Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews

J.-C. Martin[1], G. Caridakis[2], L. Devillers[1], K. Karpouzis[2], S. Abrilian[1]

1 LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
{martin, devil, abrilian}@limsi.fr
2 Image, Video and Multimedia Systems Lab, National Technical
University of Athens, Iroon Polytechniou 9, GR-157 80 Athens, Greece,
{kkarpou, gcari}@image.ece.ntua.gr

Abstract

There has been a lot of psychological researches on emotion and nonverbal communication. Yet, these studies were based mostly on acted basic emotions. This paper explores how manual annotation and image processing can cooperate towards the representation of spontaneous emotional behaviour in low resolution videos from TV. We describe a corpus of TV interviews and the manual annotations that have been defined. We explain the image processing algorithms that have been designed for the automatic estimation of movement quantity. Finally, we explore how image processing can be used for the validation of manual annotations.

1. Introduction

There has been a lot of psychological researches on emotion and nonverbal communication of facial expressions of emotions (Ekman, 1999), and on expressive body movements (Boone & Cunningham, 1998; DeMeijer, 1989; Newlove, 1993; Wallbott, 1998). Yet, these psychological studies were based mostly on acted basic emotions: anger, disgust, fear, joy, sadness, surprise. In the area of affective computing, recent studies of non-verbal behaviour during emotions are also limited with respect to the number of modalities or the spontaneity of the emotion : markers on body to recognise four acted basic emotions (Kapur et al., 2005), motion capture of static postures during acting of two nuances of four basic emotions (De Silva et al., 2005), video processing of facial expressions and upper body gestures during six acted emotional behaviours (Gunes & Piccardi, 2005).

Most of these studies are dealing with basic acted emotions, and real-life multimodal corpora are very few despite the general agreement that it is necessary to collect audio-visual databases that highlight naturalistic expressions of emotions (Douglas-Cowie et al., 2003).

Indeed, building a multimodal corpus of real-life emotions is challenging since it involves subjective perception and requires time consuming manual annotations of emotion at several levels. This manual annotation might benefit from image processing via the automatic detection of emotionally relevant video segments. Estimation of movement quantity by automatic image processing might validate the manual annotations of movements during the time-based annotation of the video, and also of emotional activation at the level of the whole video. Finally automatic annotation might ease the manual annotation process by providing movement segmentation and precise values of expressive parameters such as the speed, the spatial expansion or the fluidity of a gesture. Yet, manual annotation and image processing provide information at different levels of abstraction and their integration is not straightforward. Furthermore, most of the work in image processing of emotional behaviour has been done on high quality videos recorded in laboratory situations where emotions might be less spontaneous than during non staged TV interviews.

The goals of this paper are to explore 1) the applicability of image processing techniques to low resolution videos from TV, and 2) how image processing might be used for the validation of manual annotation of spontaneous emotional behaviour.

Section 2 describes the corpus of TV interviews that has been collected and the manual annotations that have been defined. Section 3 explains the image processing algorithms that have been designed for the automatic estimation of movement quantity. A preliminary study with 3 videos was already presented in (Martin et al., 2006). Section 4 explores several ways to compare the manual annotations and the results of image processing with the illustration of 10 videos.

2. Manual annotation of multimodal emotional behaviours

The EmoTV corpus features 50 video samples of emotional TV interviews (Abrilian et al., 2005). The videos are encoded in Cinepak Codec by CTi (720x576, 25 images/sec). The goal of the EmoTV corpus is to provide knowledge on the coordination between modalities during non-acted emotionally rich behaviours. Thus, a multilevel coding scheme has been designed and enables the representation of emotion at several levels of temporality and abstraction (Devillers et al., 2005). At the global level there is the annotation of emotion (categorical and dimensional including global activation). Similar annotations are available at the level of emotional segments of the video.

At the level of multimodal behaviours (Martin, Abrilian, & Devillers, 2005) there are tracks for each visible modality: torso, head, shoulders, facial expressions, gaze, and hand gestures. The head, torso and hand tracks contain a description of the pose and the movement of these modalities. Pose and movement annotations thus alternate. Regarding the annotation of emotional movements, we inspired our annotation scheme of the expressivity model proposed by (Hartmann et al., 2005) which describes expressivity by a set of six dimensions: spatial extent, temporal extent, power, fluidity, repetition, overall activity. Movement quality is

thus annotated for torso, head, shoulders, and hand gestures.

For hand gestures annotation, we have kept the classical attributes (Kipp, 2004; McNeill, 1992). Our coding scheme thus enables not only the annotation of movement expressivity but also the annotation of the structural descriptions ("phases") of gestures as their temporal patterns might be related to emotion: preparation (bringing arm and hand into stroke position), stroke (the most energetic part of the gesture), sequence of strokes (a number of successive strokes), hold (a phase of stillness just before or just after the stroke), and retract (movement back to rest position). We have selected the following set of gestures functions ("phrase") as they revealed to be observed in our corpus: manipulator (contact with body or object), beat (synchronized with the emphasis of the speech), deictic (arm or hand is used to point at an existing or imaginary object), illustrator (represents attributes, actions, relationships about objects and characters), emblem (movement with a precise, culturally defined meaning). Currently, the hand shape is not annotated since it is not considered as a main feature of emotional behaviour in our survey of experimental studies nor in our videos.

Whereas the annotations of emotions have been done by 3 coders and lead to computation of agreement (Devillers et al., 2005), the current protocol used for the validation of the annotations of multimodal behaviours is to have a 2nd coder check the annotations followed by discussions. Although we are also considering the validation of the annotations by the automatic computation of inter-coder agreements from the annotations of multimodal behaviours by several coders, automatic image processing might provide an alternative means for validating the manual annotation.

3. Automatic processing of videos of emotional behaviours

Image processing is used to provide estimations of head and hand movements by combining 1) the location of skin areas and 2) the estimation of movement. The task of head and hand localization in image sequences is based on detecting continuous areas of skin colour. For the given application, a very coarse model is sufficient, since there is no need for recognition of hand shape. As mentioned before the examined corpus is based on real-life situations and therefore the person's original posture is arbitrary and not subject to spatial constraints such as "right hand on the right side of the head" when the person's hands are crossed. In addition to this some skin-like regions may mislead the automatic detection and tracking algorithm. To tackle the above problems a user-assisted initialization process is required as the starting point for the tracking algorithm. During this process the user confirms the regions suggested by the system as the hands and head of the person participating in the multimodal corpora; after that, since lighting and colour conditions do not usually change within the clip, detection and tracking are performed automatically. Another usual impediment to image processing of TV videos is the fact that camera movement can be uncontrolled and may result in skin regions moving abruptly within a clip without the subject showing the relevant activity. In our approach, this can be tackled by taking into account the change of the relevant positions of the skin regions, since they will not change in the event of sudden camera movement.

The measure of movement in subsequent frames is calculated as the sum of the moving pixels in the moving skin masks, normalized over the area of the skin regions. Normalization is performed in order to discard the camera zoom factor, which may make moving skin regions appear larger without actually showing more vivid activity. Possible moving areas are found by thresholding the difference pixels between the current frame and the next, resulting to the possible motion mask. This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating further tracking only in moving image areas. Both colour and motion masks contain a large number of small objects due to the presence of noise and objects with colour similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. In the following, the moving skin mask is created by fusing the processed skin and motion masks, through the morphological reconstruction of the colour mask using the motion mask as marker.

Overall activation is considered as the quantity of movement. In our case it is computed as the sum of the motion vectors' norm (Eq. 1).

$$OA = \sum_{i=0}^{n} \left| \vec{r}(i) \right| + \left| \vec{l}(i) \right| \quad \textbf{(1)}$$

Spatial extent is modelled by expanding or condensing the entire space in front of the agent that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands (Eq. 2). The average spatial extent is also calculated for normalization reasons. The temporal parameter of the gesture determines the speed of the arm movement of a gesture's meaning carrying stroke phase and also signifies the duration of movements (e.g., quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. The power actually is identical with the first derivative of the motion vectors calculated in the first steps.

$$SE = \max\left(\left| d(\vec{r}(i) - \vec{l}(i)) \right| \right) \quad \textbf{(2)}$$

4. Comparing manual annotations and automatic processing

In this section we illustrate the comparison of manual and automatic processing on 10 videos from the EmoTV corpus (Table 1).

Video #	Informal description
01	Street ; head movement, facial expressions ; people moving in the background
02	Street ; movement (torso, hand, head) ; people moving in background
03	Street ; movement (torso, hand, head) ; facial expressions ; skin area on the torso
22	Beach ; movement (torso, hand, head) ; facial expressions ; several skin areas (swimming clothes)
36	Inside dark ; movement (hand, head) ; facial expressions ; people moving in the background
41	Inside ; head movement ; facial expressions
44	Outside ; facial expressions ; movement (head, hand)
49	Outside ; movement (hand, head) ; facial expressions ; people moving in the background
71	Inside ; movement (hand, head) ; facial expressions ; people moving in the background
72	Outside ; movement (hand, head)

Table 1. Informal description of the 10 videos used for the study

Video #	**(1) Manual annotation of emotional activation 1:low 5: high**	**(2) Automatic estimation of movement quantity**	**(3) Manual % of sec. with at least 1 manual annotation of movement in Anvil files**
Coders	3 expert coders	System	1 expert coder + checked by 2nd coder
01	4	3398,50	81,2
02	3	269,64	72,9
03	4,33	1132,80	92,6
22	4,33	3282,80	81,1
36	4,66	2240,50	94,4
41	3	959,60	73,6
44	3,33	1771,30	92,3
49	4,33	1779,00	91,2
71	2,67	904,73	86,1
72	3,33	330,92	56,7

Table 2. Manual measures (1)(3), and automatic measure (2) of global emotional activation in the 10 videos

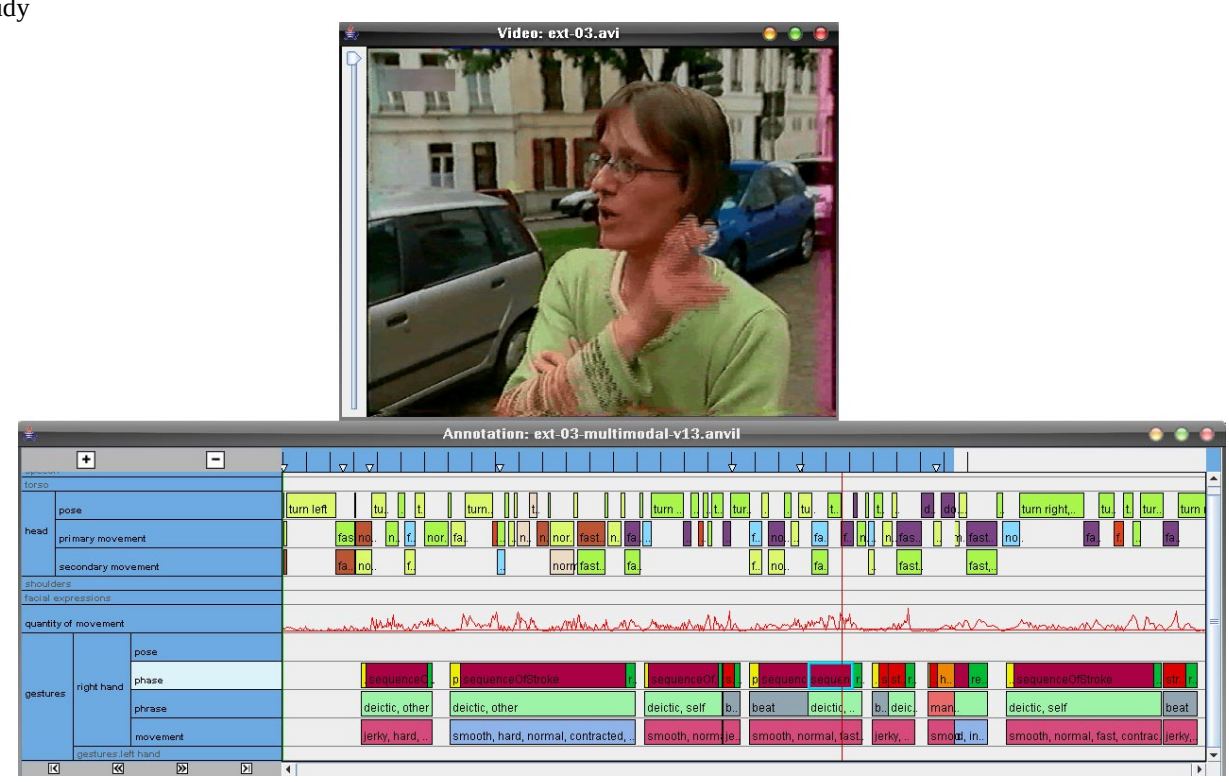

Figure 1. Integration in the Anvil tool (Kipp 2004) of manual annotations (coloured blocks) of head movements (upper tracks) and hand gestures (lower tracks), and results of image processing for the estimation of movement quantity (continuous line in the middle).

4.1. Global activation of emotional behaviours in each video

The values obtained for (1) the manual annotation of emotional activation (average of annotations by 3 expert coders), (2) the automatic estimation of movement quantity at the level of the whole video clip, 3) the % of seconds of each video for which there is at least one manual annotation of movement (either head, hand or torso) are given in Table 2.

These three values provide different estimations of the quantity of multimodal activity related to the emotion. The correlation analysis suggests that measures (1) and (2) are significatively correlated ($r = 0,64$, $p < 0,05$). This shows that the automatic processing of our 10 videos validates the manual annotation of activation at the global level of each video.

The correlation analysis also suggests that (1) and (3) may be correlated ($r = 0.49$). Finally, the correlation analysis suggests that (2) and (3) may be correlated ($r = 0,42$). However, due to the small sample size, these two measures do not reach statistical significance. More data are needed to confirm these two results.

4.2. Time-based estimation and annotation of movement

At the local time-based level, we were willing to compare the manual annotations (of the movements of the head, hands and torso) with the automatic estimation of movements. Figure 1 shows how both types of annotations have been included under the Anvil tool (Kipp, 2004).

The current image processing module enables to provide an estimation of the movement between each frame for the whole image. It does not provide separate estimations of movement for the different body parts (e.g. image areas). Thus, we compared the union of the manual annotations of movements in the head, hands and torso modalities with the automatic estimation of movements for the whole frame. When the image processing module detected a movement, we decided that there would be an agreement with the manual annotations if a movement had been manually annotated in at least one of the three body parts.

The continuous values of motion estimation provided by the image processing module need to be thresholded in order to provide a Boolean automatic annotation of movements that can be compared with the manual annotations. Setting different values to this threshold for automatic movement detection leads to different values of agreement between the manual annotations and the automatic detection of movement. The value of this amplitude threshold above which the image processing module decides that a movement has been detected should be the minimal value at which a movement should have been perceived and annotated. We evaluated the agreement between the union of the manual annotations of movements and the estimation of movement with several values of this amplitude threshold above which the image processing module decides that a movement is detected. The tested values for this threshold were between 0.1% and 40% of the maximal value of estimation of movement quantity. We use a 0,04 s. time interval for computing the agreement between manual and automatic annotations since it is the interval between 2 frames used by the automatic processing module.

The resulting confusion matrix is provided in Table 3. The agreement is the highest for videos 22 and 3 which feature many movements (head, hand) and in which the skin is visible in the upper area of the torso, and in which there is nobody moving in the background. The lowest agreement is obtained for videos 36 and 71 which feature people moving in the background, the movement of whom have not been manually annotated since we focus on interviewed people. An intermediate value is obtained for video 41 which only features slight movements of the head and a few movements of the torso. The interviews recorded outside get a higher agreement than those recorded inside, revealing the impact of video quality and lightness.

There is no systematic relations between the disagreements: for 6 videos, the number of disagreement "auto 0 – manual 1" is higher than the number of disagreement "auto 1 – manual 0".

Video #	**Threshold**	**Agreements**		
		Auto 0 Manual 0	**Auto 1 Manual 1**	**Total**
01	0,004	0,050	0,799	0,849
02	0,004	0,203	0,611	0,814
03	0,001	0,009	0,892	0,901
22	0,016	0,113	0,799	0,912
36	0,001	0,039	0,449	0,489
41	0,002	0,186	0,483	0,669
44	0,001	0,063	0,550	0,613
49	0,003	0,013	0,858	0,871
71	0,042	0,139	0,307	0,446
72	0,047	0,340	0,355	0,695
AVG	**0,012**	**0,115**	**0,610**	**0,726**

Video #	**Threshold**	**Disagreements**		
		Auto 0 Manual 1	**Auto 1 Manual 0**	**Total**
01	0,004	0,014	0,138	0,151
02	0,004	0,118	0,068	0,186
03	0,001	0,034	0,065	0,099
22	0,016	0,013	0,075	0,088
36	0,001	0,494	0,017	0,511
41	0,002	0,254	0,077	0,331
44	0,001	0,373	0,014	0,387
49	0,003	0,054	0,075	0,129
71	0,042	0,554	0,000	0,554
72	0,047	0,213	0,092	0,305
AVG	**0,012**	**0,212**	**0,062**	**0,274**

Table 3. Confusion matrix between manual annotation of movement and automatic estimation of movement quantity (for example the column "Auto 0 – Manual 0" describes the agreements "no manual annotation of movements" / "no automatic detection of movement"). The threshold is multiplied by the maximum value of movement estimation.

These 10 videos from EmoTV are rich in manual annotation of movements of either hand, torso or head (for example, the % of frames for which there is no manual annotation of movements are only 26% for video 41, 7% for video 3, and 5% for video 36).

Thus, in order to be able to compute statistical measures of the agreement between manual and automatic annotations, we balanced the number of frames with and without manual annotation by 1) computing the number of frames without any manual annotation of movement, and 2) by a random selection of the same number of frames but with a manual annotation of movement. The resulting confusion matrix is provided in Table 4. The new average agreement is higher (0,794) than the one obtained in Table 3 without a balanced number of frames (0,726). Table 4 also reveals that the disagreements are not balanced anymore: the number of frames for which there was a manual annotation of movement and for which no movement was detected by image processing is higher than the reverse for 8 of the 10 videos.

Video #	**Threshold**	**Agreements**		
		Auto 0 Manual 0	**Auto 1 Manual 1**	**Total**
01	0,047	0,353	0,358	0,711
02	0,013	0,418	0,407	0,825
03	0,076	0,442	0,423	0,865
22	0,071	0,450	0,467	0,917
36	0,034	0,500	0,300	0,800
41	0,008	0,421	0,283	0,704
44	0,010	0,460	0,316	0,776
49	0,084	0,476	0,357	0,833
71	0,048	0,500	0,283	0,783
72	0,044	0,385	0,345	0,730
AVG	**0,043**	**0,440**	**0,354**	**0,794**

Video #	**Threshold**	**Disagreements**		
		Auto 0 Manual 1	**Auto 1 Manual 0**	**Total**
01	0,047	0,142	0,147	0,289
02	0,013	0,093	0,082	0,175
03	0,076	0,077	0,058	0,135
22	0,071	0,033	0,050	0,083
36	0,034	0,200	0,000	0,200
41	0,008	0,216	0,080	0,296
44	0,010	0,185	0,039	0,224
49	0,084	0,143	0,024	0,167
71	0,048	0,217	0,000	0,217
72	0,044	0,155	0,115	0,270
AVG	**0,043**	**0,146**	**0,059**	**0,206**

Table 4. Confusion matrix between manual annotation of movement and automatic estimation of movement quantity for a balanced set of frames with or without manual annotation of movement

The maximum kappa values and the threshold for which they were obtained are listed in Table 5, column (1). The resulting kappa values range between 0,422 and 0,833 depending on the videos. These values can be considered as rather good given the resolution of our TV videos.

In the results described in Table 5 column (1), we selected the thresholds as the values providing the maximum kappa values. The differences between the corresponding thresholds obtained for the different videos show that this threshold value needs to be customised for each video, probably due to the differences between the different interviews settings and video qualities.

We explored the use of phases of each video during which no (or very little) movement is perceptually visible. We computed the average movement estimation provided by the automatic processing module during each of these phases. Using this average value as the threshold lead to a lower average kappa value (Table 5, column (2)). Further experimental explorations are thus required to study how this threshold value can be set.

	(1) Threshold corresponding to max kappa		**(2) Threshold selected from 2 s. without movement**	
Video #	**Max kappa**	**Threshold**	**Kappa**	**Threshold**
01	0,422	0,047	0,275	0,056
02	0,649	0,013	0,547	0,016
03	0,731	0,076	0,57	0,059
22	0,833	0,071	0,633	0,079
36	0,600	0,034	0,6	0,037
41	0,407	0,008	0,342	0,039
44	0,553	0,01	0,19	0,066
49	0,667	0,084	0,428	0,089
71	0,565	0,048	0,304	0,008
72	0,459	0,044	0,327	0,034
AVG	**0,589**	**0,043**	**0,421**	**0,048**

Table 5. Kappa values obtained for the same number of frames which involve a manual annotation of movement and the number of frames which do not involve a manual annotation of movement: (1) The displayed threshold is the one for which the kappa value is maximum, (2) the displayed threshold was obtained by averaging the automatic estimation of movement in a 2 s. part of the videos for which no movement can be perceived when playing the videos.

5. Conclusion

We have explored in this paper how automatic image processing can validate the manual annotation of emotional movement in TV interviews, either at the global level of the whole clip, or at the level of individual annotation of movements. Other means of comparing manual and automatic annotations will be investigated.

Future directions will include the separate estimation of movement quantity for different body parts of the image (including tracking of these areas) in order to cope with people moving in the background, the automatic extraction of values for the expressive parameters such the spatial extent (Eq. 2), the validation of the manual annotation of activation at the level of emotional segments of the videos, the relations between the estimation of movement quantity and the gesture phases (preparation, stroke, retraction), the use of temporal filters for improving the automatic detection of movements, and

finally the inclusion of torso annotation in the union of movement annotation only if it includes a skin area.

The study described in this paper may have several applications. For example, designing affective Human Computer-Interfaces such as Embodied Conversational Agents which requires modelling the relations between spontaneous emotions and behaviours in several modalities (Martin, Abrilian, Devillers et al., 2005).

Acknowledgement

This work was partly funded by the FP6 IST HUMAINE Network of Excellence (http://emotion-research.net).

6. References

Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C. (2005). *EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces.* 11th International Conference on Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July.

Boone, R. T., & Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: The development of cue attunement. *Developmental Psychology, 34*(5), 1007-1016.

De Silva, P. R., Kleinsmith, A., & Bianchi-Berthouze, N. (2005). *Towards unsupervised detection of affective body posture nuances.* 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

DeMeijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*(13), 247 - 268.

Devillers, L., Abrilian, S., & Martin, J.-C. (2005). *Representing real life emotions in audiovisual data with non basic emotional patterns and context features.* First International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech; Towards a new generation of databases. *Speech Communication*(40).

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition & Emotion* (pp. 301–320). New York: John Wiley.

Gunes, H., & Piccardi, M. (2005). *Fusing Face and Body Display for Bi-modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration.* 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

Hartmann, B., Mancini, M., & Pelachaud, C. (2005). *Implementing Expressive Gesture Synthesis for Embodied Conversational Agents.* Gesture Workshop (GW'2005), Vannes, France, May.

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P. F. (2005). *Gesture-Based Affective Computing on Motion Capture Data.* 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

Kipp, M. (2004). *Gesture Generation by Imitation. From Human Behavior to Computer Character Animation.* Florida: Boca Raton, Dissertation.com.

Martin, J.-C., Abrilian, S., & Devillers, L. (2005). *Annotating Multimodal Behaviors Occurring during Non Basic Emotions.* 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24.

Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., & Pelachaud, C. (2005). *Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs.* 5th International Working Conference On Intelligent Virtual Agents (IVA'2005), Kos, Greece, September 12-14.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., & Abrilian, S. (2006). *Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors in TV Interviews.* 3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI'2006), Athens, Greece, 7-9 June.

McNeill, D. (1992). *Hand and mind - what gestures reveal about thoughts*: University of Chicago Press, IL.

Newlove, J. (1993). *Laban for actors and dancers*. New York: Routledge.

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