Walk the Talk: Coordinating Gesture with Locomotion for Conversational Characters

Yingying Wang[1*] Kerstin Ruhland[2†] Michael Neff[1‡]
Carol O'Sullivan[2,3§]
[1]University of California, Davis
[2]Trinity College Dublin, Ireland
[3]Disney Research Los Angeles

Abstract

Communicative behaviors are a very important aspect of human behavior, and deserve special attention when simulating groups and crowds of virtual pedestrians. Previous approaches have tended to focus on generating believable gestures for individual characters and talker-listener behaviors for static groups. In this paper, we consider the problem of creating rich and varied conversational behaviors for data-driven animation of walking and jogging characters. We captured ground truth data of participants conversing in pairs while walking and jogging. Our stylized splicing method takes as input a motion captured standing gesture performance and a set of looped full body locomotion clips. Guided by the ground truth metrics, we perform stylized splicing and synchronization of gesture with locomotion to produce natural conversations of characters in motion.

Keywords: animation, motion capture, gesture synthesis

*yiwang@ucdavis.edu
†ruhlandk@scss.tcd.ie
‡mpneff@ucdavis.edu
§carol.osullivan@scss.tcd.ie

1 Introduction

While it may be a challenge for some to walk and chew gum at the same time, people frequently and effortlessly talk while they walk. This important behavior is missing, however, from most virtual character systems. Crowd simulations generally lack communicative behavior and miss natural eye, head and gesture movements of people walking. Systems focused on gesture and nonverbal communication are targeted almost exclusively at standing or sitting characters.

In general, motion capture databases consist of motions specific to a particular domain, e.g., locomotion, conversation, fighting. For combinations of motions, such as walking and talking, it would be impractical to try to capture every possible combination of both types of motion. Simulating conversational behavior for virtual characters while in locomotion is not as trivial as simply compositing separate gesture and locomotion behaviors. Gestures must be adapted to fit the natural arm swings and cadence of walking or jogging behavior. While in locomotion, attention patterns differ compared to standing as people have to pay attention to their walking path, to their conversational partner and towards points of interest.

There is a general dearth of previous research on how people adjust their gesturing behavior

while they walk or jog. We therefore began by conducting a study to obtain ground truth data on how people communicate during locomotion. Multiple subjects were video recorded standing, walking and jogging while engaged in different types of discussions and debates. According to our empirical observations, people tend to execute fewer and smaller gestures and gaze shifts toward their conversational partner while walking or jogging than they do while standing.

We take a data-driven approach to create natural animations of talking while walking or jogging. We take as input two sources of data: (i) a pre-existing locomotion database that contains various walking and jogging motions of multiple actors (without gesture); and (ii) gesticulation data from three-way standing conversations.

With our *stylized splicing* method, locomotion animation is automatically spliced with motion from the gesture database, which is intelligently adapted to match the style of the locomotion actor. For example, we temporally realign gesture emphasis to the locomotion tempo, and synthesize the typical bounce of arms seen in jogging.

Furthermore, we use an addresser-addressee relationship (AAR) to describe orientation behavior for characters engaged in a conversation with each other. In this way, we can convert standing group conversations into walking or jogging ones, by adding attention behavior and repairing the conversational partner relationship. The basis for the orientation behavior, and other conversational parameters, is derived from our annotated ground truth video. By transferring and adapting gesticulation data, our system is capable of creating conversational behaviors for individual characters and groups in locomotion.

2 Related Work

Generating conversational behaviors such as gesture, facial expressions and gaze has been a very active area of research. Utterance planning [1], prosody [2, 3], probabilistic modeling from input text [4] or real human performance [5] and rule-based systems [6] have all been used. Head movement and eye gaze of a virtual conversational partner may be used to communicate information about their internal states, attitudes, attentions and intentions [7] or to actively influence the conversation [8]. Ennis et al. [9] found that synchrony of the body motions of the conversing partners in a standing group was very important. However, the combination of conversational behaviors (e.g. gestures and gaze) for groups while walking, jogging and talking has not been explored.

Motion graphs and motion blending techniques have been proposed to reuse and combine existing motions into new motion sequences [10, 11, 12], where potential transition points in motion sequences are chosen based on a posture similarity metric and used to construct a graph structure. New sequences can be produced by stitching together the motion segments from a graph walk. Fernández-Baena et al. [13] construct a Gesture Motion Graph (GMG) from a labeled gesture database and select the graph walk that best matches the accompanying prosodic accent and the gesture timing slot. Stone et al. [5] build a linguistic network based on a character's utterance and choose optimized edges by penalizing the match of utterance and gesture, the connection of neighboring utterances and adjacent gestures. However, these approaches all treat the character state vector as a monolithic whole, taking all the Degrees of Freedom (DOFs) of data from a single clip at a time.

When splicing motions together, naive DOF replacement can produce unrealistic results as it ignores the physical and stylistic correlations between body parts [14, 15]. Mousas et al. [16] overcome the synchronization problem by using velocity based temporal alignment. Partial-body motion graphs can also be generated to splice and synchronize arm or hand motions with full-body

clips [17, 18, 19]. For example, Majkowska et al. [20] integrate separately captured hand motions into full body animation and find the corresponding splicing points using a two-pass dynamic time warping (DTW) algorithm. Our method not only adapts gesture performance spatially to the styles of the locomotion arm swing, it also temporally aligns the gesture stroke peak to the locomotion tempo.

3 Ground truth data

While there are many studies of conversational behaviors conducted with sitting or standing participants, we wish to explore how gesture and gaze behaviors differ in the case of conversers in motion. For this paper, we focus on walking and jogging scenarios.

We recorded real video footage of two sets of male and two sets of female participants, aged between 21 and 39, talking together while standing, walking and jogging. To encourage a natural and lively discussion we selected participants who knew each other and chose conversation topics that were of interest to them. Video and audio were recorded with a Sony HDR-AS100V action camera with a resolution of 1920x1080 and a frame rate of 29. In all conditions, the participants were placed next to each other and orientated toward the camera. The video camera was placed in front of the participants capturing the whole upper body for later annotation of the head rotation and gesture (see Fig. 2). In the walking and jogging conditions, a continuous path of approximately 200 meters was chosen with the camera moving in front of the subjects.

To create conversations with varying dynamics, we recorded two different conditions: *dominant speaker* conversations and *debates* (as in [9]). In the debates, each participant expressed their opinion on the topic being discussed with interruptions from the conversational partner. An informative topic was chosen for the dominant speaker conversations, where one speaker did the majority of speaking, while the other politely listened with only occasional responses or questions. In total 24 dominant speaker conversations (3 for each of the 8 participants) and 12 debates (3 with each of the 4 groups of participants) were recorded. Dominant speaker conversations lasted approximately one minute, while debates lasted between one and two minutes.

3.1 Annotation

We annotated every fifth frame of the video footage to capture all important information, such as head turns and gestures. To approximate head rotation and to compensate for camera and participants' movement, the x and y pixel coordinates of a center point on the body and the tip of the nose were marked. The participants' verbal behavior was noted as *talking* or *listening*. For each gesture, we reported the type and magnitude (0.5 and 4, with 1 being the relaxed hand position of the participant); the elbow bend (0 to 3, with 0 representing no bend and 2 a bend of 90 degrees); the arm displacement (0 representing the arm close to the body and incrementing steps thereafter); the facing direction of the palm (up, down, in or out) and the peak of the gesture.

For the type of gesture, we used the taxonomy proposed by McNeill [21]: 'Beat' - a rhythmic flick of finger, hand or arm to highlight what is being said; 'Deictic' - a pointing gesture with direction; 'Iconic' - a representation of a concrete object, or drawing with the hand; and 'Metaphoric' - a representation of an abstract concept. We also noted 'Adaptor' motions, such as crossing arms and touching the face or hair.

Each gesture is divided into 4 phases: preparation, stroke, hold, and retraction. During the stroke phase, the gesture normally peaks. The locomotion contact or flight information was annotated relative to this peak. During walking, a *contact* occurs when the front foot is close to or touching the ground, whereas one leg passing the

other represents a *flight* (see Fig. 3). For jogging, flight occurs when both legs are in the air and for contact at least one foot is on the ground.

According to empirical observations, we hypothesized that individuals would use more gestures when standing compared to walking or jogging. Similarly, we suspected the range and duration of gestures to be higher during a standing conversation. We also hypothesized that personal style would persist across gaits, and that the peak of a gesture would occur during the contact phase of locomotion.

3.2 Analysis

Frequency and duration of gaze shifts: For each participant, we averaged over the number of times they gazed at or away from their conversational partner for all conversation types. An Analysis of Variance (ANOVA) was conducted with factors gait, and gaze direction. We found an interaction effect between gait and the gaze direction ($F(2, 12) = 72.834, p \approx 0$). Participants gazed at their conversational partners more often while standing than while walking or jogging. We also conducted an ANOVA with dependent variable gaze duration and independent variables gait, and gaze direction, which showed an interaction effect ($F(2, 12) = 17.611, p \approx 0$). With increasing intensity of body motion, the average gaze duration toward the conversational partner decreases. Both of these results suggest that it is harder and at times not feasible to initiate and maintain eye contact during physical activities.

Frequency and duration of gestures: An ANOVA with the percentages of different types of gestures as the dependent variable and within factors gait, conversation type, and type of gesture was conducted. The categorical predictor was the sex of the participant. A main effect of gait ($F(2, 12) = 6.154, p = 0.015$) showed that participants gestured significantly less when jogging, and a main effect of conversation type showed that fewer gestures were used during debates than in dominant speaker conversations ($F(1, 6) = 13.172, p = 0.011$). However, an interaction between the sex of the participant and the conversation type ($F(1, 6) = 6.194, p = 0.047$) indicated that the male participants gestured significantly less during debates than dominant speaker conversations, whereas women gestured equally for both. Previous studies have also found that females gesture more during conversation [22].

A main effect of gesture type ($F(3, 18) = 14.907, p \approx 0$) indicated that the most common gesture type used was beat, followed by adaptor. There was also an interaction between gesture type and gait ($F(6, 36) = 2.979, p = 0.007$) showing that adaptors were used much more in standing than walking or jogging. An analysis on the duration of gestures gave an interaction effect between gait and participant sex ($F(2, 12) = 4.379, p = 0.037$), as male gestures were shorter during jogging, whereas the gesture duration for woman was the same for different gaits.

Gesture peaks: An ANOVA on the number of gesture peaks, with independent variables gait and locomotion phase found a main effect of locomotion phase ($F(1, 6) = 268.44, p \approx 0$), where significantly more gesture peaks happened during the contact phase. No main effect of gait was found suggesting that gesture peaks were similar across gaits. However, an interaction occurred between locomotion phase and gait ($F(1, 6) = 9.070, p = 0.024$), which shows that the percentage of gesture peaks in locomotion contact phase is significantly higher than in flight phase for both gaits.

Elbow bend: As hypothesized, an ANOVA showed a main effect of gait on the gesture space ($F(2, 12) = 10.563, p = 0.002$). During jogging, most gestures were made with the elbow bent by 90 degrees. The average elbow bend while standing and walking was significantly lower and ranged between 0 and 76.5 degrees.

From these results we can conclude that people

make fewer and smaller gestures and gaze motions with increasing body motion intensity. We also conducted a short post-experiment survey and most participants rated the question: *"How engaged in the conversation do you think your partner was in the following situations?"* the lowest for jogging. It seems reasonable to assume that the physical exertion, reduced eye-contact and gestures led to this assessment. We used these results to guide our stylized splicing and synchronization algorithms in order to generate natural conversational behaviors for conversing characters while walking or jogging.

4 Coordinated Gesture and Locomotion

In this section, we present our method for generating natural conversational behavior for pedestrians in motion, given locomotion clips from a variety of actors and standing conversational motion clips with gestures.

We use two existing motion corpuses: (i) 19 standing conversations with three male (9) or three female (10) actors, each approximately 160 sec. long, for a total performance time of 8,403 sec. [9]; and (ii) normal walking and jogging motions captured from 16 male and 16 female actors, with varied styles of arm expansion, elbow bend and swing amplitude [23]. All motions are retargeted to a 22-joint 69-DOF skeleton. All the captured motions are re-sampled to 30 fps to give a common time base.

Naively splicing two such clips generates unnatural results, thus the gesture performance needs to be customized before being spliced with the locomotion. Figure 4 illustrates the general work flow: (1) we ensure stylistic consistency between the clips by adjusting the gesture performance to match a particular locomotion style; (2) we temporally micro-synchronize the gesture phase with the locomotion cycle, and simulate arm disturbances resulting from the body's ground interaction; (3) we fully exploit the functions of gesture preparation and retraction for smooth splicing; (4) for conversations with two or more participants, we coordinate their pairings by adding head and torso orientations. Our method builds on previous splicing approaches, such as [14], by handling stylization, synchronization and conversational pairings.

To extract temporal information from the gestures, we performed the following annotation of the gesture databases. Each gesture phrase is temporally composed of preparation, stroke, hold and retraction phases [24, 25, 26], where ($gesture \rightarrow [preparation]\ [hold]\ stroke\ [hold]$). The main meaning of the gesture is carried in the *stroke* phase [26, 27]. The *preparation* phase places the arm, wrist, hand and fingers in the proper configuration to begin the stroke [26]. The *retraction* phase returns the arm to a resting position. During the annotation, we label the timing of gesture phrases and phases (t_{Pb}, t_{Pe}, t_{Sb}, t_{Se}, t_{Rb} and t_{Re}), corresponding to **b**eginning and **e**nd of **P**reparation, **S**troke and **R**etraction. The annotated gesture types are based on the taxonomy proposed by McNeil [21], as in the ground truth study. We also annotate gesturing handedness and the addresser/addressee in the group.

To extract locomotion tempo, we use a standard breakdown of locomotion into four phases, as previously mentioned in Section 3.1 (see Fig. 3). Typically, during flight the root altitude increases, whereas it decreases during contact. All our locomotion data clips are between 1.5 and 2 seconds long, starting from phase *left contact*, consisting of two full locomotion cycles and can be seamlessly looped to be any given length.

4.1 Stylization

As mentioned before, naively splicing gestures onto locomotion clips produces unrealistic results as it does not take into account the differ-

ences between gestures while standing or in motion. For example, low gestures with straight elbows are common for standing characters, but are unnatural in a jogging situation as our ground truth analysis shows that most of the gestures are made with the elbow bent to around 90 degrees. Furthermore, variation in the styles of locomotion should be transferred to the gesture style. For example, pedestrians with a larger arm swing are more likely to perform broader gestures. The goal of our stylization process is to adjust gestures to be consistent with the locomotion arm shape and swing. We consider gestures to be auxiliary actions on base motions like standing or jogging, and use the statistics of the base motions to offset the gestures.

For every locomotion clip M_L, we compute its mean arm pose B_L as a base pose, including the shoulder, elbow and wrist DOFs. Similarly for gesture performance M_G, we use its rest pose as the base B_G. B_L and B_G thus reflect the overall correspondence between the arms and the torso (see Fig. 5). The difference between the base poses, $B_D = B_L - B_G$, is then used to adjust the original gesture motion M_G. Gestures are extracted from the standing character as an offset from the base pose and then layered onto the base pose of the desired locomotion clip, which generates $M'_G = M_G + B_D$.

To incorporate the dynamic features of the locomotion arm swing, we compute the standard deviation of arm DOF d_i in the locomotion clip. Arm rotations in M'_G are constrained within $\pm c_i * STD(d_i)$, where c_i is a user specified constant, typically three. Joint rotations exceeding this active range are linearly rescaled to the allowable range. To avoid altering the pointing direction, stylization is not applied to deictic gestures.

4.2 Synchronization

Temporally, a gesture has its preparation, stroke, hold and retraction phases, while locomotion repeats its flight/contact cycles with a certain tempo. From the ground truth study, we found that both are linked, in that significantly more stroke peaks happen during contact phases. Pedestrians are therefore likely to align their stroke peaks to the locomotion contact phase. Some stroke emphases are actually caused by the arms being shaken due to ground impact during locomotion, which produces an effect of gesturing to the tempo of the locomotion cycle. We synthesize this effect to make the gesture performance more realistic during locomotion, especially for joggers.

Alignment: Unlike previous splicing research that uses Dynamic Time Warping (DTW) alignment [14] or velocity based synchronization [16] to align arm motion with the locomotion, we align gestures based on the timing of the utterance and locomotion phase. The gesture 'synchrony rules' referred to in [21] indicate that gesture strokes have been observed to consistently end at or before, but never after, the prosodic stress peak of the accompanying syllable. User studies in [28, 29] have indicated that gestures that are performed 0.2 to 0.6 second earlier w.r.t the accompanying utterance are rated highly for their naturalness. This provides us with an exact time window for gesture alignment: for a given gesture, if it has a stroke peak and does not align with the locomotion contact phase based on the utterance timing, we search 0.6 seconds behind to find the first contact phase point. We then align the stroke peak with this contact phase. We do not perform re-alignment for gestures with multiple stroke peaks, to avoid conflicts in alignment and also to preserve the original time gaps between these peaks.

Synthesis: During the contact phase of locomotion, the body hits the ground and suddenly changes its velocity. It is unlikely that a person could hold their arms as steady in this condition as a standing character could, especially for joggers. Some stroke emphases are actually caused by the shaking arms resulting from locomotion,

the effect of which could vary, due to different arm firmness and also the vigor of the jogging. To synthesize arm shake to the beat of the locomotion, we use the motion of the root to adjust the elbow. Using $R'_{elbow} = R_{elbow} + k * \Delta H_{root}$, on top of the original elbow rotation R_{elbow}, we layer the influence of root height change ΔH_{root} due to the ground impact. For a walking motion, ΔH_{root} is negligible, but for jogging motions, ΔH_{root} is large and the elbow bounce is very obvious. If k is an adjustable parameter that rescales the height changes to the elbow rotation space, then by increasing k we can synthesize jogging on a bumpy road with loose arms.

4.3 Splicing

The goal of our splicing method is to add the gesture performance M_G to locomotion clip M_L given the gesture timing, and to generate the output spliced motion M_S that naturally combines the two. We further segment the skeleton into torso, lower-body, left arm and right arm. A full-body motion sequence M^{fb} can thus be described by the union of the motion of its four parts: left arm M^{la}, right arm M^{ra}, torso M^{ts} and lower body M^{lb}, where each part should maintain close correlation with each other. As the lower-body motion is the dominant factor in locomotion, and the torso swivels to its tempo, we preserve $M_L^{lb}(t)$ and $M_L^{ts}(t)$ throughout time $t \in [0, N]$ (Eq. 1).

$$\begin{array}{rcll} M_S^{lb} & = & M_L^{lb}, & t \in [0, N] \\ M_S^{ts} & = & M_L^{ts}, & t \in [0, N] \end{array} \quad (1)$$

The stroke and hold phases carry the semantics of the gesture, thus $M_G^{la}(t)$ and $M_G^{ra}(t)$ are preserved in the spliced motion from the beginning of the stroke t_{Sb} to the end t_{Se}.

Our method takes the gap between the beginning of the gesture preparation t_{Pb} and the end of the preparation t_{Pe}, and applies spherical linear interpolation (slerp) to the arm joint rotations to transition from the locomotion swing to the gesture performance. Similarly, slerp is applied during the gap between the beginning of the gesture retraction t_{Rb} and its end t_{Re} to transition gesture performance back to locomotion swing (Eq. 2).

$$M_S^{arm} =$$

$$\begin{cases} M_L^{arm}, & t \ ! \in [t_{Pb}, t_{Re}] \\ M_G^{arm}, & t \in [t_{Sb}, t_{Se}] \\ \text{slerp}(M_L^{arm}, M_G^{arm}, \frac{t - t_{Pb}}{t_{Pe} - t_{Pb} + 1}), & t \in [t_{Pb}, t_{Pe}] \\ \text{slerp}(M_L^{arm}, M_G^{arm}, \frac{t - t_{Rb}}{t_{Re} - t_{Rb} + 1}), & t \in [t_{Rb}, t_{Re}] \end{cases} \quad (2)$$

5 Interaction

Using the motion splicing and gesture stylization methods described above, we are able to synthesize multiple gesturing characters in locomotion. To make these characters appear plausibly engaged in a conversation, head and torso orientations need to be added to generate appropriate gaze behavior. This is done based on an 'addresser-addressee relationship' (AAR) that defines the conversational interaction, where the addresser gazes toward the addresee. This AAR specification includes high level information such as labeling the addresser, addressee and the timing of gaze behavior. The system supports multiple ways of generating the AAR, including random specification, manual specification, respecting AAR from the original motion captured group conversation, or generating it based on the statistics from our ground truth study.

Gaze Generation: Gaze is implemented by first dynamically retrieving the positions of the addresser and addressee in the scene at the specified time, and then computing the θ_{yaw} that would fully rotate one character's head to look at the other, in the horizontal plane. Since gaze also involves eye movement, a complete head rotation is not always necessary. We use a distribution to determine the torso yaw angle as $r * \theta_{yaw}$ where r is randomly chosen from $[.6, 1]$. The rotation is implemented with a combination of spine and neck DOFs. If the addressee is in front of or

behind the addresser, exceeding a threshold distance (one meter in the current implementation), a small adjustment of forward/backward lean up to 15° is added to the spine joint of the addresser. After stylized splicing, $M^{torso}_{Spliced}$ is directly derived from locomotion M_{Loco}, which preserves the natural motion of the torso, timed with the locomotion tempo. The newly synthesized AAR head and torso orientations are layered on top of this torso movement.

AAR Specification: There are several ways to generate the AAR. Firstly, it may be inferred from the motion clip by assuming that the character performing a gesture is the addresser and the others are addressees. Gesture timing is used for head and torso rotation whereby the gaze direction is achieved by the start of the stroke and returns to neutral with the retraction. This method allows an originally captured standing group conversation to be transformed into one with locomotion, or any gesture specification to generate gaze behavior. Alternatively, we allow a user to fully author the AAR such that the user has complete control at the cost of some added up front labor. This flexible specification can help to pair participants from different conversations in the database to simulate a new group conversation. Our system can also randomly select addresser and addressees and pair them into AAR. This method facilitates simulating plausible crowd conversations during locomotion, with minimal user intervention.

Synthesize AAR from Ground Truth: Finally, our ground truth data may be leveraged to create more complex and realistic gaze behavior. Data analysis suggests that both the addresser and addressee will gaze at and away from each other during the conversation and that the duration of this behavior is not necessarily the same as the duration of a gesture. We analyzed the duration of the gaze at and away behavior for each of our subjects for both walking and jogging and used this for some of our experiments. Gaze behavior during the conversation is thus determined for each addresser and addressee based on the statistics of the subject model assigned to them (selected from the ground truth data). When a particular behavior is chosen, say gaze-at, it is assigned a duration based on the subject model with a small random variation. If subsequent selections of the same gaze behavior exceed a total duration greater than the average duration plus one standard deviation, the gaze behavior is forced to switch to the opposite type. Experiments were conducted with different sampling strategies, and this method was found to generate a pattern of gazing at and away behavior that appeared natural and non-repetitive.

6 Result and Applications

To evaluate the effectiveness of our method, we apply them in a number of different scenarios. Please see the supplemental videos for full animated results.

Stylized Splicing: To demonstrate the advantage of stylistic splicing, we select five distinct jogging types, detailed in Table 1. We experiment with different gesture types, varied wrist positions and stroke amplitudes.

Fig. 6 compares the results of splicing a low gesture from a standing posture on five different jogging motions. As mentioned before, low gestures with straight elbows are common for standing characters; however, splicing them directly into a locomotion clip can generate unrealistic artifacts as the base pose of jogging is quite different from the standing rest pose. In Fig. 6(a), naïve splicing not only produces an inconsistent straight arm configuration in the middle of a jogging arm swing, it also generates an identical gesture performance for the different jogging styles. Fig. 6(b) demonstrates how stylized splicing can effectively fix these problems by transferring the jogging base poses, resulting in a more believable gesture performance for the jogging locomotion.

Jogger J1 has a large elbow bend, small arm expansion and a positive swing, so the spliced gesture is also performed high and narrow in front of the body. J2 has a larger arm expansion during the jogging swing, so the gesture is also wider. J3, J4 and J5 have less elbow bend with different variances. Their elbows are more straight when performing the gesture, but J4 has narrower arm expansion for the gesture, and J5 preserves his asymmetric style.

Conversation Simulation: Based on the AAR information from the gesture database annotation, our method converts the motion captured standing conversations into locomoting ones. Figure 1 shows the same group conversation in different locomotion conditions. Head and torso orientation is calculated according to the new position of the addressee in the scene. For this experiment we used gaze timing profiles extracted from the ground truth study.

Our method can also simulate conversational relationships that vary from the original motion captured group structures. Figure 7 shows a simulated group conversation, using the same gesture and locomotion input as above, but in this case, no AAR specification is necessary from the user. By default, we randomly pick one addressee in the audience and pair it with the addresser to establish their conversational relationship. Cross-group communication different from the original motion capture data is highlighted. We allow further specification from the user for detailed control.

7 Conclusion

We have presented a novel method for generating conversational gestures for virtual pedestrians. Animators can fully reuse existing clips of locomotion and standing conversations. "Stylized splicing" flexibly adjusts gesture behavior in time and space to the locomotion style. Using AAR specification, virtual pedestrians can be dynamically paired into conversational groups, which allows the simulation of crowd conversations. Our ground truth data can also be used as a solid reference for animators to generate gestures for pedestrians.

Currently, the gesture performance and head/torso orientation ground truth is extracted from videos. In the future, analyzing motion capture data of conversations during locomotion could help to more precisely quantify the data. Our method is capable of splicing gestures into any synthesized locomotion. Integrating the technique with a general motion graph would support a larger variety of scenes and different locomotion paths. Given the extracted ground truth information, a motion retrieval algorithm could be used with our method to efficiently search the gesture database and find the most ideal performance for the sequence. We hope that our work can contribute to a new body of research on gesture synthesis for a wide set of naturalistic activities.

References

[1] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1):39–52, 2004.

[2] S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics (TOG)*, 28(5), 2009.

[3] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics (TOG)*, 29(4):124, 2010.

[4] M. Neff, M. Kipp, I. Albrecht, and H. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker

style. *ACM Transactions on Graphics (TOG)*, 27(1), 2008.

[5] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513, 2004.

[6] J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. Beat: The behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 477–486. ACM, 2001.

[7] S. Marsella, J. Gratch, and J. Rickel. Expressive behaviors for virtual worlds. In *Life-like characters*, pages 317–360. Springer, 2004.

[8] D. Bohus and E. Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 5:1–5:8. ACM, 2010.

[9] C. Ennis, R. McDonnell, and C. O'Sullivan. Seeing is believing: Body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)*, 29(4):91:1–91:9, 2010.

[10] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. *ACM Transactions on Graphics (TOG)*, 21(3):473–482, 2002.

[11] A. Safonova and J.K. Hodgins. Construction and optimal search of interpolated motion graphs. *ACM Transactions on Graphics (TOG)*, 26(3), 2007.

[12] M. Gleicher, H. Shin, L. Kovar, and A. Jepsen. Snap-together motion: assembling run-time animations. In *ACM SIGGRAPH 2008 Classes*, SIGGRAPH '08, pages 52:1–52:9. ACM, 2008.

[13] A. Fernández-Baena, M. Antonijoan, R. Montaño, A. Fusté, and J. Amores. Bodyspeech: A configurable facial and gesture animation system for speaking avatars. *Proceedings of the International Conference on Computer Graphics and Virtual Reality (CGVR)*, page 3, 2013.

[14] R. Heck, L. Kovar, and M. Gleicher. Splicing upper-body actions with locomotion. *Computer Graphics Forum*, 25(3):459–466, 2006.

[15] Pengcheng Luo and Michael Neff. A perceptual study of the relationship between posture and gesture for virtual characters. In *Motion in Games*, pages 254–265. Springer Berlin Heidelberg, 2012.

[16] C. Mousas, P. Newbury, and C. Anagnostopoulos. Splicing of concurrent upper-body motion spaces with locomotion. *Procedia Computer Science*, 25:348–359, 2013.

[17] K. Tamada, S. Kitaoka, and Y. Kitamura. Splicing motion graphs: Interactive generation of character animation. *Short papers of Computer Graphics International*, 3, 2010.

[18] W. Ng, C. Choy, D. Lun, and L. Chau. Synchronized partial-body motion graphs. In *ACM SIGGRAPH ASIA 2010 Sketches*, pages 28:1–28:2. ACM, 2010.

[19] N. Al-Ghreimil and J.K. Hahn. Combined partial motion clips. In *WSCG'03*, 2003.

[20] A. Majkowska, V.B. Zordan, and P. Faloutsos. Automatic splicing for hand and body animations. In *Proceedings of the*

2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pages 309–316, 2006.

[21] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[22] N.J. Briton and J.A. Hall. Beliefs about female and male nonverbal communication. *Sex Roles*, 32(1-2):79–90, 1995.

[23] L. Hoyet, K. Ryall, K. Zibrek, H. Park, J. Lee, J.K. Hodgins, and C. O'Sullivan. Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. *ACM Transactions on Graphics (TOG)*, 32(6):204:1–204:11, 2013.

[24] D. Efron. *Gesture and environment*. Nkig's Crown Press, 1941.

[25] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25:207–227, 1980.

[26] Sotaro Kita, Ingeborg Van Gijn, and Harry Van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and sign language in human-computer interaction*, pages 23–35. Springer, 1997.

[27] D. McNeill. *Gesture and thought*. University of Chicago Press, 2005.

[28] C. Kirchhof. On the audiovisual integration of speech and gesture. *Int. Soc. Gesture Studies (ISGS)*, 2012.

[29] Y. Wang and M. Neff. The influence of prosody on the requirements for gesture-text alignment. In *Intelligent Virtual Agents (IVA)*, pages 180–188, 2013.

Authors' Biographies:

Yingying Wang is a Ph.D. student in the Department of Computer Science at the University of California, Davis. Her main research is on character animation, motion capture, retrieval and perception.

Kerstin Ruhland received her Degree in Computer Science from the Ludwig Maximilian University of Munich in 2009. Since March 2013, she has been working toward a Ph.D degree at Trinity College Dublin, under the supervision of Dr. Rachel McDonnell. Her main focus is on realistic facial and eye animation.

Michael Neff is an associate professor in Computer Science and Cinema & Digital Media at the University of California, Davis where he directs the Motion Lab. He holds a Ph.D. from the University of Toronto and is also a Certified Laban Movement Analyst. His interests include character animation tools, especially modeling expressive movement, physics-based animation, gesture and applying performing arts and psychology research to animation. He received an NSF CAREER Award and the Alain Fournier Award

Carol O'Sullivan is the Professor of Visual Computing at Trinity College Dublin, where she has been on the faculty since 1997. From 2013-2016 she was a Senior Research Scientist with Disney Research Los Angeles, based in Glendale, CA and a Visiting Professor in Seoul National University from 2012-2013. Her research interests include perception, animation, virtual humans, and crowds.

Figure 1: Standing conversation (l) converted to walking (m) and jogging (r) groups.

Figure 2: Two participants walking

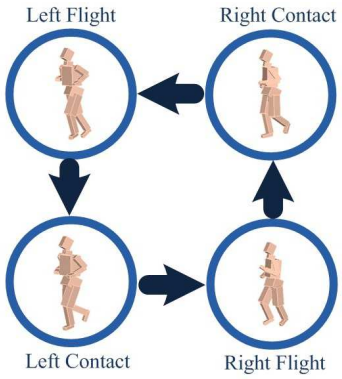


Figure 3: Locomotion Cycle: jogging

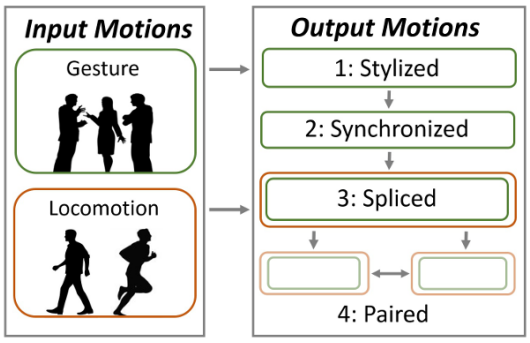


Figure 4: Overall work flow for performing stylized splicing and synchronization of conversing characters when walking or jogging.

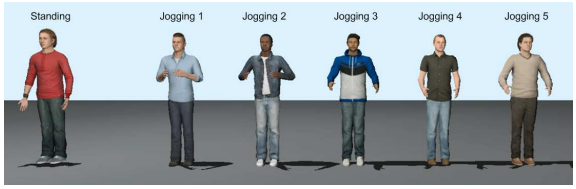
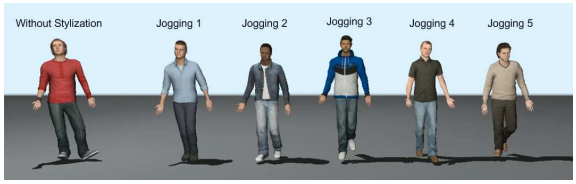


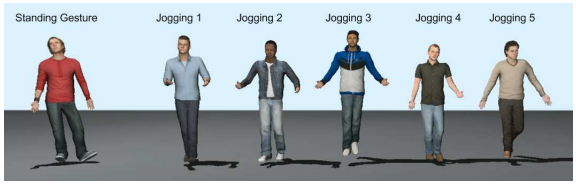
Figure 5: Base arm poses for standing and 5 jogging styles (Table 1)

ID	Expansion(y)	Swing(z)	Bend(z)
J1	(13.8 ± 1.1)	(27.3 ± 4.5)	(113.8 ± 3.1)
J2	(30.6 ± 3.5)	(-6.4 ± 13.7)	(109.9 ± 6.3)
J3	(21.5 ± 0.9)	(-3.5 ± 14.6)	(82.7 ± 11.8)
J4	(9.9 ± 3.1)	(-3 ± 8.6)	(68.4 ± 7.5)
J5	(22.5 ± 1.8)	(-2.9 ± 13.9)	(74.7 ± 13.9)

Table 1: Quantized Jogging Styles (approx. mean and stdev in deg.)



(a) Without stylization, the spliced arm strokes look stiff and inconsistent.



(b) Gestures consistent with locomotion styles

Figure 6: Comparison demonstrating the effect of gesture stylization.

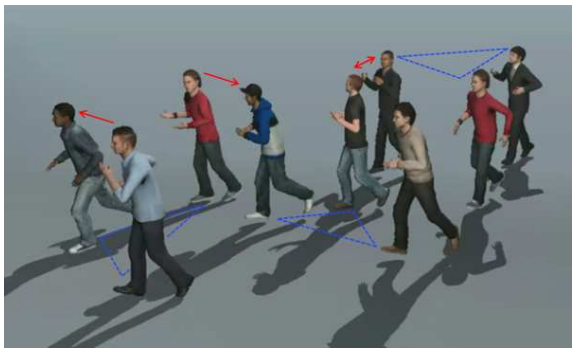


Figure 7: Simulated large group of joggers conversing with random AARs: original (blue), cross group (red).