Modeling emotional expressions as sequences of behaviors

Radosław Niewiadomski[1], Sylwia Hyniewska[2], and Catherine Pelachaud[3]

[1] Telecom ParisTech, Paris, France
niewiado@telecom-paristech.fr
[2] Telecom ParisTech, Paris, France
hyniewska@telecom-paristech.fr
[3] CNRS-LTCI, Paris, France
pelachau@telecom-paristech.fr

Abstract. In this paper we present a system which allows a virtual character to display multimodal sequential expressions i.e. expressions that are composed of different signals partially ordered in time and belonging to different nonverbal communicative channels. It is composed of a language for the description of such expressions from real data and of an algorithm that uses this description to automatically generate emotional displays. We explain in detail the process of creating multimodal sequential expressions, from the annotation to the synthesis of the behavior.

Key words: virtual characters, emotional expressions, multimodality

1 Introduction

In this paper a novel approach to the generation of emotional displays of a virtual character is presented. The aim is to develop a model of multimodal emotional behaviors that is based on data from literature and on the annotation of a video-corpus. For this purpose a language was developed to describe the appearance in time of single signals as well as the relations between them.

Recent studies (e.g. [1–3]) show that several emotions are expressed by a set of different nonverbal behaviors which include different modalities: facial expressions, head and gaze movements, gestures, torso movements and posture. We call *multimodal sequential expressions of emotions* emotional displays that go beyond the description of facial expressions of emotions in their apex. They might be composed of nonverbal behaviors (called in this paper *signals*) displayed over different modalities or/and as a sequence of behaviors.

Several models of emotional expressions have been proposed to enrich virtual characters' behavior. A tool to modify manually the course of the animation of any single facial parameter was proposed in [4]. In Paleari and Lisetti [5] and Malatesta et al. [6] the emotional expressions are manually created from sequences predicted in Scherer's appraisal theory [7]. Both papers focus on the temporal relations between different facial actions predicted by the theory. In

Xueni Pan et al. [8] a motion graph is used to generate emotional displays from sequences of signals like facial expressions and head movements. The arcs of the graph correspond to the observed sequences of signals while nods are possible transitions between them. New animations can be generated by reordering the observed displays. Finally Lance and Marcella [9] model head and body movements in emotional displays focusing on the correlation between them.

In next sections we present a system that allows the generation of multimodal sequential expressions.

2 Multimodal sequential expressions language

In this section we present the representation scheme that encompasses the dynamics of emotional behaviors. The scheme is based in observational studies. We use a symbolic high level notation which gives us flexibility in the generation of possible behaviors. Our XML-based language defines multimodal sequential expressions in two steps: *behavior set* and *constraint set*. Single signals like a *Duchenne smile*, *shake* or *bow* are described in the repositories of the character's nonverbal behaviors. Each of them may belong to one or more *behavior sets*. Each emotional state has its own behavior set, which contains signals that might by used by the character to display that emotion. A number of regularities occur in expressions that concern the signal duration and the order of displaying (see e.g. [1, 2]). Consequently for each signal in a behavior set one may define the following five characteristics: *probability_start* and *probability_end* - probability of occurrence at the beginning (resp. towards the end) of an expression (a value in the interval [0..1]), *min_duration* and *max_duration* - minimum (resp. maximum) duration of the signal (in seconds), *repetitivity* - number of repetitions during an expression.

According to the observational studies (e.g. [2]) the signals occurrence in an emotional display is not accidental. The relations that occur between the signals of one behavior set are more precisely described in the *constraint sets*. This set introduces a set of constraints on the occurrence and duration (i.e. on the values for start_{s_i} and stop_{s_i}) of the signal s_i in relation to others signals. We introduced two types of constraints:

- *temporal constraints* define relations on the start time and end time of a signal using arithmetical relations: $<$, $>$ and $=$;
- *appearance constraints* describe more general relations between signals like inclusion or exclusion e.g. "signals s_i and s_j cannot co-occur" or "signal s_j cannot occur without signal s_i".

The constraints of both types are composed using the logical operators: *and, or, not*. The constraints take one or two arguments.

Three types of *temporal constraints* are used *morethan*, *lessthan*, and *equal*. These arithmetical relations may involve one or two signals: for example the observation: "signal s_i cannot start at the beginning of animation" will be expressed as following $start_{s_i} > 0$, while "signal s_i starts immediately after the signal s_j finishes" will be $start_{s_i} = stop_{s_j}$.

In addition, five types of *appearance constraints* were introduced:

- $exists(s_i)$ - is true if the s_i appears in the animation;
- $includes(s_i, s_j)$ - is true if s_i starts before the signal s_j and ends after the s_j ends;
- $excludes(s_i, s_j)$ - is true if s_i and s_j do not co-occur at the same time t_k i.e.: if $start_{s_i} < t_k < stop_{s_i}$ then $stop_{s_j} < t_k$ or $start_{s_j} > t_k$ and if $start_{s_j} < t_k < stop_{s_j}$ then $stop_{s_i} < t_k$ or $start_{s_i} > t_k$;
- $precedes(s_i, s_j)$ - is true if s_i ends before s_j starts;
- $rightincludes(s_i, s_j)$ is true if s_i starts before the signal s_j ends, but s_j ends before s_i ends.

During the computation of the animation constraints are instantiated with signals appearance times (i.e. $start_{s_i}$ and $stop_{s_i}$). By the convention the constraints that cannot be instantiated (i.e. one of the arguments does not appear in the animation) are ignored. An animation is consistent if there is no constraint that is not satisfied.

3 From annotation to behavior representation

In this section we present how the definition of behavior and constraint sets are created from the manual annotation. One coder annotated the modalities of the face, head, gaze and body movements. The facial changes have been annotated with FACS [10], while the head, gaze and body movements were described verbally. For practical reasons a *signal* is defined as a configuration of body actions that can occur at the same time in a particular modality. Thus one signal per modality is displayed at a time. Usually different body actions of one modality were defined as independent signals, e.g. *a hand touching the face* and *a hand hiding the mouth* gestures are two signals. The same body actions can be part of several signals, if they can occur in different configurations and with different co-occurrences, e.g. a smile is a signal, a smile with an open mouth is another one even if they have some AUs in common.

Figure 1 presents the FACS annotation of a segment of a panic fear expression. The following signals were individuated in this sample:

- *signal 1*: eyebrows very raised and drawn together, eyes extremely open, mouth extremely open,
- *signal 2*: inner eyebrows slightly raised, slightly drawn together, mouth slightly open with lowered mouth corners,
- *signal 3*: upper lid raised widening the eye, mouth open,
- *signal 4*: eyebrows drawn together, upper lid raised to widen the eye, lower lid raised,
- *signal 5*: outer eyebrows raised, eyebrows drawn together, mouth open with lowered corners.

In three of the four panic fear videos, extreme facial displays of fear with the very widely open eyes and mouth (signal 1) were followed by milder facial expressions

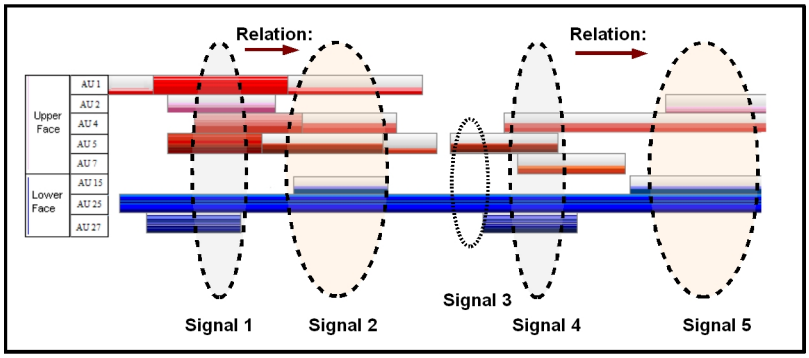

Fig. 1. Annotation of facial expressions of panic fear.

of fear like an open mouth with falling lip corners and the inner or outer part of the eyebrows raised (signal 2). As a consequence, a rule was written to state besides others that *signal 2* occurs *immediately* after *signal 1* (see Figure 1). This information might be described in the constraint set as a *temporal* constraint: the start time of *signal 2* is equal to the end time of *signal 1* (i.e. $stop_{signal1} = start_{signal2}$).

Then analyzing expressions of panic fear across different modalities the following signals were individuated: hands to the face (*signal 6*), shoulders up (*signal 7*), hand suspended in the air (*signal 8*), hand that hides the mouth than comes to rest on the chest (*signal 9*), both hands on the chest (*signal 10*), hand hides the mouth (*signal 11*), head (*signal 12*) and gaze turns to the side (*signal 13*). By looking at the panic fear videos one could argue that *signal 6* (i.e. a gesture with hands to the face) is a signal that cannot occur in panic fear if *signal 1* (i.e. facial expressions of eyebrows very raised and drawn together, eyes and mouth extremely opened) has not started before. It also stops before the end of *signal 1*. This information might be described in the constraint set by *appearance constraints* of the type *includes* and *exists*. When *signal 6* cannot appear without *signal 1* we obtain the constraint: *exists(signal 1) and includes(signal1,signal6)*.

4 Generation of multimodal sequential expressions

In our model, the behavior and constraint sets are used to generate multimodal sequential expressions of emotions. The input to the system is one emotional label e (e.g. *panic fear* or *embarrassment*) from a predefined set of emotional labels and its expected duration, t. Our system generates sequences of multimodal expressions, i.e. the animation A of a given duration t composed of a sequence of signals $s_{i(j)}$ on different modalities. It does so by choosing a coherent subset of signals from the behavior set BS_e as well as their timing $start_{s_i}$, $stop_{s_i}$.

4.1 Algorithm

Let A be the animation to be displayed by a virtual character. A can be seen as a set of triples $A = \{(s_i, start_{s_i}, stop_{s_i})\}, start_{s_i}, stop_{s_i} \in [0..t], start_{s_i} < stop_{s_i}$ where s_i is the name of the signal, $start_{s_i}$ is the start time of the signal s_i and $stop_{s_i}$ is its stop time. At the beginning A is empty. In the first step the algorithm chooses the behavior set $BS_e = \{s_k\}$, the constraint set $CS_e = \{c_m\}$ corresponding to the emotional state e, and the number n of uniform intervals, time stamps, for which the t is divided. Next, at each time step, t_j, (j=0..n-1, t_n=t), the system randomly chooses a signal-candidate s_c between the signals of the behavior set BS_e considering their probabilities of occurrence. For this purpose it manages a table of probabilities that contains, for each signal s_k, its current probability value $p_{k(tj)}$. Obviously, at the first time stamp, $t_0 = 0$, the values of this table are equal to the values of the variable *probability_start*, while at the last time stamp t_{n-1} the probabilities are equal to the *probability_end*. At each time stamp, t_j, the probabilities $p_{k(tj)}$ of each signal $s_k \in BS_e$ are updated. The candidate for a signal to be displayed s_c in a time stamp t_j is chosen using the values $p_{k(tj)}$. Next, the start time $start_c$ is chosen from the interval $[t_j, t_{j+1})$ and the consistence of CS_e with the partial animation $A(t_{j-1}) \cup (s_c, start_{s_c}, \emptyset)$ is checked. If all the constraints are satisfied the stop time $stop_c$ is randomly chosen between two values:

$$stop_{c1} = min_{s_c} + R * \frac{M}{2}, \qquad stop_{c2} = max_{s_c} - R * \frac{M}{2}, \qquad M = max_{s_c} - min_{s_c}$$

while R is a value from the interval [0..1], $\max_{s_c}$ is the maximum duration of s_c while $\min_{s_c}$ is the minimum duration of s_c. Otherwise, i.e. if there is a constraint that is not satisfied, another signal from BS_e is chosen as candidate. The consistency of the triple $(s_c, start_{s_c}, stop_{s_c})$ with the partial animation $A(t_{j-1})$ is checked again. If all the constraints are satisfied the signal s_c starting at $start_{s_c}$ and ending at $stop_{s_c}$ is added to A. The table of probabilities is updated and the algorithm chooses another signal, moves to the next time stamp, or finishes generating the animation.

In our approach we do not scale the timing of an observed sequence of behaviors to t. Rather the system chooses between the available signals of a behavior set in order to generate animations. The choice of our approach is motivated by research results showing that the duration of signals is related to their meaning. For example, spontaneous facial expressions of felt emotional states are usually not longer than four seconds, while the facial display of surprise is much shorter [11]. Similarly, gestures have also a minimum duration. Moreover the same gesture performed with different expressivity parameters (e.g. velocity) might convey different meanings.

The algorithm intentionally does not use any backtracking mechanism as it is implemented in near real-time applications that generate the animation rapidly. Thus in each computational step it adds a new signal that starts not earlier than the previous one. It allows the animation generation pipeline to be more efficient.

The algorithm is able to generate a number of animations that is consistent with the constraints. In this way we avoid the repetitiveness of the character's behavior and we obtain a variety of animations, each of which is consistent with the annotator's observation but go beyond a set of annotated cases.

4.2 Example

a b c d

Fig. 2. An example of a multimodal expression, based on the annotation of panic fear.

We used the Greta agent [12] to generate animations using our model. In Figure 2 an animation generated by our algorithm from the description of panic fear (see section 3). The following signals are displayed: 2a) *signal 1* - eyes and mouth extremely open, 2b) *signal 1* with *signal 6* - hand to the eyes, 2c) *signal 1* and *signal 10* i.e. hands on chest, 2d) *signal 2* accompanied by *signal 12* and *signal 13* - eyebrows slightly raised and drawn together, mouth open with lowered lip corners, head and gaze turns to the side.

5 Conclusions

In this paper a multimodal sequential expressions model for a virtual character was introduced. These expressions go beyond facial displays defined in their apex. For this purpose a language was proposed that allows formalizing the observational data and an algorithm that generates multimodal sequential expressions coherent with their descriptions. We have conducted a perceptual study [13] of eight expressions (anger, anxiety, cheerfulness, embarrassment, panic fear, pride, relief, tension) generated by our algorithm. The results show that multimodal sequential expressions enable the recognition of affective states, such as relief, that are not prototypical expressions of basic emotions. In the case of all emotions the recognition rate surpassed chance level [13].

Acknowledgement
Part of this research is supported by the EU FP6 Integrated Project CALLAS IP-CALLAS IST-034800.

References

1. Haidt, J., Keltner, D.: Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. Cognition and Emotion **13**(3) (1999) 225–266
2. Keltner, D.: Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology **68** (1995) 441–454
3. Wallbott, H.: Bodily expression of emotion. European Journal of Social Psychology **28** (1998) 879–896
4. Ruttkay, Z.: Constraint-based facial animation. International Journal of Constraints **6** (2001) 85–113
5. Paleari, M., Lisetti, C.: Psychologically grounded avatars expressions. In: First Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, Bremen, Germany (2006)
6. Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.D.: Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. Appl. Intell. **30**(1) (2009) 58–64
7. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking. In Scherer, K., Schorr, A., Johnstone, T., eds.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press (2001) 92–119
8. Pan, X., Gillies, M., Sezgin, T.M., Loscos, C.: Expressing complex mental states through facial expressions. In: Second International Conference on Affective Computing and Intelligent Interaction (ACII), Springer (2007) 745–746
9. Lance, B., Marsella, S.: Emotionally expressive head and body movements during gaze shifts. In: Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA), Springer (2007) 72–85
10. Ekman, P., Friesen, W.: Facial Action Coding System. Consulting Psychologists Press (1978)
11. Ekman, P., Friesen, W.: Unmasking the Face. A guide to recognizing emotions from facial clues. Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1975)
12. Bevacqua, E., Mancini, M., Niewiadomski, R., Pelachaud, C.: An expressive ECA showing complex emotions. In: Proceedings of the AISB Annual Convention, Newcastle, UK (2007) 208–216
13. Niewiadomski, R., Hyniewska, S., Pelachaud, C.: Evaluation of multimodal sequential expressions of emotions in ECA. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), Amsterdam, Holland, Springer (2009)