Radosław Niewiadomski[1], Sathish Pammi[1], Abhishek Sharma[1], Jennifer Hofmann[2], Tracey Platt[2], Richard Thomas Cruz[3], Bingqing Qu[1]

## Visual laughter synthesis: Initial approaches

*[1]Telecom ParisTech, Rue Dareau, 37-39, 75014 Paris, France*
*[2]Universität Zürich, Binzmuhlestrasse, 14/7, 8050 Zurich, Switzerland*
*[3]De la Salle University, Manila, Philippines*

Visual laughter synthesis is a challenging task that was only rarely explored and empirical investigations are scarce. For the purpose of building a virtual agent able to laugh naturally we exploit different animation techniques such as a procedural animation or based on motion capture and we apply them to visual laughter synthesis. At the moment we focus on three approaches: procedural animation based on manual annotation of facial behavior; motion capture driven animation and animation generated from automatic facial movements detection. For the purpose of this study we use the Greta agent (Niewiadomski et al., 2011) that can be driven by both high-level anatomically inspired facial behavior description based on the Facial Action Coding System (FACS; Ekman, et al., 1978) or low-level facial animation parameterization (FAPs) that is a part of MPEG-4 standard (Ostermann, 2002) for facial animation. We also use two video corpora: AVLC database (Urbain et al., 2011) containing mocap, video and audio data of 24 subjects showing spontaneous amusement laugh responses and Queen's University Belfast's dataset of same sex interaction dyads during the watching of funny stimuli. We present all approaches in detail.

**Manual annotation of action units.** FACS is a comprehensive anatomically based system for measuring all visually discernible facial movement. It describes all distinguishable facial activity on the basis of 44 unique Action Units (AUs), as well as several categories for head and eye positions/movements and miscellaneous actions. Using FACS and viewing digital-recorded facial behavior at frame rate and in slow motion, certified FACS coders are able to distinguish and code all visually discernible facial expressions. Utilizing this technique, a selection of twenty pre-recorded, laboratory stimulated, laughter events were coded. These codes were then used to model the facial behavior on the Greta agent which is able to display any configuration of AUs. For 3 virtual characters single AUs were defined and validated by certified FACS coders under the constraints of the technology. A Behavior Markup Language (BML) implemented in Greta permits the control of each AU of the agent (its duration and intensity) independently. The animation of any AU is linearly interpolated according to Attack-Decay-Sustain-Release model (Ekman, et al., 1978). Next, the symbolic intensity values are converted to low-level facial animation parameters (FAP) which, finally, are used to deform a mesh of the virtual model. We also developed a tool that automatically converts manual annotation files created with Noldus Observer XT, a commercial tool for manual videos annotation, to BML. Consequently any file containing manual annotation of AUs can be easily displayed with the Greta agent.

**Animation from automatic facial movements detection.** The Greta agent uses facial animation parameters (FAPs) to realize low-level facial behavior. FAPs represent movements of MPEG-4 facial points compared to the 'neutral' face. In order to estimate FAPs of natural facial expressions, we made use of an open-source face-tracking tool – FaceTracker (Saragih et al., 2010) – to track facial landmark localizations. It uses a Constrained Local Model (CLM) fitting approach that includes a Regularized Landmark Mean-Shift (RLMS) optimization strategy. It can detect 66 facial landmark coordinates within real-time latency depending on the system's configuration.

Facial geometry differs from one human to another one. Therefore, it is difficult to estimate FAPs without neutral face calibration. To compute FAPS from facial landmarks, a neutral face model is created with the help of 50 neutral faces of different persons. With the help of this model, FAPs are estimated as the distance between facial landmarks and neutral face landmarks. In case of user-specific FAP estimation in a real-time scenario, the neutral face is estimated from a few seconds of video by explicitly requesting the user to hold the face still. However, the better estimation of FAPs requires manual intervention for tweaking weights to map landmarks and FAPs, which is a downside of this methodology.

The landmark coordinates produced by the FaceTracker are observed as noisy due to the discontinuities and outliers in each facial point localization. Especially, the realized behavior is unnatural on a virtual model when we re-target the observed behavior onto the Greta agent. To smooth the face-tracking parameters, a temporal regression strategy has been applied on individual landmarks by fitting 3rd order polynomials using a sliding window, where the sliding window size and its shifting rate are 0.67 seconds and 0.33 seconds respectively.

**Animation from motion capture data.** AVLC corpus (Urbain et al., 2011) contains motion capture data of laugh episodes that have to be retargeted to the virtual model. The main problem in these kinds of approaches consists in finding appropriate mappings for each participant's face geometry and different virtual models. Many existing

solutions are typically linear (e.g., methods based on blend shape mapping) and do not take into account dynamical aspects of the facial motion itself. Recently, Matthew Zeiler and colleagues (2011) proposed to apply variants of Temporal Restricted Boltzmann Machines (TRBM) to the facial retargeting problem. TRBM are a family of models that permit tractable inference and allows complicated structures to be extracted from time series data. These models can encode a complex nonlinear mapping from the motion of one individual to another, which captures facial geometry and dynamics of both source and target. In the original application (Zeiler et al., 2011) these models were trained on a dataset of facial motion capture data of two subjects, asked to perform a set of isolated facial movements based on FACS. The first subject had 313 markers (939 dimensions per frame) and the second subject had 332 markers (996 dimensions per frame). Interestingly there was no correspondence between marker sets. They were able to retarget the motion with a RMS error of 2 %. However, they only evaluated their results on slow facial movements.

We use TRBM models for our project, which involves retargeting from an individual to a virtual character. In our case, we take the input as the AVLC mocap data and output the corresponding facial animation parameters (FAP) values. This task has two interesting aspects. First, the model performance was previously evaluated only on retargeting an isolated slow expression whereas our case involves transitions from laughter to some other expression (smile or neutral) as well as very fast movements. Second, we use less markers compared to the original application. Our mocap data had only 27 markers on face, which is very sparse.

So far we used the AVLC data of one participant. As a training set we used two sequences, one of 250 frames and another one of 150 frames. Target data (i.e., facial animation parameters) for this training set was generated using manual retargeting procedures explained in Urbain et al. (2011). Both the input and output data vectors were reduced to 32 dimensions by retaining only their first 32 principal components. Since this model typically learns much better on scaled data (around [-1,1]), the data was then normalized to have zero mean and scaled by the average standard deviation of all the elements in the training set. Having trained the model, we used it to generate facial animation parameters values for 2 minutes long mocap data (2500 frames coming from the same participant). The first results are promising but more variability in the training set is needed to retarget more precisely different type of movements.

**Conclusion.** These three approaches offer different degrees of flexibility and control over the expression, different levels of realism and precision of the movements. We expect, for instance, that the mocap-based animation should be richer in movements and consequently it may be perceived as more realistic. Also using mocap data should permit to maintain the temporal and dynamic characteristics of the original laugh. On the other hand animation generated with this method is difficult to control manually (e.g., its duration, intensity, communicative function). Moreover the mocap procedure is invasive, recourse- and time consuming. On the other hand, describing animation by action units allows one to control precisely an animation and its meaning (e.g., by adding or removing AU6, a marker of the Duchenne smile) but has all the weaknesses of procedural approaches to facial animation. The animation is poor in details and the dynamics of the movements is very simplistic. Finally, a solution based on the automatic facial action detection combines advantages of both solutions: it should be sufficiently rich in the details (it depends highly on the quality of the face tracker applied). At the same time one can manually control and edit the final animation by adding or removing some facial actions. Still it requires that recordings be taken in controlled conditions (e.g., good lighting).

Future works will consist of a set of perceptive studies that we want to develop in order to check the quality of the animations and compare our 3 methods. For this purpose we use just one set of laugh episodes and will generate animations with these 3 different approaches. The factors considered in the evaluation will be believability and naturalness of the animations.

**Bibliography**


Ekman, P., Friesen, W.V., & Hager, J. C. (1978). *Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto: Consulting Psychologists Press.

Niewiadomski, R., Bevacqua, E., Quoc Anh Le, Obaid, M., Looser, J., & Pelachaud, C. (2011). Cross-media agent platform. *Web3D ACM Conference*, Paris, France (pp. 11-19).

Ostermann, J. (2002). Face animation in MPEG-4. In I. Pandzic and R. Forchheimer (eds.), *MPEG-4 Facial Animation - The Standard Implementation and Applications* (pp. 17–55), England: Wiley.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision 91* 200 – 215.

Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., & Wagner, J. (2010). AVLaughterCycle. Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation. *Journal of Multimodal User Interfaces*, *4*, 47-58.

Zeiler, M.D., Taylor, G.W., Sigal, L., Matthews, I., & Fergus, R. (2011). Facial Expression Transfer with Input Output Temporal Restricted Boltzmann Machines. *Neural Information Processing Systems Conference NIPS 2011*, Granada, Spain. (pp. 1629-1637).