Perceiving human motion variety

Martin Pražák*
Trinity College Dublin

Carol O'Sullivan†
Trinity College Dublin

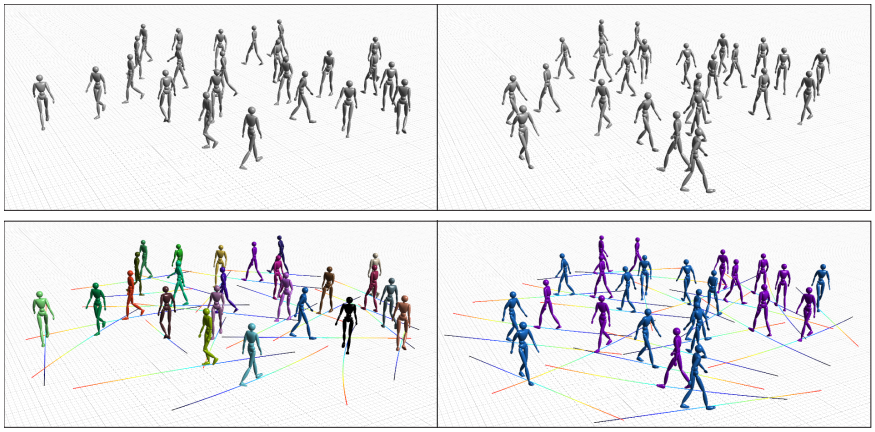

Figure 1: *The experiment stimuli and their creation – the stimuli as presented to the subject (top) and their creation, with trajectories and characters coloured according to their respective animations (bottom). In the depicted trial, the gold standard is shown on left.*

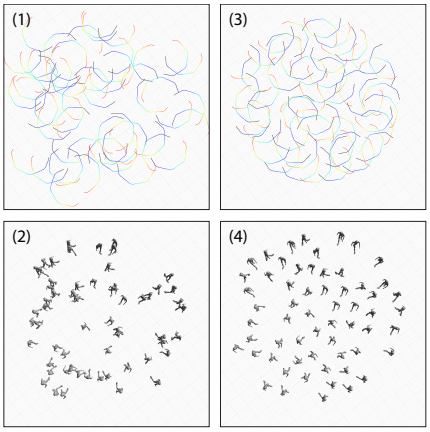

Figure 2: *Trajectory optimisation – random trajectories result in a large number of intersections (1), causing many collisions (2). Our optimisation approach ensures a collision-free scenario (3, 4).*

Abstract

In order to simulate plausible groups or crowds of virtual characters, it is important to ensure that the individuals in a crowd do not look, move, behave or sound identical to each other. Such obvious 'cloning' can be disconcerting and reduce the engagement of the viewer with an animated movie, virtual environment or game. In this paper, we focus in particular on the problem of motion cloning, i.e., where the motion from one person is used to animate more than one virtual character model. Using our database of motions captured from 83 actors (45M and 38F), we present an experimental framework for evaluating human motion, which allows both the *static* (e.g., skeletal structure) and *dynamic* aspects (e.g., walking style) of an animation to be controlled. This framework enables the creation of crowd scenarios using captured human motions, thereby generating simulations similar to those found in commercial games and movies, while allowing full control over the parameters that affect the perceived variety of the individual motions in a crowd. We use the framework to perform an experiment on the perception of characteristic walking motions in a crowd, and conclude that the minimum number of individual motions needed for a crowd to look varied could be as low as three. While the focus of this paper was on the dynamic aspects of animation, our framework is general enough to be used to explore a much wider range of factors that affect the perception of characteristic human motion.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation;

Keywords: crowd animation, animation variety, animation perception

1 Introduction

Crowds and groups of virtual characters are an important element of many games, virtual environments and movies. However, simulating large numbers of realistic virtual humans requires considerable resources, and optimisations are often needed when rendering, animating and simulating the behaviour of crowds. For example, significant savings can be made in terms of computational effort, memory footprint and artists' time by using a small number of human models or motions and replicating them many times to give the impression of a large crowd of people. However, this can result in the disturbing effect of a crowd full of identical individuals, or *clones*, an effect that has been recently explored by McDonnell et al. [2008]. In this paper, we build on this work and present a framework that enables a large range of scenarios to be created and configured to test the perception of human motion in groups and crowds. Using this framework, we conduct a perceptual study to find the number of individual motions that are needed to make the movements of a crowd of virtual humans look varied. Our results can be used to derive guidelines regarding the resources required to create, store and process captured motions. Although the focus in this paper is on the dynamic aspects of the animation (e.g., the way a character walks), our framework could also be used to study the static properties (e.g., skeletal structure variations) and the effects of a character's general appearance (e.g., 3D skinned model).

2 Background

McDonnell et al. [2008] explored the perception of variety in crowd simulations. They found that *appearance clones*, i.e, those that

*e-mail: prazakm@tcd.ie
†e-mail:carol.osullivan@scss.tcd.ie

use the same human body model, are easier to detect than *motion clones*, i.e., when the same animation is used for multiple characters. They also determined that the cloned motions of certain individuals were easier to spot than others, and that playing motions out-of-step made appearance clones more difficult to detect. However, one of the limiting factors of a crowd simulator is the total number of individual animations used as the basis of the motions for the full crowd, and their experiments were not designed to determine this factor in a realistic scenario. In this paper, we aim to determine the number of individual animations needed to create the illusion of a varied crowd of walking characters, in a realistic scenario similar to that seen in crowd applications.

The perception of human motion has been studied by many researchers. Johansson [1973] used a stimulus called a *pointlight walker*, which consisted of 10-12 light markers placed on the joints of a moving human subject. This allowed the motion of an individual to be represented independently of its appearance, while still being easily identifiable as a moving human. In fact, people are not only able to recognize simple actions from such pointlight displays [Dittrich 1993], but also specific gaits [Cutting and Kozlowski 1977], a walker's gender [Kozlowski and Cutting 1977], and simple emotions [Dittrich et al. 1996], provided that exposure and number of points used are sufficient [Barclay et al. 1978]. Therefore, it is clear that even a very reduced representation of a human body in motion can still convey a lot of identifying information about the actor.

The separation of the motion of a human from its appearance is therefore important in order to explore its properties. Barclay et al. [1978] distinguished between the *static* (i.e., posture and body shape) and *dynamic* (movement) properties of motion and showed that both properties affected the perceived sex of pointlight walkers. Troje [2002] took this process one step further by performing Principal Component Analysis on both the static and dynamic properties of human motion. His generic linear statistical framework for analysis and data-driven synthesis of pointlight walker stimuli can approximate any motion property. This distinction between the static and dynamic aspects of human motion is also inherent in models for data-driven computer animation. The static aspects of an individual's motion are represented by a skeleton structure (i.e., a hierarchy of rigid-body transformations, or 'bones') and its T-pose (i.e., the default pose of the skeleton). The dynamic aspects of the motion are its constituent keyframes (i.e., sets of rigid-body transformations, one for each bone, placed on a time scale describing the relative deformations of the skeleton's pose).

For a complete description of an animated human, a third property of *model type*, ranging from point-light and stick-figure to a highly realistic skinned model, can be defined. The model type has been shown to affect the perception of motion differences, which were easier to detect on a more realistic model [Hodgins et al. 1998]. A strong interaction between model, static and dynamic motion properties was also found when determining the perceived sex of a virtual human's motion [McDonnell et al. 2007; McDonnell et al. 2009]. While distinct in terms of computation, it is clear that these three properties are closely linked perceptually, and hence we pay particular attention to them in the design of our framework.

3 Crowd Animation Evaluation Framework

In this section we present our framework for evaluating the perception of crowd animation. The first step involves creating the model to be used to represent the appearance of the character. Raw captured motions can seldom be used without some preprocessing, so our second step is to process the animations while making minimal adjustments to the original motion. Finally, different crowd scenarios must be created, which involves the generation of natural and collision-free trajectories for different sized crowds of virtual characters. In all three steps, our aim is to make the framework as configurable and generic as possible.

3.1 Human Model Creation

We wished to create a parametric human model for our framework, so that it could be used to represent both the static and dynamic properties of motion, if required. However, in this paper we are interested in determining the number of characteristic motions needed to create an illusion of variety, and hence we also need a way to view the dynamic properties in isolation.

The static properties of the animation are fully represented by the skeleton and its T-pose, so we can use this information to construct a model to represent the appearance of the characters.

The simplest models are constructed by displaying the model's joints as dots (i.e., a point-light walker). The hierarchy of bones is then represented only implicitly by the dynamic information of the animation. Even though the human visual system (HVS) can easily derive the hierarchy from such a moving model [Johansson 1973], in practical applications in computer graphics, the skeleton is always represented explicitly. The simplest representation is to connect the dots by sticks, creating a stick-figure, while more advanced models make use of rigid-body segments and skinning techniques, thereby creating a realistic representation of the human body. Unfortunately, these representations are not perceptually equivalent, as it has been shown that motion properties are more easily perceived on realistic human models [McDonnell et al. 2007; Hodgins et al. 1998]. However, with more realistic models, confounding visual cues can be introduced, such as gender, skinning, bodyshape and texture. For that reason, we use a simple mannequin character composed of rigid segments, which has been shown to be gender-neutral while still retaining the body structure and therefore represents motion properties well [McDonnell et al. 2007]. We chose this model in order to focus on the motion properties of the animation, though for future experiments our framework could easily be extended to include a larger variety of appearance models.

We could use the same model for every motion, but in that case every animation has to be retargeted to fit the skeleton of the model, thereby potentially introducing motion artifacts. Whilst all retargeting methods aim to achieve minimal changes to the original motion, a parametric representation of the mannequin model would allow these changes to be reduced even further. Our animation framework therefore creates the mannequin model by building a simple 'box' model with parameters directly derived from bone lengths (Figure 3, middle). By applying several iterations of the Catmull-Clarke subdivision scheme and flattening several critical surfaces (soles of feet, palms), we create a fully parameterised mannequin model (Figure 3, right).

This parametric model of the character provides the required flexibility in terms of displaying both the static and dynamic properties of any motion in exactly the way they were captured.

However, in the experiment described in this paper, we are interested in examining the dynamic properties of the motion alone, and therefore wish to eliminate any static cues that could identify an individual's body shape. Therefore, the skeletons have to be identical and to minimize retargeting errors for a set of animations, the target skeleton should then be the average of the source skeletons. Assuming an identical pose of all source skeletons (which is true for all motion capture systems), the only differences are in the bone lengths (i.e., the translational parts of the rigid body transformations used to describe the skeleton). To create an average skeleton for our experiment, we can therefore simply average the corresponding bone lengths.

3.2 Animation Preprocessing

In order to use captured animations in a real application, a certain amount of preprocessing is inevitable. In our case, we need to ensure that a character follows a specific trajectory for a given length of time, while its motion is displayed using the mannequin model described above. Our aim was to change the animation data as little as possible, while allowing the required level of experimental control.

A simple representation of an animation performed on a flat surface is its *trajectory*. This is also the usual representation for solving the collision avoidance problem (see Section 3.3). The simplest derivation approach is to project the root bone onto the ground plane, creating a rough approximation of centre-of-mass movement as projected on the ground (Figure 4). By embedding the skeleton with a new root bone (*trajectory bone*), setting the original root as its only parent and using the Y rotation and X-Z transformation of the keyframe data (assuming Y as the "up" vector), we create a new animation that corresponds to the original animation but explicitly represents the trajectory information.

As animations for use in a crowd application or as perceptual stimuli usually need to be much longer than what can be obtained as a continuous sequence using motion capture technology (due to limited capture space), it is often necessary to extend the original data. If the source animation is an intrinsically *periodic animation* (e.g., locomotion), we can lengthen it by repeating the periodic part. First, by embedding the animation with a trajectory bone and setting all the trajectory-bone transformations to identity, we obtain an on-the-spot version of the original animation. Using the sum of squared distances between corresponding bones on this data, we can create a simple metric to compare pairs of frames [Kovar et al. 2002a]. Describing all comparisons as a matrix, where each row and column represent the indices of the compared frames, we obtain a symmetrical matrix with zeros on the diagonal. The minimal value sufficiently distant from the diagonal then represents the frame indices of the beginning and the end of the period (Figure 5). After cropping the animation to one period, we can represent the differences between the first and last frame as a set of transformations, one for each bone. Motion blending [Kovar et al. 2002a] is then used to blend-in these differences, thereby creating an animation with the same first and last frame. By simply concatenating the resulting animation, we can create a motion of arbitrary length, while retaining the properties of the original.

We complete the motion preprocessing step by adjusting the animation to fit the average skeleton, according to the adaptation described by Kovar et al. [2002b]. As we choose skeletons with minimal bone-length differences for our experiment (Section 4.1), we do not require the lengthening step for any of our source animations, though a greater range of body shapes would necessitate this step in future experiments.

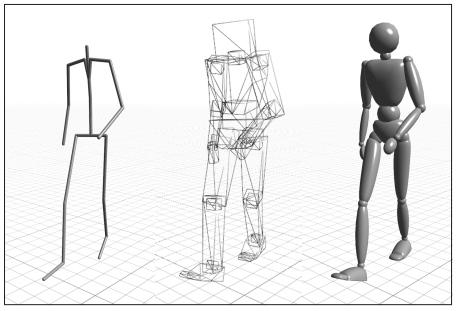

Figure 3: *Mannequin character building steps. A stick-figure skeleton (left) is used as a starting-point to generate a parametric box-figure (middle), which is then subdivided using Catmull-Clarke subdivision to provide a smooth mannequin character (right).*

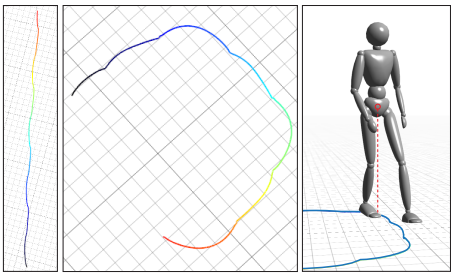

Figure 4: *The trajectory reconstruction and trajectory-bone embedding. The 2D trajectory (left, middle) is extracted by projecting the root-bone on the ground (right) and creating a new parent bone following this trajectory.*

3.3 Scenario Generation

Ideally, we would like to create scenarios in which animated characters are placed randomly and are given random trajectories, with a different scenario generated for every trial. However, without any further processing, the generated trajectories would often intersect in the corresponding animation frames, which would lead to collisions between characters (Figure 2, left). In order to achieve maximal randomness while avoiding disconcerting character collisions, we introduce a simple iterative least-squares optimisation scheme based on rigid-body fitting. First, we place the animations into random positions and directions in the desired area. With each trajectory represented as a set of 2D vectors $\mathbf{v}_i(t)$ (where i is the animation's index and t is the index of the keyframe), we can define a vector $\mathbf{d}_{a,b}(t)$ describing the difference between two corresponding points on the trajectories a and b as $\mathbf{d}_{a,b}(t) = \mathbf{v}_b(t) - \mathbf{v}_a(t)$. We can also create similar displacement vectors for enforcing the trajectories into a certain area. The displacement force vector $\mathbf{f}_{a,b}(t)$ is then created from vector $\mathbf{d}_{a,b}(t)$ as

$$\mathbf{f}_{a,b}(t) = \frac{\mathbf{d}_{a,b}(t)}{\|\mathbf{d}_{a,b}(t)\|} \left(max\left(0, f_{min} - \|\mathbf{d}_{a,b}(t)\|\right)\right)^2$$

where f_{min} is the minimal enforced distance. This equation describes a simple repulsion force of a particle system (f_{min} was set to 1.5 in all our experimental scenarios described below). By adding the force vector $\mathbf{f}_{a,b}(t)$ to each point $\mathbf{v}_a(t)$ of the trajectory a, we create a deformed version of the trajectory data. Subsequently, we can fit the original trajectory data into the deformed trajectory using rigid-body least squares fit, thereby creating a displaced version of the original trajectory. By iteratively applying this process, we minimize the sum of all norms of vectors $\mathbf{f}_{a,b}(t)$, leading to a set of collision-free trajectories (Figure 2, right).

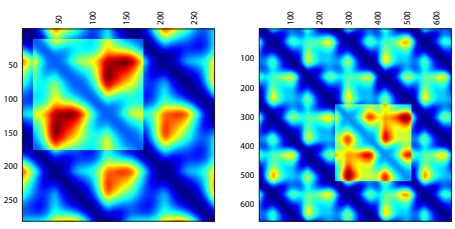

Figure 5: *A visualisation of comparison matrix between each pair of frames of an animation and the detection of the animation's period (shown as the lighter rectangle).*

4 Experiment Design

We used the framework described in Section 3 to determine the minimum number of individual motions required to create the impression of a varied crowd of walking humans. We also explored how this number was affected by the size of the crowd and the speed of motion. The stimuli consisted of short animations, each depicting a 'crowd' of walking mannequins. We hypothesised that using **a higher number of individual animations will make the crowd look more varied**. We therefore tested four values of factor *motion consistency*, where 100% motion consistency meant that all characters in the scene used the same animation, and the lowest value meant that four animations were each used to animate 25% of the crowd. We also tested whether **the size of the crowd will affect the perception of cloned animations**, as more characters on the screen could either mask their motion similarities or else provide more examples to aid in spotting the animation clones. Therefore, we tested the factor *crowd size* with values low (8 or 9 characters), medium (15 or 16 characters) or high (24 characters). The +/- 1 is to ensure that the number of characters is divisible by the total number of animation used in the stimulus, in order to ensure an even distribution of the number of simultaneously displayed animations. Finally, we tested whether **the speed of motion will influence the perceived variety of the crowd**, i.e., whether the personal characteristics of the actors have different effects on the perception of motion differences depending on their *speed* (slow, with one period of $1.32 \pm 0.03s$, normal, with one period of $1.06 \pm 0.01s$ or fast, with one period of $0.90 \pm 0.01s$).

4.1 Motion Stimuli

As the input data for our experiment, we used straight walks from our motion capture database (currently containing motions from 83 actors (45M and 38F), with ages ranging from 14 to 50, weights from 41 to 102 kg and heights from 1.53 to 1.96 m).

In order to avoid any motions with artifacts caused by an outlying body build, we selected the 24 required actors for each gender by performing Principal Component Analysis on the weight and height data (separately for each gender), constructed an elliptical Gaussian curve around the mean point, with directions determined by the recovered eigenvectors and standard deviations proportional to the eigenvalues, and selected the 24 actors projecting to the highest values on this curve (see Figure 6). This procedure provided a subset with sufficient variance, whilst avoiding any weight/height outliers.

For the input motion data, we used 3 straight walking motions from each actor (according to the *speed* factor). The motions were preprocessed by extracting one period (see Section 3.2), and looping to make the total length of each clip $5 + \alpha$ seconds, where α was the length of one period of the motion. By selecting the start of each motion randomly from the interval $[0..\alpha)$, we achieve full desynchronisation of the displayed motions.

4.2 Experiment

The stimuli were presented on two separate 19-inch LCD screens using their native resolution of 1280x1024 pixels and refresh rate of 60Hz. On one screen, an animation was displayed where every member of the crowd was animated using an individual motion (i.e., the gold standard). On the other screen, the same animation was shown, but with cloned motions based on the *motion consistency* factor. We counterbalanced the *position* of the gold standard across all trials, and randomised the order of trials for each participant. We also tested whether the *gender* of the actors influenced the perception of variety in the motions and therefore showed only all male or all female motions simultaneously in each of the scenarios, also counterbalanced. The scene data were generated before each

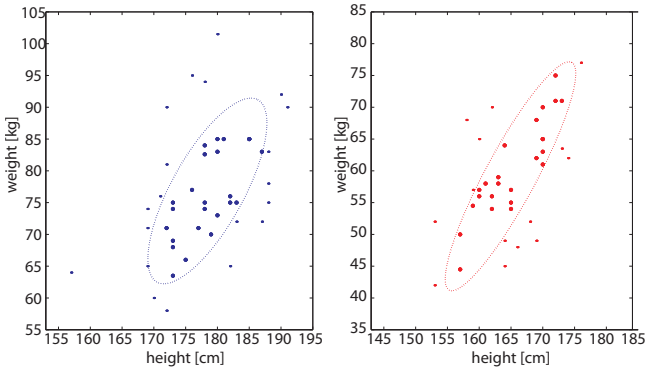

Figure 6: *The selection of motion clips based on actors' body shapes (left – males, right – females). The non-standard body builds were rejected (outside the ellipse), while accounting for different height and weight trends (a similar approach is used for computing BMI, and our selected subset would then correspond to participants closest to 'average').*

trial, with the placement of trajectories first randomised and subsequently optimised to avoid any collisions, as described in Section 3.3. The camera view was centered in the middle of the scene and oriented at an angle of 20^o above the ground plane with the view angle set to 35^o vertically to encompass the full scene, which was always 6m in diameter.

The scene was rendered using the OpenGL rendering system with vertical synchronisation turned on to avoid any possible flickering artifacts. The experiment itself consisted of 144 trials: 4 (*motion consistency*) $\times$ 3 (*crowd size*) $\times$ 3 (*speed*) $\times$ 2 (counterbalanced for actor *gender*) $\times$ 2 (counterbalanced for gold standard *position*). Each trial lasted 10 seconds and consisted of two repetitions of the 5 second animations. The borders of both screens started blinking 2.5 seconds before the end of each trial (at 0.5 second intervals), to indicate the end of the time limit. Following each trial, there was a 2 second delay before the next display of stimuli, during which the remaining number of trials was displayed at the centre of each screen.

Twelve volunteers (7M, 5F) from students and staff of our university participated in our experiment. All participants were naive to the purpose of the experiment and had normal or corrected to normal vision. They were asked to indicate *on which screen was every character walking differently* by pressing one of two clearly indicated buttons on the keyboard. We recorded both the accuracy of their answer and their response time. If the participant did not answer in the allocated time limit, the progression of trials was paused until the answer was provided. The total duration of the experiment was about 30 minutes per participant. After the experiment was finished, participants were asked to fill a simple questionnaire, aimed at determining the discrimination method they adopted.

5 Results

We analysed both the percentage of correct answers and the response times recorded for all participants – see Figure 7. As expected, we found no effect of either the gender of the actors or the position of the gold standard, so we averaged across these factors. We then performed a repeated measures ANalysis Of VAriance (ANOVA) on the three factors: *motion consistency*, *crowd size* and *speed*.

For answer accuracy, we found a main effect of *motion consistency* ($F_{3,33} = 7.6517, p < 0.0006$) and a three-way interaction between all three factors ($F_{12,132} = 2.1858, p < 0.02$). Post-hoc analy-

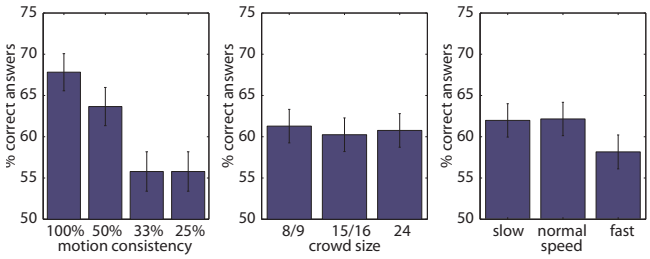

Figure 7: *Results from the experiment on the perception of crowd variety*

sis of the main effect using Neuman-Keuls comparison of means showed that there were two distinct groupings of values, with no significant difference within these sets. Motion consistencies of 100% and 50%, i.e., where either 100% or 50% of the characters were animated using the same cloned motion, were significantly easier to detect than the more varied scenarios.

Neuman-Keuls post-hoc analysis of the three way interaction did not provide any more information (no combination of factors was significantly different), and a further study with a larger number of participants would be required to explain this effect.

Trials lasted 10 seconds with a warning sign shown at 7.5 seconds, and the average response time was 7.548s with standard error of 0.054s. The ANOVA on response times showed a main effect of *crowd size* ($F_{2,22} = 4.3230$, $p = 0.03$). Neuman-Keuls post-hoc comparisons showed response times were significantly slower when the largest number of characters was displayed. The difference of means was, however, only 0.27s.

In an informal questionnaire presented to each participant after finishing the experiment, participants reported that the discrimination method they adopted during the experiment was based on searching for pairs of similar motions. If they succeeded in finding such a pair on either of the screen, they reported the other as the correct answer. Failing that, they resolved to "general feeling" of the stimulus and selected the one seemingly more varied.

Our results strongly suggest that the dominant factor in the perception of crowd variety is the number of cloned motions present in the scenario. The factors of crowd size and motion speed have, at best, only a weak and unpredictable effect on the perception of motion variety in a crowd. We found that the maximum number of individual motions successfully detectable was 2, i.e., when 50% of the crowd was animated using one motion, and the other 50% also all animated using a second motion.

6 Conclusions and Future Work

The main guideline we can provide from this study is that, to preserve the perception of motion variety in a typical pedestrian crowd, there is no need for more than 3 different characteristic animations per gender of the displayed characters (whereas 2 motions are not enough). By setting a 10s limit for the answer, our results only indicate the participants' immediate impression of the crowd motion variety. These results are therefore probably most valid for a classical game-like scenario, but may not hold under careful scrutiny of the animations. In future work, we would like to investigate this aspect of animation perception. As our crowd animation evaluation framework is very general and configurable, we are planning to use it for exploring other factors that can affect motion perception, e.g., camera angles, animation types, animation artifacts, the effect of different types of appearance models and other effects.

Acknowledgments

We wish to thank the reviewers for their comments, and the participants in our experiments. This work was sponsored by Science Foundation Ireland as part of NaturalMovers project.

References

BARCLAY, C. D., CUTTING, J. E., AND KOZLOWSKI, L. T. 1978. Temporal and spatial factors in gait perception that influence gender recognition. *Perception And Psychophysics 23*, 2, 145–152.

CUTTING, J. E., AND KOZLOWSKI, L. T. 1977. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society 9*, 5, 353–356.

DITTRICH, W. H., TROSCIANKO, T., LEA, S. E., AND MORGAN, D., 1996. Perception of emotion from dynamic point-light displays represented in dance.

DITTRICH, W. H. 1993. Action categories and the perception of biological motion. *Perception 22*, 1, 15–22.

HODGINS, J. K., O'BRIEN, J. F., AND TUMBLIN, J. 1998. Perception of human motion with different geometric models. *IEEE Transactions on Visualization and Computer Graphics 4*, 4, 307–316.

JOHANSSON, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception And Psychophysics 14*, 2, 201–211.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002. Motion graphs. *ACM Transactions on Graphics 21*, 3.

KOVAR, L., SCHREINER, J., AND GLEICHER, M. 2002. Footskate cleanup for motion capture editing. *Proceedings of the 2002 ACM SIGGRAPHEurographics symposium on Computer animation SCA 02*, i, 97.

KOZLOWSKI, L. T., AND CUTTING, J. E. 1977. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics 21*, 6, 575–580.

MCDONNELL, R., JÖRG, S., HODGINS, J. K., NEWELL, F., AND O'SULLIVAN, C. 2007. Virtual shapers & movers: form and motion affect sex perception. *Applied Perception in Graphics and Visualization Vol 253*.

MCDONNELL, R., LARKIN, M., DOBBYN, S., COLLINS, S., AND O'SULLIVAN, C. 2008. Clone attack! Perception of crowd variety. *ACM Transactions on Graphics 27*, 3, 1.

MCDONNELL, R., JÖRG, S., HODGINS, J. K., AND NEWELL, F. N. 2009. Evaluating the effect of motion and body shape on the perceived sex of virtual characters. *ACM Transactions on Applied Perception 5*, 4, 1–14.

TROJE, N. F. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait. *Journal of Vision 2*, 10, 371–387.