factors, and 4 related to the perimeter variations (see Table 2 and 3). These results indicate that energy features are more sensitive to emotion condition compared to perimeter features. These results are in line with previous studies [1,7,8] where the quantity of motion value distinguishes distinguish between emotions. Our results also show that the short movements (an average of 2 seconds duration) are large enough to reveal differences among conditions in energy and its related features.

A significant effect concerns the attack and release parts of the gesture: the features *Final gesture release* (for energy and perimeter), *largest peak attack* and *largest peak release* (for energy) varied significantly according to the emotional condition. The critical role of attack and release parts to discriminate emotions was already revealed for the acoustic channel by Juslin et al. in music performance (speaking tone attacks and decays) [16].

Another significant energy related feature was *maximum value / peak duration*, which is a measure of the impulsiveness in movement. This feature, jointly with *Largest peak duration / Gesture duration* for energy, highlighted that the shape of the largest peak is sensitive to the emotional condition. The largest peak changed its shape across the conditions especially in terms of steepness (i.e. impulsiveness) and in its relative importance (i.e. proportion of duration with respect to the whole gesture).

The data analysis shows that energy is a meaningful motion cue for assessing acted emotions as the emotion conditions and that they have a significant effect on nine energy features. Results concerning the perimeter are less convincing since emotion conditions have a significant effect on only four perimeter features. However, energy significant features are complementary with these four perimeter related features. The number of peaks in the variations of the perimeter for example are a significant effect that indicates the number of transitions between contracted and expanded posture. This conclusion confirmed previous studies that revealed the significant role of body spatial occupation to discriminate between emotions [1,6].

Our results also show more specific differences between emotion conditions. Post-hoc analyses (Bonferroni correction) of different emotional expressions were performed with respect to the significant features previously identified. Anger and Joy (high arousal) present significantly higher values than Relief and Sadness (low arousal) in the majority of energy and perimeter features (see Table 4 and 5).

Emotion (I)	Emotion (J)	Mean Difference (I-J)			
		Largest peak release of energy	***Max value of energy***	***Mean value of energy***	***Standard deviation value of energy***
Anger	Relief	-226.92*	46.4*	17.00*	10.96*
	Sad	-237.3*	58.6*	20.00*	14.15*
Joy	Relief	-239.76*	53.3*	15.09*	13.33*
	Sad	-250.15*	65.5*	18.09*	16.52*

* $p<.05$

Table 4 Significant differences for motion cue **energy**

In *standard deviation of energy*, Anger and Joy were also higher revealing a more important dispersion of the value of energy. This superior value of variance means that when changes occur, they are of larger size.

In *Largest peak release*, Anger and Joy were also higher than Relief and Sadness. It means that in the anger and joy conditions, as expected, ending movements following the largest peak decreased more abruptly.

Emotion (I)	Emotion (J)	Mean Difference (I-J)	
		Max value of perimeter	***Peaks number***
Anger	Relief	24.75*	3.20*
	Sad	33.34*	3.60*
Joy	Relief	35.80*	3.30*
	Sad	44.39*	3.70*

* $p<.05$

Table 5 Significant differences for motion cue **perimeter**

Anger and Joy also differed from Relief and Sadness in two perimeter related features. In the Anger and Joy conditions, the amplitude of perimeters changes is higher (*max value for perimeter*) and the number of local maxima increases (*number of peaks for perimeter*). This means that the spatial occupation of the actors also varies accordingly to the activation level that characterizes the two groups of emotions. It also means that these variations are more impulsive and frequent when this activation level is higher.

These findings show that the sum of modules of velocities (energy) and the variations of a bounding triangle perimeter enable to distinguish between active and passive emotions (i.e. between joy, anger and relief and sadness). These results also confirm, in a quantitative way, the previous findings by Walbott: a collection of emotions characterized by their arousal can be related to the qualitative aspects of gestures like the total activation of the movement and by the space occupation of the person. Further analysis we are currently carrying out aim at detailing the significant effect of the emotion at the level of the head and the hands separately.

5. Conclusion

The intention of this paper is to introduce the concept of an automated video analysis of human gesture dynamics for emotion recognition. The described experiment's non-intrusive set up allowed us to analyze human emotional behavior in a more flexible environment with regard to lighting and actor appearance, therefore eliminating the need for body markers.

We presented and analysed the results of our approach and found that the energy cue is the most significant cue in differentiating between the emotions, with a minor but significant role played by the perimeter cue.

With this study we investigated the role of movement expressivity rather than describing the form of the gesture itself. Results showed how expressive motion cues allow one to discriminate between "high" and "low arousal" emotions. Such analysis will be extended to a larger core-set of 150 portrayals.

Rating studies are also planned to confirm the validity of our motor cues on the perceptual level. Point light animations consisting of the three blobs corresponding to the head and the hands of the actors will be used as stimuli. Participants will be evaluated on their ability to discriminate between the selected four emotions on the basis of such impoverished display.

Further work will consist of performing multimodal analysis by integrating these results with analysis of prosody and facial expression of the same portrayals.

References

[1] Wallbott, H. G. (1998). Bodily expression of emotion. European Journal of Social Psychology, 28, 879-896.

[2] Picard, R.W.: Affective Computing. The MIT Press, Cambridge (1997)
[3] Blake, R., & Shiffrar, M., (2007) Perception of Human Motion Annual Review of Psychology, Vol. 58, January 2007
[4] Banziger, T., & Scherer, K.R. (2007). Using actors portrayals to systematically study multimodal emotion expression: the GEMEP corpus. In A. Paiva, R. Prada & R.W. Picard (Eds.), Lecture Notes in Computer Science: vol. 4738. Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings (pp. 59 - 70). Berlin: Springer Verlag.
[5] Pollick, F. E., Paterson, H., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. Cognition, 82. B51-B61.
[6] De Meijer, M.: The contribution of general features of body movement to the attribution of emotions. Journal of Nonverbal Behavior 13, 247–268 (1989)
[7] Ekman, P.Rosenberg, E.L (1997) What the Face Reveals Basic and Applied Studies of Spontaneous Expression. Oxford University Press
[8] Camurri, A., Lagerlöf, I, & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. International Journal of Human-Computer Studies, 59(1-2), 213-225.
[9] Kapur, A., Kapur, A., Babul, N.V., Tzanetakis, G., Driessen, P.F.: Gesture-based affective computing on motion capture data. In: ACII, pp. 1–7 (2005)
[10] Castellano, G, Villalba, S. D., & Camurri, A. (2007). Recognising Human Emotions from Body Movement and Gesture Dynamics. In A. Paiva, R. Prada & R.W. Picard (Eds.), Lecture Notes in Computer Science: vol. 4738. Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon
[11] Bernhardt, D., & Robinson, P. (2007). Detecting affect from non-stylised body motions. In A. Paiva, R. Prada & R.W. Picard (Eds.), Lecture Notes in Computer Science: vol. 4738. Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings (pp. 59 - 70). Berlin: Springer Verlag.
[12] A. Camurri, A., Coletta, P., Varni, G., & Ghisio, S. (2007). Developing multimodal interactive systems with EyesWeb XMI. Proceedings of the 2007 conference on new interfaces for musical expression (NIME07) (pp. 305-308). New York, USA.
[13] Wundt, W. (1896). Lecture on human and animal psychology. New York:MacMillan
[14] Russell, J.A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology: 39, 1161–1178.
[15] Jennings, J. R., & Wood, C. C., (1976). The e-adjustment procedure for repeated-measures analyses of variance. Psychophysiology, 13, 277-278.
[16] Gabrielsson,A. & Juslin, P.N. (2003) Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. Psychology of Music, Vol. 24, No. 1, 68-91, 1996.