Synthesizing Realistic Image-based Avatars by Body Sway Analysis

Masashi Nishiyama, Tsubasa Miyauchi, Hiroki Yoshimura, Yoshio Iwai
Graduate School of Engineering, Tottori University
Tottori , Japan
nishiyama@eecs.tottori-u.ac.jp

ABSTRACT
We propose a method for synthesizing body sway to give human-like movement to image-based avatars. This method is based on an analysis of body sway in real people. Existing methods mainly handle the action states of avatars without sufficiently considering the wait states that exist between them. The wait state is essential for filling the periods before and after interaction. Users require both wait and action states to naturally communicate with avatars in interactive systems. Our method measures temporal changes in the body sway motion of each body part of a standing subject using a single-camera video sequence. We are able to synthesize a new video sequence with body sway over an arbitrary length of time by randomly transitioning between points in the sequence when the motion is close to zero. The results of a subjective assessment show that avatars with body sway synthesized by our method appeared more alive to users than those using baseline methods.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Image-based Avatar, Body sway, Human-like movement

INTRODUCTION
Interactive systems that use human-like avatars are an active topic in human-agent interaction research, and have many potential applications, such as conversational avatars in a museum [15], avatars talking about the past [2], and speech-interactive guidance avatars [10]. These avatars have the potential to actively utilize the large-sized displays that we often see many places, e.g., in stations, shopping plazas, office entrances, and airports. These interactive avatars can automatically communicate with users without the constraints of time and place (e.g., at an automatic information desk late at night). In this paper, we focus on synthesizing human-like realistic avatars by exploiting an image-based technique [2, 9] in which life-sized video sequences interact with users through large-sized displays.

This is the draft version of our paper of HAI 2016.

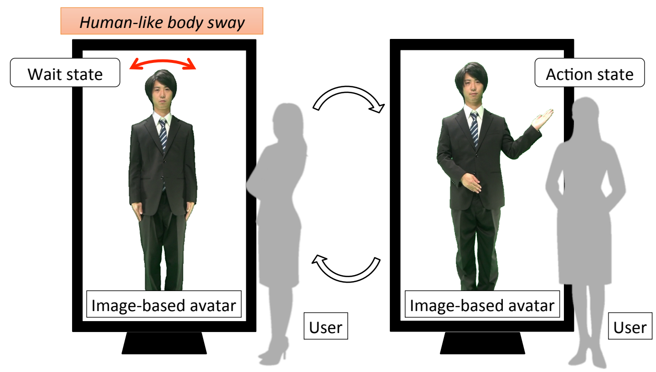


Figure 1. We consider the body sway of an image-based avatar in the wait state to enhance the nonverbal communication between it and a user in an interactive system.

In order to synthesize realistic avatars for interactive systems, we need to consider two states. The first is the action state (e.g., speaking with a user) and the second is the wait state (e.g., maintaining a standing pose). In particular, the wait state plays an important role in the interaction between avatars and users because the avatars do not continuously communicate with the users. The wait state is required to handle periods before and after interaction with users. For instance, if the wait state is not adequately applied in the avatars, the users cannot judge whether to begin communication with the avatars or not. However, existing methods [2, 9] do not take the wait state into consideration; they mainly handle the action state.

Our key idea is to exploit the wait state of the avatars to enhance the nonverbal communication between users and avatars, as illustrated in Figure 1. To achieve this, we synthesize human-like micro movements. Consider a standing pose in the wait state; we often see standing receptionists at airports, hotels, and information offices. When people maintain a standing pose, they continuously move their body slightly around its center line. As described in [11], people unconsciously spread the burden of standing across their muscles by making micro movements to avoid overloading one set of muscles. This physiological phenomenon is called body sway.

In this paper, we propose a novel method for synthesizing human-like body sway for image-based avatars in the wait state. To analyze the body sway of real people, we first acquire video sequences of standing subjects and measure the movement of each body part. We extract the characteristics of body sway from the temporal changes of this movement. Using these characteristics, we can synthesize image-based avatars with body sway for an arbitrary length of time using identified transitioning times in the video sequences.

RELATED WORK

Interaction systems need to have an intuitive feel so that users can naturally communicate with avatars as if they were communicating with real people. As described in [6, 7], to design human-like avatars for interaction systems, video synthesis and speech synthesis are key techniques. In particular, video synthesis techniques are directly linked to the appearance of the avatars in nonverbal communication [3, 5, 4]. Researchers have developed various video synthesis techniques for generating realistic avatars, and there are currently two main approaches. The first exploits a computer graphics-based avatar [15, 10] and the second exploits an image-based avatar [2, 9]. In general, the appearance of image-based avatars is more human-like than that of computer graphics-based avatars. However, image-based avatars can only replay video sequences stored in databases. Thus, image-based avatars sometimes cannot work well when users perform unexpected actions. To overcome this difficulty, a technique [12, 1, 14, 17, 8] that synthesizes various video action sequences has been proposed. Some methods [12, 1] synthesize facial video sequences in conversation. Okwechime et al. [14] proposed synthesizing full-body video sequences of actions in conversation. Shi et al. [17] proposed editing video sequences of interactive actions between subjects, and Huang et al. [8] developed a method to seamlessly combine various actions. However, these existing methods do not sufficiently consider synthesizing video sequences of the wait state. Our method tackles the synthesis of realistic image-based avatars that contains human-like body sway.

ANALYSIS OF BODY SWAY

Design of measurement of body sway

Here, we consider a scenario in which standing people are always visible to the public, such as receptionists in public spaces. In order to synthesize this kind of body sway on human-like avatars, we need to measure the movement of each body part of a person in a standing pose. We also need to check whether the body parts move synchronously or not.

To analyze the body sway of real people, a footing force plate is widely used. However, this device measures temporal changes of the body's center of gravity and cannot acquire the motion of specific body parts. Nashner [13] measured sway motions using acceleration sensors attached to the body. These sensors have the disadvantages of needing to be attached to a user and are difficult to temporally synchronize. Recently,

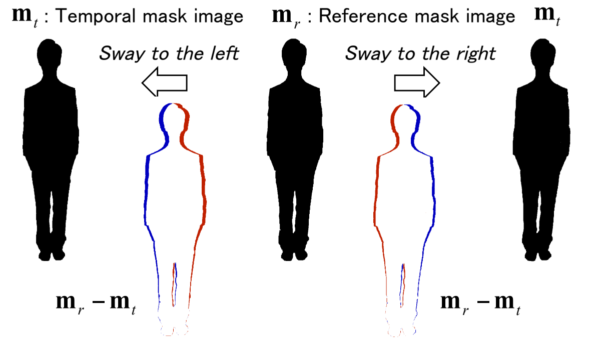


Figure 2. A reference mask image and temporal mask images. Red represents a region converted from body to background and blue represents the opposite case.

Wang et al. [19] inferred the position of the center of gravity of the whole body using multiple video cameras and Yeung et al. [21] inferred the center of gravity position from body joints using common gaming sensors. However, these existing methods handle only the center of gravity position of the body and insufficiently consider the micro movements of each body part. We aim to measure the amounts of movement of each body part using a simple video camera without body-attached sensors. The details of our method are described below.

Algorithm for measuring body sway

Our method uses a binary mask image in which a pixel within a body region is 1; otherwise it is 0. We acquire a video sequence of N frames from a standing subject. Given a reference time r, representing the center of the body sway in the video sequence, we determine a reference mask image $\mathbf{m}_r$. We compare the reference mask image with temporal mask images $\mathbf{m}_t (t \in 1, \ldots, N)$, as illustrated in Figure 2. Our method computes the amount d_i of body sway movement as

$$d_i = \sum_{x \in \mathrm{parts}(i)} (\mathbf{m}_r(x) - \mathbf{m}_t(x)) , \quad (1)$$

where $\mathbf{m}_r(x)$ and $\mathbf{m}_t(x)$ indicate the pixel value at position x in $\mathbf{m}_r$ and $\mathbf{m}_t$, respectively, and $\mathrm{parts}(i)$ is a body part region indicated by i, such as the regions shown in Figure 4. The value of d_i is positive when the number of pixels converted from the background to the body is higher and negative when the number of pixels converted from the body to the background is higher. Note that we compute reference time r using Algorithm 1 to determine reference mask image $\mathbf{m}_r$.

Algorithm 1 Determine reference time r.

```
for r̃ = 1 to N do
    D_r̃ ← 0
    for t = 1 to N do
        compute d̃_i using m_r̃, m_t
        D_r̃ ← D_r̃ + Σ|d̃_i| for all body parts
    end for
end for
r ← arg min D_r̃
```

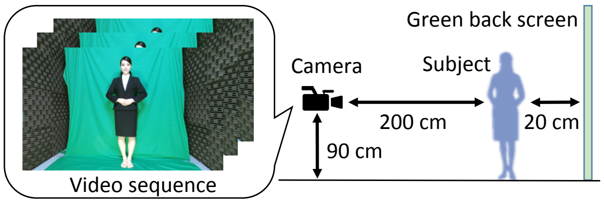


Figure 3. Setup for body sway measurement.

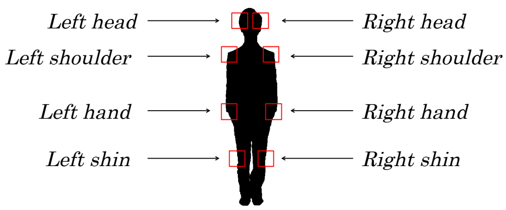


Figure 4. Body part regions for measuring body sway movement.

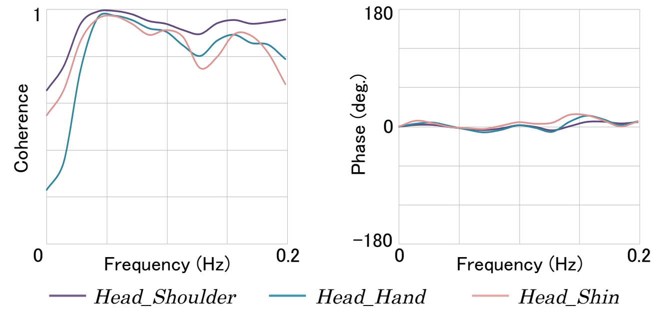


Figure 5. Coherence analysis of movements (180 s) of right body parts of subject A. The values of phase between these parts were nearly identical though the values of coherence were different.

Body part	Test Subject *A*	*B*	*C*	*D*	*E*	*Avg.*
Left head	289	132	185	237	197	208
Right head	296	107	191	265	200	212
Left shoulder	205	114	157	156	159	158
Right shoulder	232	86	140	198	155	162
Left hand	212	75	145	111	119	132
Right hand	191	63	88	107	126	115
Left shin	73	31	54	56	57	54
Right shin	89	32	50	53	55	56

Table 1. Standard deviation [pixel] of body sway movement $|d_i|$ for each body part.

Evaluation of body sway measurement

To evaluate the effectiveness of our method, we captured five video sequences of five test subjects (average age 21.4 ± 0.5 years, average height 166.9 ± 7.1 cm). Each test subject stood in front of a camera. Each sequence lasted 180 s. We used a video camera with 30 frames per second and $1,920 \times 1,080$ pixel resolution. We set the camera at a height of 90 cm and distance of 200 cm from the subject (Figure 3). The bounding box of the subjects is hence about 300 $\times$ 950 pixels. We generated mask images from the sequences with a simple background subtraction technique using a green back screen. We manually identified the head, shoulder, hand, and shin regions (Figure 4) for each video sequence. We set the size of each body part to 70 $\times$ 70 pixels.

Table 1 shows the standard deviation of body sway movement $|d_i|$ for each body part of each subject. We can see that these amounts were dynamically different between subjects. We can also see that the movements of the upper body parts were larger than the those of the lower body parts. We believe that all body parts synchronously moved in all subjects. We also observed the same tendency in Figure 5. Note that a change in d_i of about 380 pixels corresponds to a body part movement of 1 cm.

Figure 6 shows the temporal changes in the amount of movement for each body part. We can see that the temporal changes were different for each subject because the shape of the waves notably varies between subjects. We can also see that the temporal changes were nearly identical for all body parts; for instance, the head moved to the left and then the other parts moved in the same direction. Similarly, when the right parts moved to the left, the corresponding left parts moved in the inverse direction. We observed this same tendency among all subjects. We believe that the motion synchronously changes for all body parts.

Furthermore, we verified the temporal characteristics of the movement. Figure 7 shows temporal changes over 60 s for the left side of the head of subject A. We can see that the zero-crossing points (representing the reference times) repetitively appear and the temporal changes between the zero-crossing points forms a multi-modal wave. We observed the same tendency for all subjects. We believe that the temporal changes are periodic and contain the sway motion between the reference times.

Measurement error evaluation

We evaluated the error of our motion measurement. We used five reference mask images of the five subjects and 15 temporal mask images randomly selected from the video sequences. We manually calculated the ground truth of the body regions to generate the mask images. Figure 8 compares the amount of movement $|d_i|$ using the ground truth with that measured using the automatically extracted mask images. These results show that the error of measurement is very small for each body part. We thus believe that the results of our method are not affected by error from the mask images generated using the background subtraction technique.

Next, we compared the performance of our method with that of an existing method [21] that measures the 3D center of gravity using the body joints of Microsoft Kinect v2. We computed the 2D positions of the centers of gravity projected from the 3D positions because the purpose of our method is to measure the 2D variations of body sway in a video sequence. We compared the Root Mean Squared Error (RMSE) between the centers of gravity of manually labeled body regions and those of our method. The RMSE of our method is 1.1 pixels and that of the existing method is 36.5 pixels. Hence, our

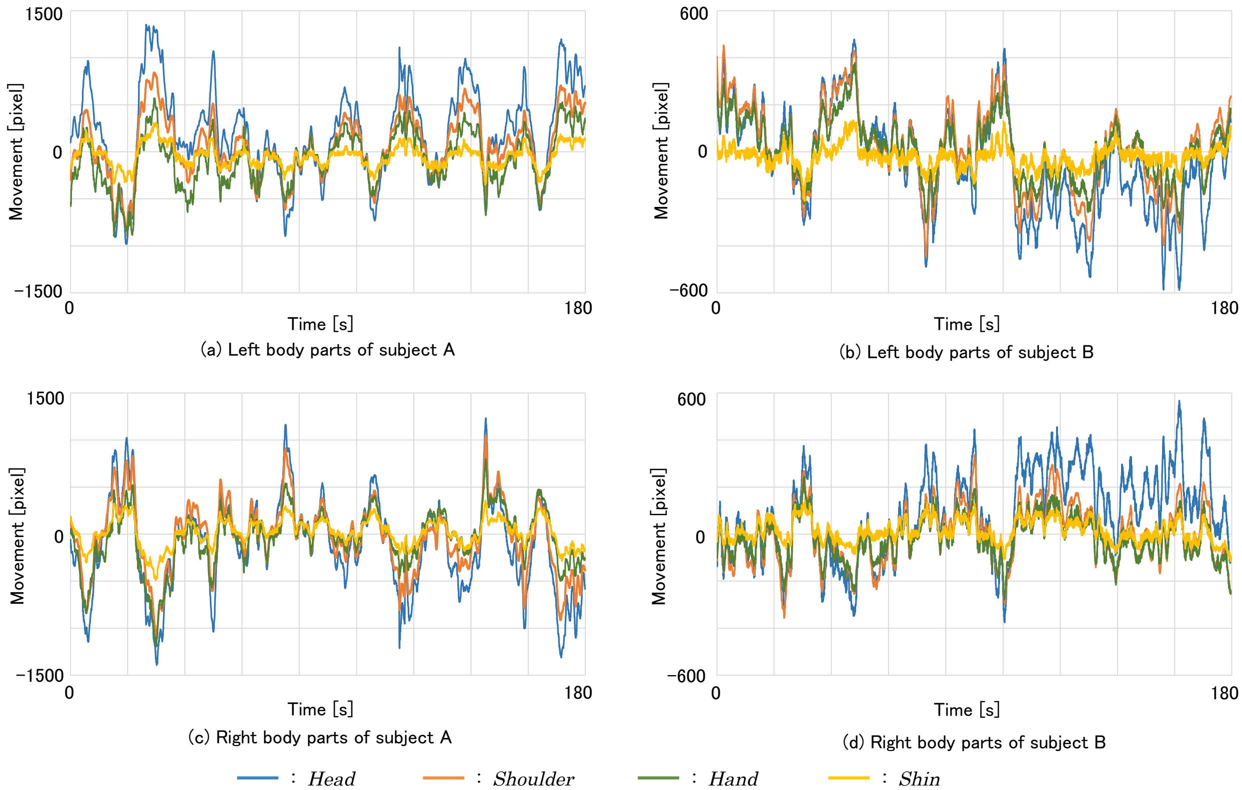


Figure 6. Measuring temporal movement caused by body sway for each body part.

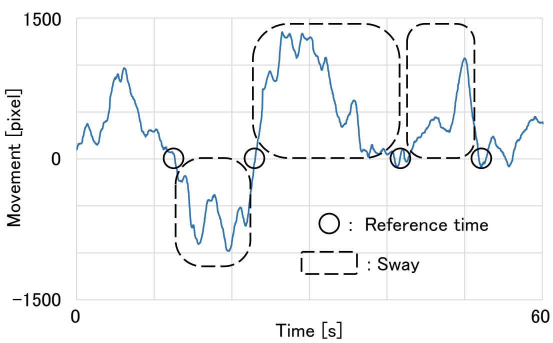


Figure 7. Temporal characteristics of body sway motion. First is that there is a plurality of reference times and second is that there is sways between the reference times.

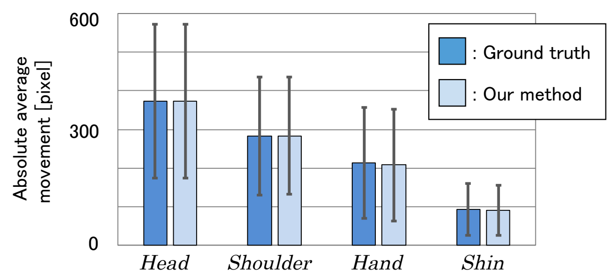


Figure 8. Performance of movement measurement using the simple background subtraction method. Error bars represent standard deviations.

method outperforms the existing method when measuring the 2D positions of body part centers of gravity.

SYNTHESIS OF IMAGE-BASED AVATARS WITH HUMAN-LIKE BODY SWAY

Design for generating body sway

We now describe how we synthesize body sway for image-based avatars. We consider exploiting the characteristics of the temporal changes of movement shown in Figure 7. Our method extracts the reference times of body sway from a video sequence captured from a subject for an image-based avatar and replays the video sequence by randomly transitioning between the reference times, as illustrated in Figure 9. Using this process, we can synthesize a body sway video sequence of arbitrary time length. However, we only handle the mask image representing the shape of the body in the discussion in the above section. This causes the problem that incorrect reference times are extracted, i.e., the shape of the body is similar but its texture is not. In order to synthesize a human-like body sway, we should take both the appearance and shape of the body into consideration. We also exploit the appearance of the face because users are sensitive to facial expressions. The details of this synthesis are described below.

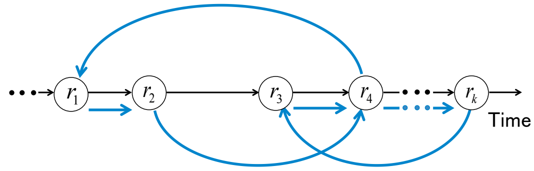


Figure 9. Synthesizing a body sway video sequence using a set of reference times r_k. Our method replays a video sequence of body sway by randomly transitioning between the reference times. The black arrows represent the flow of the original video sequence and blue arrows represent the flow of the synthesized video sequence.

Algorithm for computing shape and appearance similarities

To synthesize body sway, our method computes the similarity of the body shape, body appearance, and face appearance at each time t. We use mask image $\mathbf{m}_t$ to represent body shape, color image $\mathbf{a}_t$ to represent body appearance, and facial image $\mathbf{f}_t$ to represent facial appearance. Our method generates $\mathbf{m}_t$ using the same technique described in the section on measurement. Here, $\mathbf{a}_t$ refers to each frame of a video sequence. We generate the pose-aligned $\mathbf{f}_t$ using facial feature points inferred by a face tracking technique [16]. Given an initial reference time r_0, our method computes similarity $s_{t,s}$ of a body shape as

$$s_{t,s} \quad = \quad e^{-\lambda \sum |d_i|} \,, \tag{2}$$

where λ is a constant and d_i is the amount of motion measured using $\mathbf{m}_{r_0}$ and $\mathbf{m}_t$. Equation (2) uses the summation of d_i for all body parts. We use $s_{t,s}$ to extract the times that shifts between body and background regions do not occur. Next, we compute the similarity $s_{t,a}$ of body appearance as

$$s_{t,a} \quad = \quad \mathrm{SSIM}(\mathbf{a}_{r_0}, \mathbf{a}_t) \,, \tag{3}$$

where SSIM is the Structural SIMilarity [20], used to predict the perceived quality of frames of a video sequence. We use $s_{t,a}$ to extract the times of unvarying texture over the whole body. Furthermore, we compute similarity $s_{t,f}$ for the face appearance as

$$s_{t,f} \quad = \quad \mathrm{FaceSimilarity}(\mathbf{f}_{r_0}, \mathbf{f}_t) \,, \tag{4}$$

where FaceSimilarity is a correlation value using the edge-based feature vectors of $\mathbf{f}_{r_0}$ and $\mathbf{f}_t$. We use $s_{t,f}$ to extract the times of unchanging facial expressions such as between blinks. We compute the final integrated similarity s_t as

$$s_t \quad = \quad \alpha s_{t,s} + \beta s_{t,a} + \gamma s_{t,f} \,, \tag{5}$$

where $\alpha + \beta + \gamma = 1$. Using s_t, we aim to extract the reference times at which the shapes and appearances do not change much over the whole-body. Note that we determine initial reference time r_0 using Algorithm 2.

Transition between reference times

As described above, the reference times are suitable for seamlessly transitioning through the video sequence. We thus exploit the set of reference times r_k to synthesize human-like body sway. Our method generates the set of reference times by searching for times at which the shape and appearance of the body are close to each other (Algorithm 3). At these times,

Algorithm 2 Determine initial reference time r_0.

```
for r̃0 = 1 to N do
    S_r̃0 ← 0
    for t = 1 to N do
        compute s̃_t using m_r̃0, a_r̃0, f_r̃0, m_t, a_t, f_t
        S_r̃0 ← S_r̃0 + s̃_t
    end for
end for
r0 ← arg max S_r̃0
```

s_t reaches a local maximum and is greater than a threshold T_1. To prevent extracting reference times with short intervals, we also add a condition that reference times are selected more than an interval threshold of T_2 apart.

Algorithm 3 Generate set of reference times.

```
for t = 1 to N do
    compute s_t using m_r0, a_r0, f_r0, m_t, a_t, f_t
    if s_t is local maximum, s_t > T1, interval > T2 then
        add time t to set of reference times
    end if
end for
```

We next describe a video transition technique using the set of reference times (Algorithm 4). We randomly select a reference time r_k from the set and display color images $\mathbf{a}_t$ from the selected reference time r_k to the next reference time r_{k+1}. By repeating the random selection, we can synthesize realistic video sequences of body sway for image-based avatars.

Algorithm 4 Synthesize new video sequence.

```
while true do
    select r_k randomly from set of reference times
    for t = r_k to r_{k+1} do
        display color image a_t
    end for
end while
```

Evaluation of synthesis of body sway

We evaluated the body sway of image-based avatars synthesized by our method. We used male and female avatars. We acquired video sequences of 90 s for each standing person, extracted eight reference times the video sequences and synthesized new sequences of 20 s video to generate the avatars. We set $\alpha = 0.5, \beta = 0.25, \gamma = 0.25, \lambda = 1.0 \times 10^{-5}, T_1 = 0.97$, and $T_2 = 1$. The parameters were experimentally determined to gain smoothness of synthesized videos. We conducted the subjective assessment on the following methods:

V1: (Ideal) directly replaying the first of 20 s of the acquired video sequences

V2: (Our method) replaying the synthesized video sequences

V3: (Baseline 1) replaying one short video sequences of 4 s randomly chopped from the acquired video sequences

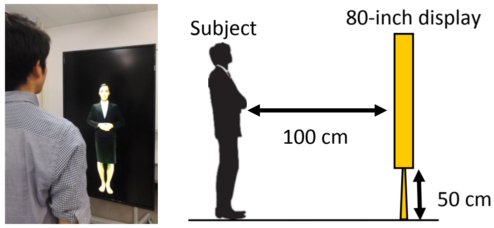


Figure 10. Setup for the subjective assessment.

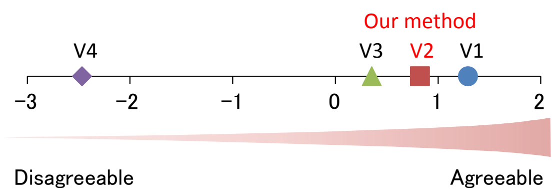


Figure 11. Results of the subjective assessment of synthesized body sway of avatars. We used the paired comparisons method.

V4: (Baseline 2) displaying a single shot from the acquired video sequences

Ten subjects (7 males, 3 females, aged 22.2 ± 1.1 years) participated in the subjective assessment using the paired comparisons method proposed by Thurstone [18]. The subject stood in front of a vertical 80-inch display at a distance of 1 m (Figure 10). We asked each subject to look at a pair of video sequences 6 $(= {}_4C_2)$ times, and select one video sequence from the pair that the subject felt more closely represented human-like movement.

Figure 11 shows the result of the subjective assessment. A higher subjective score indicates agreement from the subjects and vice versa. As we can see, our method (V2) scores higher than the baseline methods (V3, and V4). We took free comments from the subjects after the paired comparisons. One of the subjects said that V3 was unnatural because discontinuous movements appeared suddenly, and V4 was unnatural because the avatar completely stopped. Note that the ideal video sequence of V1 obtained a higher subjective score than V2. For this reason, we believe that V2 contained some small discontinuities at the transition points. Figures 12 shows examples of video sequences for the subjective assessment and a visualization of the movement amounts. We can see that the movement was large at the transition point in V3, the movement was small at the transition point in V2 as well as V1, and bodies of the avatars swayed to the left and right in V2. We believe that extracting reference times by our method is useful for increasing the reality of the image-based avatars.

Next, we compared the synthesized video sequences of our method with a video that replayed a single short video sequence of 2 s from one reference time to the next reference time. In the alternative video sequence, the same movements were repeated within a short time, although the movement was smooth at the transition point (Figure 13). We asked the subjects to select the video sequence that they felt more closely represented human-like movement. Figure 14 shows the comparison of the number of votes between our method and the alternative method. Our method obtained 18 votes and the alternative method obtained 2 votes. We believe that the choosing random transitions at reference times is effective for synthesizing human-like movements.

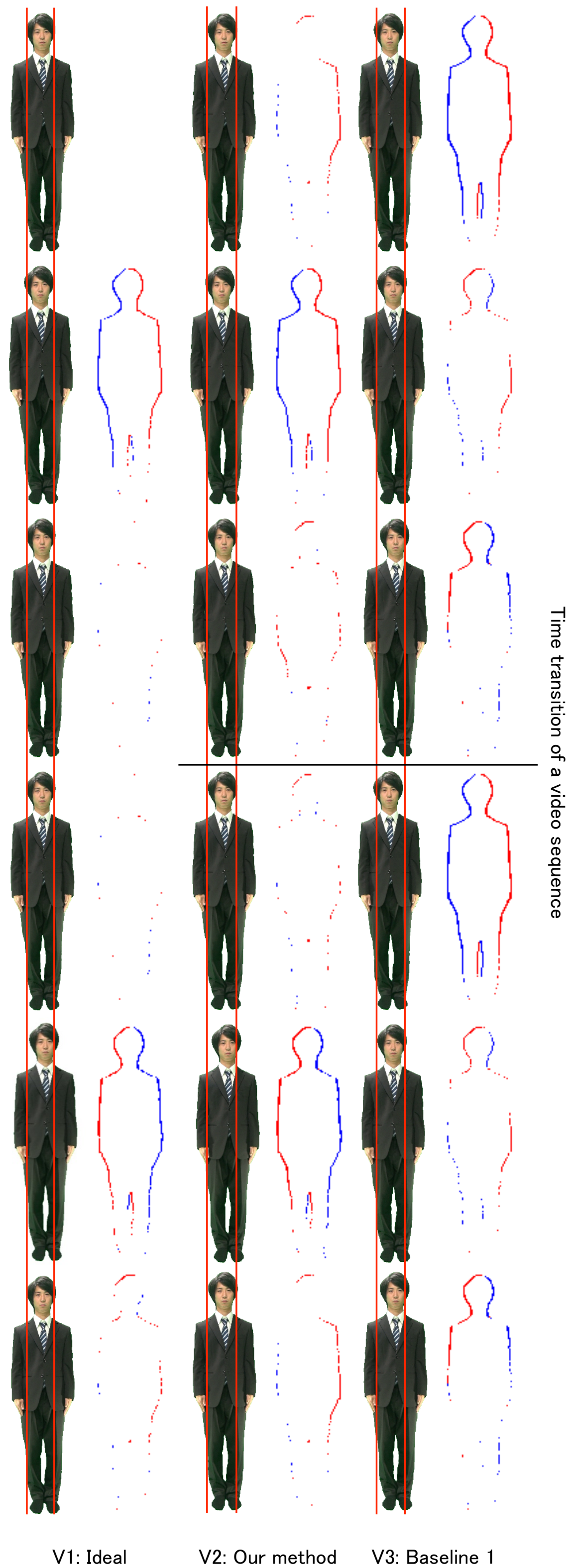
CONCLUSION

We proposed a method for analyzing and synthesizing body sway to generate human-like wait states of image-based avatars. By measuring the amount of movement for each body part, we observed that reference times of body sway repeatedly appear and all body parts continuously move between these reference times. We synthesized an arbitrary length time of body sway by randomly transitioning between the reference times. A subjective experiment demonstrated that our method is more agreeable than the baseline methods for subjective assessment.

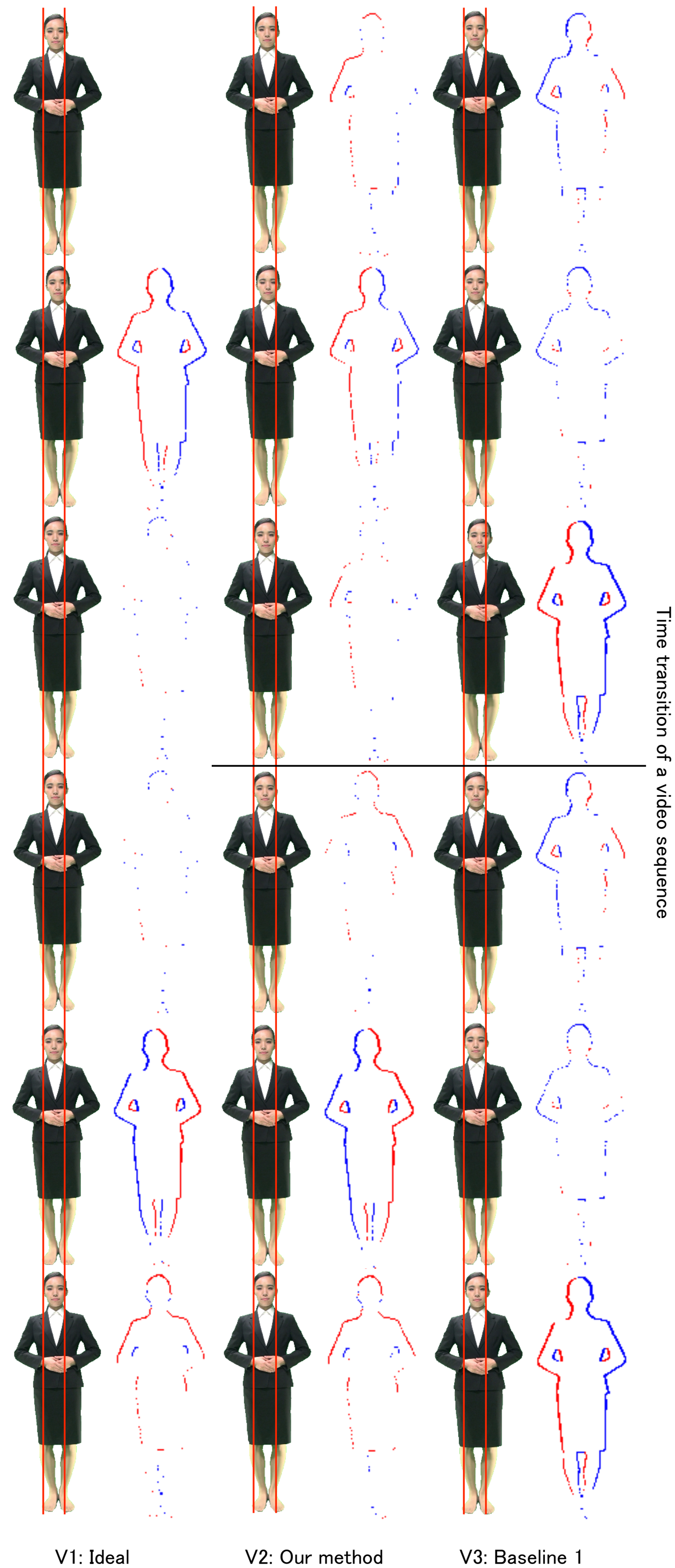
As part of our future work, we will expand our analysis on interactive systems and intend to consider the transitions between wait and action states.

REFERENCES

1. R. Anderson, B. Stenger, V. Wan, and R. Cipolla. 2013. Expressive Visual Text-to-Speech Using Active Appearance Models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3382–3389.
2. R. Artstein, D. Traum, O. Alexander, A. Leuski, A. Jones, K. Georgila, P. Debevec, W. Swartout, H. Maio, and S. Smith. 2014. Time-Offset Interaction with a Holocaust Survivor. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI)*. 163–168.
3. T.K. Capin, I.S. Pandzic, N.M. Thalmann, and D. Thalmann. 1997. Realistic Avatars and Autonomous Virtual Humans in VLNET Networked Virtual Environments. In *From desktop to webtop: virtual environments on the internet, WWW and networks, international conference*.
4. J. Cassell. 2000. Embodied Conversational Interface Agents. *Commun. ACM* 43, 4 (2000), 70–78.
5. M. Fabri, D.J. Moore, and D.J. Hobbs. 1999. The Emotional Avatar: Non-verbal Communication Between Inhabitants of Collaborative Virtual Environments. In *Gesture-Based Communication in Human-Computer Interaction: International GestureWorkshop (GW)*. 269–273.
6. J. Gratch, J. Rickel, E. Andre, J. Cassell, E. Petajan, and N. Badler. 2002. Creating interactive virtual humans: some assembly required. *IEEE Intelligent Systems* 17, 4 (2002), 54–63.
7. A. Hartholt, D. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L. P. Morency, and J. Gratch. 2013. All Together Now Introducing the Virtual Human Toolkit. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents (IVA)*. 368–381.



(a) Male avatar



(b) Female avatar

Figure 12. Comparison of synthesized video sequences of a male and a female avatars. Red represents a region converted from body to background and blue represents the opposite case.

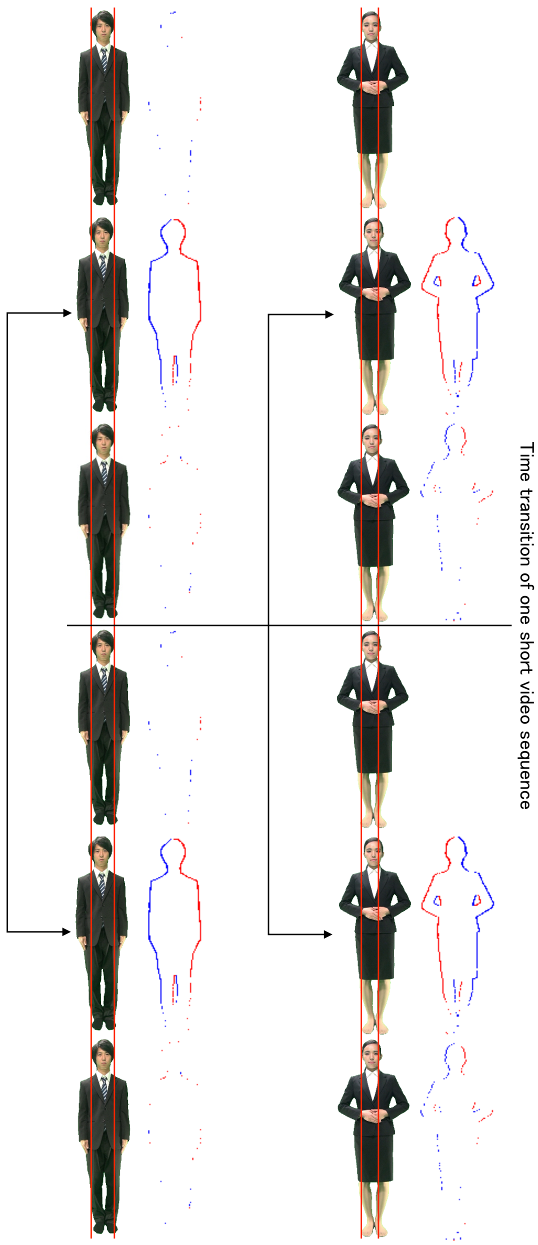


Figure 13. Video sequences using the alternative method. The same movements were repeated within a short time.

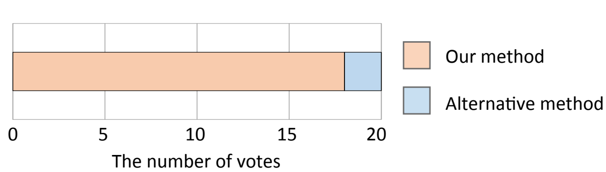


Figure 14. Comparison of the number of votes between our method and the alternative method in subjective assessment.

8. P. Huang, M. Tejera, J. Collomosse, and A. Hilton. 2015. Hybrid Skeletal-Surface Motion Graphs for Character Animation from 4D Performance Capture. *ACM Transactions on Graphics* 34, 2 (2015), 17:1–17:14.

9. A. Jones, J. Unger, K. Nagano, J. Busch, X. Yu, H. I. Peng, O. Alexander, M. Bolas, and P. Debevec. 2015. An Automultiscopic Projector Array for Interactive Digital Humans. In *ACM SIGGRAPH 2015 Emerging Technologies (SIGGRAPH)*. 6:1.

10. A. Lee, K. Oura, and K. Tokuda. 2013. MMDAgent - A Fully Open-Source Toolkit for Voice Interaction Systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 8382–8385.

11. L. S. Mark, J. S. Warm, and R. L. Huston. 1987. *Ergonomics and Human Factors: Recent Research*. Springer.

12. W. Mattheyses and W. Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication* 66 (2015), 182–217.

13. L. M. Nashner. 1972. Vestibular postural control model. *Kybernetik* 10, 2 (1972), 106–110.

14. D. Okwechime, E. Ong, A. Gilbert, and Richard Bowden. 2010. Social Interactive Human Video Synthesis. In *Proceedings of the 10th Asian Conference on Computer Vision (ACCV)*. 256–270.

15. S. Robinson, D. Traum, I. Ittycheriah, and J. Henderer. 2008. What would you Ask a conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. 28–30.

16. J. M. Saragih, S. Lucey, and J. F. Cohn. 2011. Deformable model fitting by regularized landmark mean-shift. *IEEE international Journal of Computer Vision* 91, 2 (2011), 200–215.

17. Q. Shi, S. Nobuhara, and T. Matsuyama. 2015. Augmented Motion History Volume for Spatiotemporal Editing of 3-D Video in Multiparty Interaction Scenes. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 1 (2015), 63–76.

18. L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review* 34, 4 (1927), 273–286.

19. F. Wang, M. Skubic, C. Abbott, and J. M. Keller. 2010. Body sway measurement for fall risk assessment using inexpensive webcams. In *Proceedings of International Conference of Engineering in Medicine and Biology Society (EMBC)*. 2225–2229.

20. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

21. L.F. Yeung, Kenneth C. Cheng, C.H. Fong, Winson C.C. Lee, and Kai-Yu Tong. 2014. Evaluation of the Microsoft Kinect as a clinical assessment tool of body sway. *Gait & Posture* 40, 4 (2014), 532–538.