Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors in TV Interviews

Jean-Claude Martin[1], George Caridakis[2], Laurence Devillers[1],
Kostas Karpouzis[2], Sarkis 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. Designing affective Human Computer-Interfaces such as Embodied Conversational Agents requires modeling the relations between spontaneous emotions and behaviors in several modalities. There have 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 might cooperate towards the representation of spontaneous emotional behavior 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 several ways to compare the manual annotations and the cues extracted by image processing.

1 Introduction

Designing affective Human Computer-Interfaces such as Embodied Conversational Agents requires modeling the relations between spontaneous emotions and behaviors in several modalities. There has been a lot of psychological researches on emotion and nonverbal communication of facial expressions of emotions [8], and on expressive body movements [2, 5, 17, 18]. 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 behavior during emotions are also limited with respect to the number of modalities or the spontaneity of the emotion. For example, cameras are used by [13] to capture markers placed on

Please use the following format when citing this chapter:
Martin, Jean-Claude, Caridakis, George, Devillers, Laurence, Karpouzis, Kostas, Abrilian, Sarkis, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 369–377

various points of the whole body in order to recognize four acted basic emotions (sadness, joy, anger, fear).

With respect to other modalities than facial expressions, static postures were recorded by De Silva et al. [4] using a motion capture system during acted emotions (two nuances for each of four basic emotions; e.g. upset and angry as nuances of anger). In Gunes et al. [11] the video processing of facial expressions and upper body gestures are fused in order to recognize six acted emotional behaviors (anxiety, anger, disgust, fear, happiness, uncertainty). A vision based system that infers acted mental states (agreeing, concentrating, disagreeing, interested, thinking, and unsure) from head movements and facial expressions is described in el Kaliouby et al. [10]. Choi et al. [3] describe how video processing of facial expressions and gaze are mapped onto combinations of emotions (neutral, surprise, fear, sadness, anger, disgust, happiness).

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 [7].

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. Automatic processing might provide finer numerical values which are not possible with manual annotations. 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. 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 behavior 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 1) to explore the applicability of image processing techniques for low resolution videos from TV, and 2) explore how manual annotation and image processing might cooperate towards the representation of spontaneous emotional behavior. 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. Section 3 explores several ways to compare the manual annotations and the results of image processing with the illustration of three video samples.

2 Manual annotation of multimodal emotional behaviors

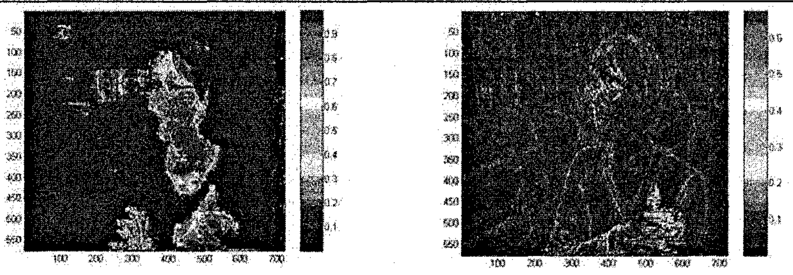
The EmoTV corpus features 50 video samples of emotional TV interviews [1]. The videos are encoded in Cinepak Codec by CTi (720x576, 25 images/sec). The goal of the EmoTV corpus is to provide knowledge on the coordination between modalities during non-acted emotionally rich behaviors. A multilevel coding scheme has been designed and enables the representation of emotion at several levels of temporality and abstraction [6]. 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 behaviors [15] 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 movements, we inspired our annotation scheme of the expressivity model proposed by [12] 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 gestures annotation, we have kept the classical attributes used for gesture annotation [14, 16] but focused on repetitive and manipulator gestures which occur frequently in EmoTV. 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 behavior 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 [6], the current protocol used for the validation of the annotations of multimodal behaviors is to have a 2nd coder check the annotations followed by discussions. We are considering the validation of the annotations by the automatic computation of inter-coder agreements from the annotations by several coders.

3 Automatic processing of videos of emotional behaviors

Image processing is used to provide estimations of head and hand movements by combining 1) location of skin areas and 2) the estimation of movement (Fig. 1). The task of head and hand localization in image sequences is based on detecting continuous areas of skin color. 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 color 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.



(a) skin detection (b) Movement estimation of all colors

Fig. 1. Steps in image processing for automatic estimation of movement quantity: (a) skin detection, (b) movement estimation of all colors. The next step is to compute the intersection of (a) and (b) for estimating movement of skin areas

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 color and motion masks contain a large number of small objects due to the presence of noise and objects with color 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 color mask using the motion mask as marker.

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

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

Spatial extent is modeled 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.

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

We illustrate our approach on the combination of image processing and manual annotation on a video of the EmoTV corpus (duration 29 seconds, frame rate 25 fps, 722 frames). The image processing module provides information related to emotional behavior at two levels: 1) a global level of the whole video clip, and 2) a local level (e.g. between two frames). At the global level of the whole video, an estimation of the overall activation is computed. For the video 3, this overall activation (Eq. 1) normalized by the number of frames is 1340. It was compared with the results obtained with two laboratory recorded videos with different behaviors but similar viewpoint. The overall activation for a video with fewer movements (showing a single gesture) was smaller (44). For a video with more activation (showing several repetitive gestures), this value was higher (2167).

After the user-assisted initialization step the tracking algorithm is responsible for classifying the skin regions in the following frames of the examined video. Skin region size, distance wrt the previous classified position of the region, flow alignment and spatial constraints. These criteria ensure that the next region selected to replace the current one is approximately the same size, close to the last position and moves along the same direction as the previous one as long as the instantaneous speed is above a certain threshold. As a result each candidate region is being awarded a bonus for satisfying these criteria or is being penalized for failing to comply with the restrictions applied. The winner region, the one that collects the most points during this process, is appointed as the reference region for the next frame. The criteria don't have an eliminating effect, meaning that if a region fails to satisfy one of them is not being excluded from the process, and the bonus or penalty given to the region is relative to the score achieved in every criterion test. The finally selected region's score is thresholded so that poor scoring winning regions are excluded. In this case the position of the body part is unchanged wrt that in the previous frame. This feature is especially useful in occlusion cases when the position of the body part remains the same as just before occlusion occurs. After a certain number of frames the whole process is reinitialized so that a possible misclassification is not propagated.

4 Comparing manual annotations and automatic processing

In this section we illustrate the comparison of manual and automatic processing on three videos from the EmoTV corpus. These three investigated videos have different profiles. Video 41 which includes only head movement. Video 3 includes torso and hand movements. Video 36 includes movements of other people in the background.

4.1 Global activation of behaviors in each video

The values obtained for 1) the manual annotation of emotional activation, 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 1.

Video #	41	3	36
(1) Emotional activation (manual annotation) 1:low activation, 5: high activation	3	4,33	4,66
(2) Estimation of movement quantity (automatic image processing)	959,60	1132,80	2240,50
(3) % of sec. for which there is at least one manual annotation of movement (head, hand or torso)	73,6	92,6	94,4

Table 1. Values of three manual and automatic measures of emotional activation in three videos

These three values provide different estimations of the quantity of multimodal activity related to the emotions. These values are consistent with the different profiles of these videos. The correlation analysis suggests that measures (1) and (3) may be correlated (r = 0,97). This might reveal a consistency in the manual annotation process. The correlation analysis also suggests that (1) and (2) may be correlated (r = 0.74). This shows that the automatic processing of videos might be useful for validating the manual annotation of activation at the global level of each video. Finally, the correlation analysis suggests that (2) and (3) may be correlated (r = 0,58). However, due to the small sample size, these three values do not reach statistical significance. More data are needed to confirm such a result.

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. The current state of the image processing module enables to provide an estimation of the movement between each frame for the whole image. The current image processing module 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. 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 modalities.

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 the 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 has been 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. interval for computing the agreement between manual and automatic annotations since it is the interval between 2 frames used by the automatic processing.

The resulting confusion matrix is provided in Table 2. The agreement is the highest for video 3 which features many movements of the head, hand and the upper area of the torso where the skin is visible. The lowest agreement is obtained with video 36 which features 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. These three videos from EmoTV are rich in annotation of movements of either hand, torso or head. The % of frames for which there is no manual annotation of movements are 26% for video 41, 7% for video 3, and 5% for video 36.

Table 2. 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)

Video #		**Agreements**			**Disagreements**	
	Auto 0 **Manual 0**	**Auto 1** **Manual 1**	**Total**	**Auto 0** **Manual 1**	**Auto 1** **Manual 0**	**Total**
3	1%	89%	90%	3,5%	6,5%	10%
41	19%	48%	67%	25%	8%	26%
36	4%	45%	49%	49%	2%	51%

In order 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 with a manual annotation of movement. For video 3, a threshold for motion detection of 8% of maximum movement quantity, leads to a maximum kappa (0,71). For video 36, the maximum kappa is 0,6 (threshold 9%). For video 41, the maximum kappa is 0,425 (threshold 0.6%). For the three videos the kappa is very low when the threshold is too low (the system considers that there is always a movement from automatic processing, and when compared to the manual annotation, the agreement is lower). Then the kappa increases until reaching its highest value (e.g. the values described above), and then decreases as the threshold becomes higher.

5 Conclusions and future directions

In this paper we have described an exploratory approach aiming at computing various relations between manual and automatic annotations of videos of multimodal emotional behaviors. We observed that some dimensions of manual annotations and results of automatic might be correlated. A next step is to consider other videos in order to reach statistical significance in the comparison of manual annotation of activation and the automatic estimation of movement quantity. We will consider videos with very few movements in order to be able to compare the manual and the automatic annotations with classical kappa measures for all the annotations.

Future direction also include the use of temporal filters for improving the automatic detection of movements, the separate estimation of movement quantity for different body parts of the image (including tracking of these areas), 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 segment, the relations between the estimation of movement quantity and gesture phases (preparation, stroke, retraction), the inclusion of torso annotation in the union of movement annotation only if it includes a skin area.

Acknowledgments

This research is partly supported by the EC Project HUMAINE (IST-507422).

References

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

2. Boone, R. T., Cunningham, J. G.: Children's decoding of emotion in expressive body movement: The development of cue attunement. Developmental Psychology 34 5 (1998)
3. Choi, S. M., Kim, Y. G.: An Affective User Interface Based on Facial Expression Recognition and Eye Gaze Tracking. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 907-915
4. De Silva, P. R., Kleinsmith, A., Bianchi-Berthouze, N.: Towards unsupervised detection of affective body posture nuances. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 32-40
5. DeMeijer, M.: The contribution of general features of body movement to the attribution of emotions. Journal of Nonverbal Behavior 13 (1989)
6. Devillers, L., Abrilian, S., Martin, J.-C.: 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) (2005) Beijing, China 519-526
7. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech; Towards a new generation of databases. Speech Communication 40 (2003)
8. Ekman, P.: Basic emotions. Handbook of Cognition & Emotion. J. Wiley (1999)
9. Ekman, P., Walla, F.: Facial Action Coding System (FACS). (1978)
10. el Kaliouby, R., Robinson, P.: Generalization of a Vision based Computational Model of Mind Reading. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 582-590
11. Gunes, H., Piccardi, M.: 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) (2005) Beijing, China 102-110
12. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop (GW'2005) (2005) Vannes, France
13. Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., Driessen, P. F.: Gesture Based Affective Computing on Motion Capture Data. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 1-8
14. Kipp, M.: Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Boca Raton, Dissertation.com Florida (2004)
15. Martin, J.-C., Abrilian, S., Devillers, L.: Annotating Multimodal Behaviors Occurring during Non Basic Emotions. 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005) (2005) Beijing, China 550-557
16. McNeill, D.: Hand and mind - what gestures reveal about thoughts. University of Chicago Press, IL (1992)
17. Newlove, J.: Laban for actors and dancers. Routledge New York (1993)
18. Wallbott, H. G.: Bodily expression of emotion. European Journal of Social Psychology 28 (1998)