Virtually Cloning Real Human with Motion Style

Manoj kumar Rajagopal[1], Patrick Horain[1], Catherine Pelachaud[2]

[1] Institut Telecom, Telecom SudParis, 9 rue Charles Fourier, 91000 Evry, France
[2] CNRS Telecom ParisTech, 37-39 rue Dareau, 75014 Paris, France

Abstract. Our goal is to capture style from real human motion so it can be rendered with a virtual agent that represents this human user. We used expressivity parameters to describe motion style. As a first contribution, we propose an approach to estimate a subset of expressivity parameters defined in the literature (namely spatial extent and temporal extent) from captured motion trajectories. Second, we capture the expressivity of real users and then output it to the Greta engine that animates a virtual agent representing the user. We experimentally demonstrate that expressivity can be another clue for identifiable virtual clones of real humans.

Keywords: Motion Style, Expressivity, Virtual Clone, Embodied Conversational Agent (ECA), Greta

1 Body Motion Style and Expressivity

Non-verbal behavior, including body language, is an important communication channel. It conveys user's mental and affective states. Non-verbal behavior reflects the cognitive and emotional states[16]. In face-to-face communication, body language can convey up to 55 % of the information on feelings and intentions[21]. Hand gesture is part of the cognitive process in communication[2].
Gallaher[13] defined "style", or expressive movement, in reference to the way behavior is performed. She found style has four dimensions namely expressiveness, animation, expansiveness and coordination. Human motion style and human expressivity are interlinked to each other[8].

1.1 Motion style

To our knowledge, Tenanbaum *et al.* [27] were the first who separated motion style and content. They used a bilinear model to identify the font from hand written characters.
Elgammal *et al.* [11] proposed to separate style and content through non-linear dimensionality reduction.
Hsu *et al.* [17] proposed to transform the style of human motion while preserving its original content, using a linear time invariant model.

Wang *et al.* [31] used multi-factor Gaussian process models for style and content separation in human motion. They used a non-linear basis function to generalize multi linear models and a Gaussian process latent variable model to marginalize over the weights.
Human identification based on gait analysis has been widely studied [18] [6] [10]. Gait, which can be regarded as a walking style, does convey information on the identity of a person.
Noot and Ruttkay [23] developed a markup language called GESTYLE to specify the gesturing style of an Embodied Conversational Agent (ECA).They defined style for virtual human in terms of style dictionaries. Style dictionary contains typical for an individual, professional or cultural group or people of certain age, sex or personality (e.g. to accommodate meanings depending on his culture, specific meanings belonging to his profession, personal habits, etc.).
Since all these approaches work frame-by-frame, they consider neither motion speed nor acceleration.

1.2 Expressivity

Some studies [22] [3] decode the emotional states to movement qualities and investigate from expressive body movements. Drosopoulos *et al.* [9] analyzed and classified expressive gesture from image processing (body movements and facial expressions) for multimodal emotion recognition.Camurri *et al.* [5] [4] classified expressive gesture in human body movement during music and dance performances. Wallbott *et al.* [30] classified the bodily behavior using five criterions: slow or fast, small or expensive, weak or energetic, small or large movement activity and unpleasant or pleasant. Behavior expressivity has been correlated to energy in communication, to the relation with characteristics of gestures and/or personality/emotion. For Wallbott [29], it is related to the mental state (e.g. emotion). Behaviors encode not only content information (information communicated through gesture shape) but also expressive information (the how it is communicated, the manner of execution).
Following Wallbott's approach, Hartmann *et al.* [15] focused on expressivity (behavior expressivity) based on surface realization of movement from hand motion. They have defined a set of parameters that describe expressivity for a virtual agent based on the wrists 3D trajectories. They integrated expressivity control in the animation engine of the Greta ECA [24], a three-dimensional virtual agent that is able to communicate with real humans using a rich palette of verbal and nonverbal behaviors.
Expressivity is also conveyed by face expressions. Vasilescu and Terzopoulos [28] proposed a multi-linear analysis of tensor faces based on a tensor extension of the conventional matrix Singular Value Decomposition (SVD). Mancini *et al.* [19] map acoustic cues and emotion to expressivity parameters they use to control the expressive virtual head of the Greta ECA. Pelachaud and Poggi [25] aim at combining the Greta facial expressions in a complex and subtle way, just like humans do, by assessing and managing the multimodal communicative behavior

of a person when different communicative functions have to be displayed at the same time.

1.3 Our Approach

We aim at animating a virtual agent representing a real human with the expressivity learnt from that person. In this work, we propose an approach to estimate two of the Hartmann's *et al.* [15] expressivity parameters from 3D motion data. We validate this approach against synthesis animations with controlled expressivity generated using Greta. Then we estimate expressivity from motion captured from real users, and we feed the user's expressivity to the extended Greta ECA animation engine. Finally, we discuss a possible application for a user virtual representative that would clone gesturing expressivity, in complement to visual appearance and voice, turning expressivity into another key clue for user identity in virtual inhabited worlds.

1.4 Organizaion of the Paper

In this paper, section 2 defines Hartmann *et al.* [15] spatial and temporal extent parameters which is input for Greta ECA. Also in this section we have given the inside view of expressivity parameters. In section 3, we explain our method of determining the spatial and temporal parameters from real human motion. In section 4, we validate our spatial and temporal extent estimation against synthesized motion. In section 5, we explain our experiments conducted with real human data. In the process of experiments we animated virtual agent with estimated spatial and temporal extent as input. Also in this section we have given user reviews for animated virtual agent. In section 6 we have given the conclusion and future work.

2 Expressivity Parameters

Hartmann *et al.* [15] animated the Greta ECA [24] with a set of six expressivity features: overall activation, spatial extent, temporal extent, fluidity, power and repetitiveness. Each of the parameters is float-valued and defined over the interval [-1, 1], where zero point corresponds to the action without expressivity control. We briefly introduce those two parameters we are using, namely spatial extent and temporal extent.

2.1 Spatial Extent

Spatial extent describes the space used by a person for making gestures. Hartmann *et al.* [15] define spatial extent as a parameter controlling the centers of the McNeill's [20] sectors where hand movement occurs. Spatial extent value +1.0 (resp. 1.0) means the wrist moves further (resp. closer) to the coordinate origin, which is set at the sacroiliac vertebra.

They replace the wrist positions $p = (p_x, p_y, p_z)^T$ in the neutral spatial extent trajectory with positions $p' = (p'_x, p'_y, p'_z)^T$:

$$\begin{bmatrix} p'_x = (1 + SPC \cdot Agent_x)p_x \\ p'_y = (1 + SPC \cdot Agent_y)p_y \\ p'_z = (1 + SPC \cdot Agent_z)p_z \end{bmatrix} \quad (1)$$

where $\cdot_x, \cdot_y and \cdot_z$ refer to the lateral, vertical and frontal directions, Agent are scaling factors in the respective directions used in the Greta animation engine(resp. 1.3, 0.6 and 0.25 resp. in the x, y and z directions and positive SPC, and resp. 0.7, 0.25 and 0.25 resp. in the x, y and z directions and negative SPC), and SPC is the spatial extent in the range [1.0, +1.0].

2.2 Temporal Extent

Temporal extent (TMP) describes how fast a gesture is performed. According to Hartmann *et al.* [15] time for performing a gesture is segmented into three intervals, namely preparation, stroke and retraction. A stroke is that part of an expressive gesture that conveys meaning. It is preceded by a stroke preparation period when the wrist moves from the initial position, and followed by retraction where the wrist returns to the rest position after completing the stroke.
Temporal extent is related to the stroke time interval. Hartmann *et al.* [15] derive the time taken for each segment from a simplification of Fitt's law [12] :

$$T = a + b\log_2(\left\|x_{(n)} - x_{(n+1)}\right\| + 1) \quad (2)$$

where T is the duration of the stroke, a is a time offset, b is a velocity coefficient, x_n is the wrist positions at stroke start, x_{n+1} is the wrist positions at the stroke end. The velocity coefficient is defined by Hartmann *et al.* [15] as:

$$b = (1 + 0.2 \cdot TMP)/10 \quad (3)$$

where TMP is the temporal extent where 1.0 TMP corresponds to lower speed and +1.0 TMP to higher speeds.

3 Estimating Expressivity from a Motion Trajectory

We estimate the spatial and temporal extents from 3D motion trajectories described as sequences of poses p for the two wrists. We use the Greta animation engine to generate example motions with controlled expressivity, from which we tune parameters for estimation.

3.1 Estimating Spatial Extent

The spatial extent defines a scale of coordinates to the wrists positions $p'_.$ with respect to an origin. Following the convention of the Greta animation engine, we set that origin at the sacroiliac vertebra, which stands approximately between

the rest positions of the two wrists.
Taking the logarithm, equation (1) turns into an additive expression, where the term SPC is to be estimated:

$$Log(p'.) = Log(1 + SPC \cdot Agent.) + Log(p.) \tag{4}$$

Taking the mean over input samples, we get:

$$Log(1 + SPC \cdot Agent.) = \overline{Log(p'.)} - \overline{Log(p.)}$$

where $\overline{Log(p.)}$ can be learnt from a set of motion trajectories generated with null spatial extent. Finally, we estimate the spatial extent SPC of our input trajectory as the mean of the directional spatial extents:

$$SPC. = \frac{e^{\left[\overline{Log(p'.)} - \overline{Log(p.)}\right]} - 1}{Agent.} \tag{5}$$

where . stands for either x, y or z and is $[\overline{---}]$ the mean of $[---]$ and $\overline{Log(p.)}$ is learnt from a trajectory generated with null spatial extent. Finally, we estimate the spatial extent SPC of our input trajectory as the mean of the directional spatial extents:

$$SPC = \frac{1}{3}(SPC_x + SPC_y + SPC_z) \tag{6}$$

3.2 Estimating Temporal Extent

As already mentioned, the higher the temporal extent TMP, the higher the speed is. From this monotonous relation, we derive a heuristic to estimate TMP from the observed instant speeds along the trajectory, i.e. the distance between poses at successive time intervals.
In the work by Hartman's *et al.* [15] expressivity parameters affect synthesized gestures during stroke time only. Rather than segmenting strokes, which is a difficult problem [26] we estimate expressivity parameter from whole motion trajectories. From synthesis trajectories, we find that high TMP values give high speed only during the strokes, while the preparation and retraction exhibit low speeds. Therefore, TMP can be estimated by considering only some quantile of the highest speeds.
Based on this, we sort the instant speeds of example trajectories in descending order, and we set the quantile limit at the sample with highest correlation with the TMP values that were originally input to the animation engine for generating the example trajectories. From the learnt trajectories we considered, we found that the mean speed up to 7 % upper quantile determines the temporal extent.The correlation for mean speed up to 7 % upper quantile is 93%. After rejecting 2 % upper quantile, the correlation improves to 97 %. This improvement is because there may be a chance for discontinuity while combining gestures to form a motion, this decreases the correlation value. So we reject initial 2 % of upper quantile for the mean calculation and mean speed with the remaining 5%

of upper quantile we achieve best mapping to temporal extent in the range [1.0, +1.0] through linear regression (Fig.1). Using the obtained linear regression from learnt trajectories, we estimated temporal extent for other motion trajectories. The experimental results show that our estimation method is robust.

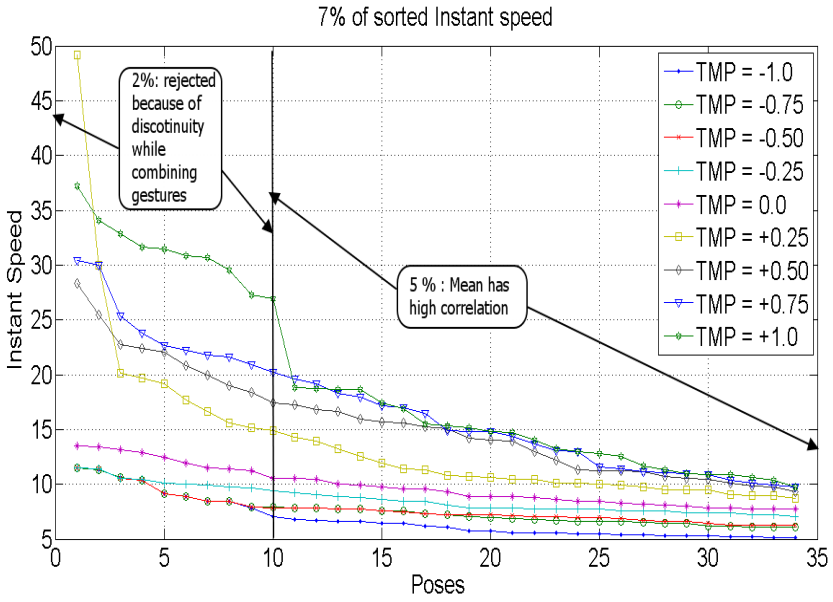

Fig. 1. Highest instant speeds from example synthesized trajectories.

4 Validating Expressivity Estimation

We evaluate the above approach for estimating the previous two expressivity parameters against a corpus of ten communicative gestures with controlled expressivity that are generated using the Greta animation engine [24]. Each of the ten gestures has been rendered with nine different values of spatial extent (-1.0, -0.75, -0.5, -0.25, 0.0, +0.25, +0.5, +0.75 and +1.0). We calculate the wrist positions of the synthesized motions through forward kinematics [7]. From the wrist positions, spatial extent is calculated as described in paragraph 3.1. The estimated spatial extent and actual spatial values are plotted as shown in Fig.2.

From Fig.2 we see that estimated spatial extent not exactly lying with actual spatial extent due to error in the estimation process. The error in the estimation process is calculated in terms of absolute mean error. The absolute mean error for spatial extent estimation is 0.14. In the range [-1.0,+1.0] there are 20 samples of size 0.1. It is understood that our spatial extent estimation process produces the result with 7 % of error.

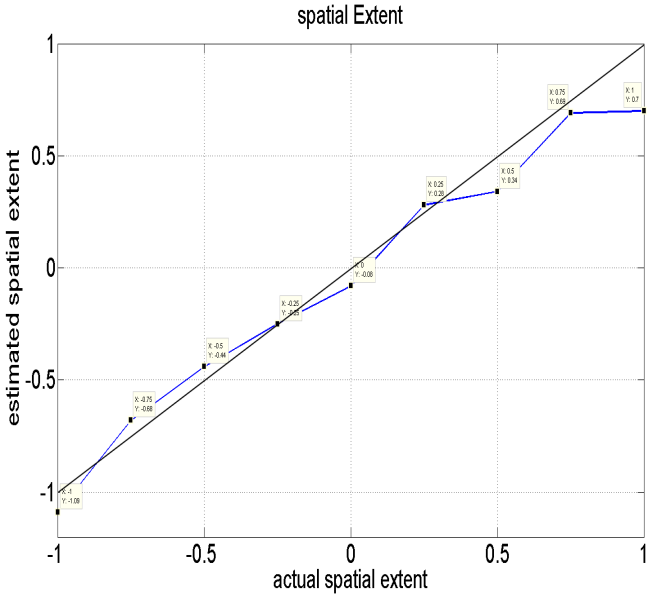

Fig. 2. Estimated spatial extent *vs.* input spatial extent for synthesized gestures

The estimation of TMP is also validated against synthesized motion with varying controlled expressivity. Instant speeds of the wrists are calculated using forward kinematics and are sorted in increasing order. We map instant speed to temporal extent values using the linear regression found at paragraph 3.2. The estimated and actual temporal extent values are plotted in Fig.3.
Similar to the spatial extent, the error measure for TMP is also done by absolute mean error. Absolute mean error for TMP estimation is 0.15. (i.e.,) our TMP estimation method causes 7.5% error in determining the TMP from 3D motion data.

5 Experiments

We are interested in reproducing real user's expressivity in an ECA animation. Starting from input 3D motion data, we estimate expressivity parameters that we feed to the Greta ECA and then compare the synthesized animation against user's motion.

5.1 Estimating Expressivity

Our input is 3D human motion data that can be captured e.g. using a consumer system such as newly available Kinect[1]. Alternatively, we used motion data captured by computer vision from two video lectures (hereafter named V1 and

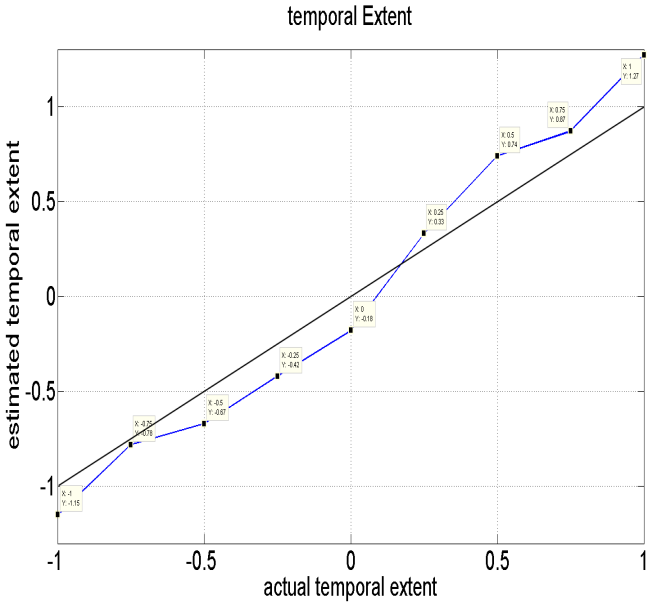

Fig. 3. Estimated temporal extent *vs.* input temporal extent for synthesized gestures

V2 , see Fig.4 and Fig.5) using software developed by Gómez Jáuregui *et al.* [14] that works by registering a 3D articulated model of the human body onto videos and outputs the upper body joints angles. 3D wrist positions are obtained from upper body joint angles through forward kinematics [7].
The spatial extent is estimated as explained in section 3.1 for both videos of V1 and V2. It yields the spatial extent for V1 as +0.8 and the spatial extent for V2 as +0.6. By seeing the V1 and V2 snap shots, we can say V1 has more spatial extent than V2. Experiment results also confirm this.

Fig. 4. Set of poses of real human motion in video sequence V1

Fig. 5. Set of poses of real human motion in video sequence V2

The estimated TMP is -1.0 for V1 and -0.7 for V2. This shows the user in V2 has more temporal extent than user in V1. The estimated TMP cannot be compared from the snap shots of V1 and V2. The TMP estimation for V1 and V2 is analyzed in section 5.2.

5.2 Animating Virtual Agent

We animate a virtual agent with new gesture based on spatial and temporal extent value which we estimated based on section 3.1 and 3.2. From this estimation we synthesize new gesture with Greta having the spatial extent of V1 and of V2 (resp. Fig.6 and Fig.7).

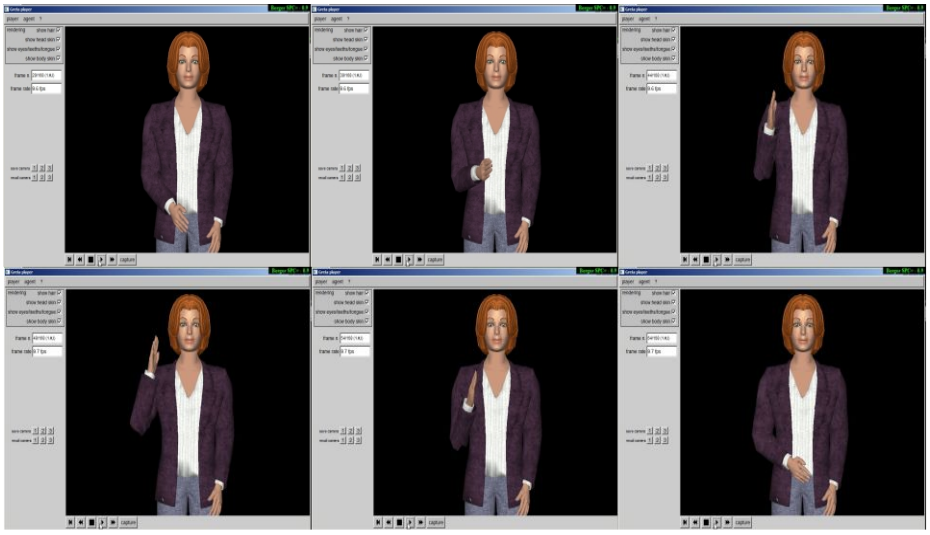

Fig. 6. Animated gesture using Greta for the spatial extent of V1.

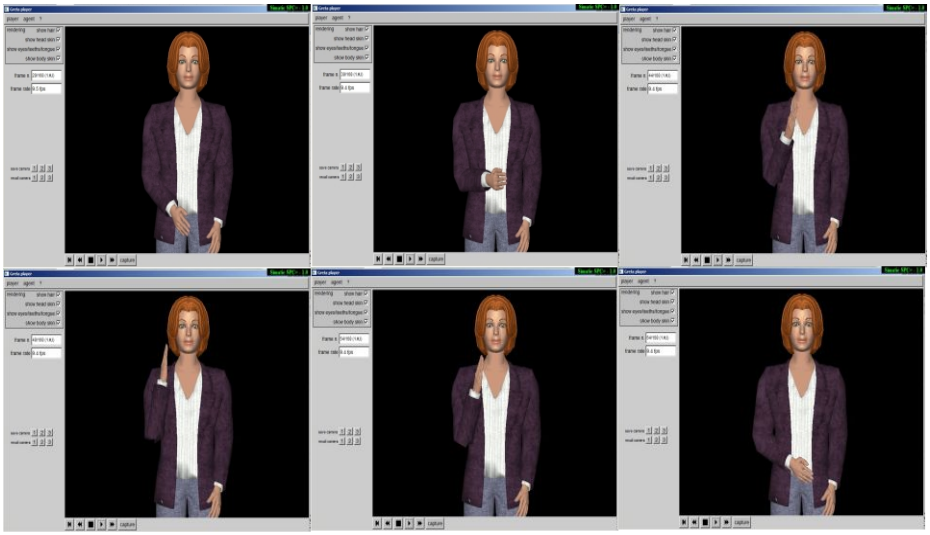

Fig. 7. Animated gesture using Greta for the spatial extent of V2.

Using estimated temporal extent for the users in videos V1 and V2, motion trajectories are synthesized. Variation of temporal extent for synthesized trajectories is explained by Y position variation of wrist with respect to chest. From the synthesized motion, the variation of Y position of the wrist with respect to chest is shown in Fig.8. In Fig.8, the higher TMP attains peak earlier and finishes the gesture earlier than lower TMP motion. As we know the gesture which is done with higher speed will end earlier, this is reflected in the Fig.8. Also this shows that user in V2 has higher temporal extent than user in V1 comparatively.

5.3 User Reviews on Animation

We validated our results with user reviews. We have chosen some of our friends who do not work on expressivity parameters or animation field as our users to validate our results. We believe their reviews will be from the lay man view on animated virtual agent. So far we have seventeen users reviewed our results. We have conducted three tests. In the first test, the animated video contains only spatial extent parameter variation in the generated motion. In the second test, the synthesized motion has variation based on temporal extent parameter. In the third test, we generated synthesized video having variation of both spatial and temporal extent variation. Users are asked to identify whether they can recognize the synthesized motion similar to real human motion. First test result reveals that out of seventeen users, thirteen users correctly identified the synthesized motion with real human motion. In case of second test, fourteen users rightly recognized the animated video to the real human video. In the third test, fourteen

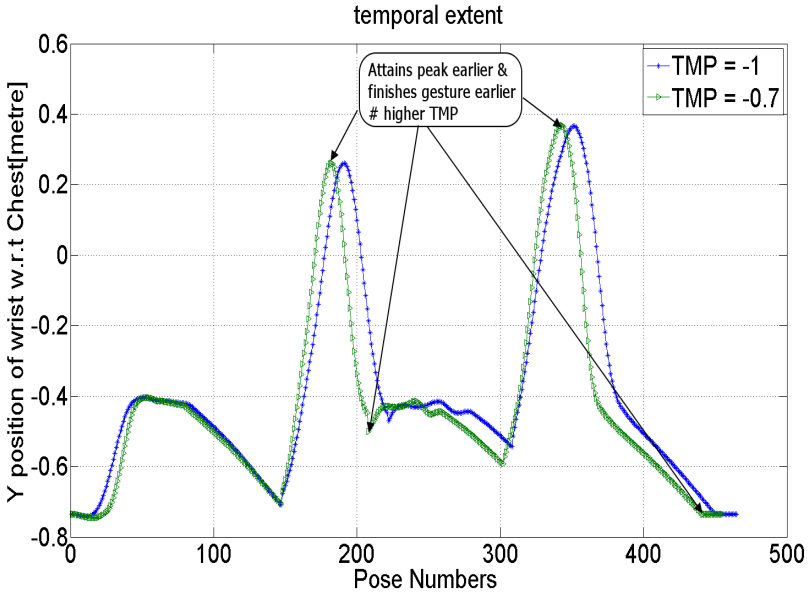

Fig. 8. Using temporal extent from V1 and V2 a motion is animated using Greta.

users could recognize the synthesized motion matching with real human motion. These results show that, our result has more than 75 % to 83 % consistency among defining the style of a human.

6 Conclusion and Future Work

By estimating the spatial and temporal extents from motion by a real user, we can then animate the virtual agent to render the expressivity captured from a real user. This animation can be played virtually when the user is not available to control the avatar. Rendering the expressivity parameters allows generating personalized animations, so that the viewer can have the feeling of interacting with an expressive virtual human. In the future, more expressivity parameters such as power and fluidity can be considered for fine control of the virtual character.

References

1. Kinect (2010), http://www.xbox.com/en-GB/kinect
2. Bavelas, J.B., Chovil, N.: Visible acts of meaning:an integrated message model of language in face to face dialogue. vol. 19, pp. 163–494 (2000)
3. Boone, R.T., Cunningham, J.G.: Children's decoding of emotion in expressive body movement: the development of cue attunement. vol. 34, pp. 1007–1016 (1998)

4. Camurri, A., Castellano, G., Ricchetti, M., Volpe, G.: Subject interfaces: measuring bodily activation during an emotional experience of music. vol. 3881, pp. 268–279 (2006)
5. Camurri, A., Lagerlf, I., Volpe, G.: Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. vol. 59, pp. 213–225 (2003)
6. Chellappa, R., Roy-Chowdhury, A.K., Kale, A.: Human identification using gait and face. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–2. Minneapolis, MN, USA (2007)
7. Craig, J.J.: Introduction to robotics : mechanics and control. Prentice hall, 3rd edition edn. (1986)
8. Davis, J.W., Gao, H.: An expressive three-mode principal components model of human action style. vol. 21, pp. 1001–1016 (2003)
9. Drosopoulos, A., Mpalomenos, T., Ioannou, S., Karpouzis, K., Kollias, S.: Emotionally-rich man-machine interaction based on gesture analysis. vol. 4, pp. 1372–1376 (2003)
10. Ekinci, M.: A new approach for human identification using gait recognition. In: ICCSA. pp. 1216–1225. Glasgow , UK (2006)
11. Elgammal, A., Lee, C.S.: Separating style and content on a nonlinear manifold. In: Computer Vision and Pattern Recognition (CVPR). pp. 478–485 (2004)
12. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude. In: Journal of Experimental Psychology. vol. 47, pp. 381–391 (June 1954)
13. Gallaher, P.E.: Individual differences in nonverbal behavior: Dimensions of style. vol. 63, pp. 133–145 (1992)
14. Gómez Jáuregui, D.A., Horain, P., Rajagopal, M.K., Karri, S.S.K.: Real-time particle filtering with heuristics for 3d motion capture by monocular vision. In: Multimedia Signal Processing. pp. 139–144. Saint-Malo, France (2010)
15. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture Workshop,LNAI. Springer (2005)
16. Hassin, Ran R; James, U.S.B.J.A. (ed.): The New Unconcious. Oxford University Press (2005)
17. Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. vol. 24, pp. 1082–1089 (2005)
18. Lee, L., Grimson, E.: Gait analysis for recognition and classification. In: Fifth IEEE International Conference on Automatic Face Gesture Recognition. pp. 734–742. Washington,USA (2002)
19. Mancini, M., Bresin, R., Pelachaud, C.: A virtual-agent head driven by musical performance. In: IEEE Transactions on Audio, Speech and Language Processing. vol. 15, pp. 1883–1841 (2007)
20. McNeill, D.: Hand and Mind what gestures reveal about thought. The university press of chicago press, Chicago,USA (1992)
21. Mehrabian, A., Wiener, M.: Decoding of inconsistent communications. vol. 6, pp. 109–114 (1967)
22. Meijer, M.: The contribution of general features of body movement to the attribution of emotions. vol. 13, pp. 247–268 (1989)
23. Noot, H., Ruttkay, Z.: The gestyle language. In: International Journal of Human-Computer Studies - Special issue: Subtle expressivity for characters and robots. vol. 62 (2005)

24. Pelachaud, C.: Greta, http://perso.telecom-paristech.fr/ pelachau/Greta/
25. Pelachaud, C., Poggi, I.: Subtleties of facial expressions in embodied agents. In: The Journal of Visualization and Computer Animation. vol. 13, pp. 301–312 (2002)
26. Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H.: Gesture, speech, and gaze cues for discourse segmentation. In: Computer Vision and Pattern Recoginition(CVPR). vol. 2, pp. 247–254 (2000)
27. Tenanbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. vol. 12, pp. 1247–1283 (2000)
28. Vasilescu, M.O., Terzopoulos, D.: Multilinear analysis of image ensembles: Tensorfaces. In: ECCV. pp. 447–460 (2002)
29. Wallbott, H.G.: Bodily expression of emotion. In: European Journal of Social Psychology. vol. 28, pp. 879–896 (1998)
30. Wallbott, H.G., Scherer, K.R.: Cues and channels in emotion recognition. In: Journal of Personality and Social Psychology. vol. 51, pp. 690–699. American Psychological Association (1986)
31. Wang, J.M., Fleet, D.J., Hertzmann, A.: Multifactor gaussian process models for style-content seperation. In: ICML. pp. 975–982. Corvallis, OR , USA. (2007)