Artificial Life Manuscript Submission

Crowd-Sourced Identification of Characteristics of Collective Human Motion

Martyn Amos [1] (@martyn_amos), Jamie Webster [1],

Corresponding: Martyn Amos (martyn.amos@northumbria.ac.uk)

1. Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom.

Abstract. Crowd simulations are used extensively to study the dynamics of human collectives. Such studies are underpinned by specific movement models, which encode rules and assumptions about how people navigate a space and handle interactions with others. These models often give rise to macroscopic simulated crowd behaviours that are statistically valid, but which lack the noisy microscopic behaviours that are the signature of believable real crowds. In this paper, we use an existing "Turing test" for crowds to identify realistic features of real crowds that are generally omitted from simulation models. Our previous study using this test established that untrained individuals have difficulty in classifying movies of crowds as real or simulated, and that such people often have an idealised view of how crowds move. In this follow-up study (with new participants) we perform a second trial, which now includes a training phase (showing participants movies of real crowds). We find that classification performance significantly improves after training, confirming the *existence* of features that allow participants to identify real crowds. High-performing individuals are able to *identify* the features of real crowds that should be incorporated into future simulations if they are to be considered realistic.

Keywords: crowds, simulation, realism, agents, Turing test

1 Introduction

A significant amount of artificial life research is concerned with studying the collective dynamics of *mobile agents* operating in a spatially-explicit environment. Relevant domains include the flocking behaviour of birds and other "animats" ("boids" being the archetypal example (Reynolds, 1987)), the power of distributed swarm robotics (Brambilla et al., 2013), and the engineering of biological cell populations (Gorochowski, 2016). In all such cases, agents (whether simulated or physically realised) are situated in Cartesian space, and may interact both with one another and with their environment.

One specific area of growing interest is the study of *crowd dynamics* (Adrian et al., 2019); that is, the behaviour of large numbers of human individuals moving through and interacting in a given environment. The need to understand collective human behaviour in physical space is pressing, as it has significant implications for events planning and management (Crociani et al., 2016), urban design (Feng et al., 2016), and incident response and analysis (Harding et al., 2011; Pretorius et al., 2015). During and after the COVID pandemic, with potentially long-lasting and profound structural and behavioural changes being made, the need to understand the crowd will persist (Pouw et al., 2020).

Due to the inherent difficulty of performing large-scale experiments with human participants, *crowd simulations* (Thalmann & Musse, 2013) (usually using an agent-based approach) are often used to investigate collective behaviour and the impact of physical or behavioural interventions on crowd dynamics. Two features of simulations are of interest; *validity* and *believability*. Validity describes how closely the output of the model matches data obtained from the real world (Klüpfel, 2007; Pettré et al., 2009; Seer et al., 2014). *Believability* is subtly different, and concerns the human perception of whether or not a crowd's behaviour is *realistic*, or plausible. We are not concerned with "cinematic", photo-realistic believability of the rendering of a crowd, but whether or not observers are able to detect characteristic *patterns of behaviour* in real crowds which are absent in simulated crowds. Fundamentally, we assume that a simulation is valid, and are interested in whether or not it also *looks realistic*.

The rest of the paper is organised as follows; we give some background motivation, outline our hypothesis, and describe our crowd Turing test framework for its investigation. We then describe our experimental method for the current study, and describe our results. We conclude with a discussion of the implications of our findings, and suggest possible future work.

2 Background and Motivation

Crowd simulations are now used extensively in a wide range of application domains, from urban planning (Aschwanden et al., 2011), emergency response (Mahmood et al., 2017), games and training simulations (Mckenzie et al., 2008), and the CGI generation of Hollywood movie scenes (a classic example being the large-scale battle scenes in *The Lord of the Rings* series) (Ricks, 2013). Most crowd simulations are underpinned by a behavioural/movement model, which makes simplifying assumptions about individuals, and which is used by agents to determine their trajectories through the simulated space.

The Social Forces Model (SFM) (Helbing & Molnar, 1995) lies at the heart of many scientific and commercial crowd simulation packages, such as FDS+EVAC (Korhonen et al., 2010), PedSim (Gloor, 2016), SimWalk (Kimura et al., 2003) and MassMotion (Rivers et al., 2014). However, there are well-established deficiencies in this and other existing movement models. As (Lerner et al., 2007) argue, “While such approaches may capture the broad overall behaviour of the crowd, they often miss the subtle details displayed by the individuals. The range of individual behaviours that may be observed in a real crowd is typically too complex for a simple behavioural model… Simple things such as walking in pairs, stopping to talk to someone, changing one’s mind and heading off in a different direction or aimlessly wandering about, are just a few examples which are difficult to capture.” The emphasis here is less on the locomotion model of avatars or the cosmetic appearance of the agents, and more on the *patterns* and “quirks” of movement that distinguish a real crowd from a simulated one.

Why is this important? After all, emergency planners (to take one significant user group) will generally be satisfied if the overall outcome of a simulation (in terms of the time required to evacuate a stadium, for example) is broadly valid, and will usually not concern themselves with micro-level “turbulence” and other localised phenomena. However, as (Fuchsberger et al., 2017) argue, crowd simulations still meet with resistance from decision makers in some significant industrial and societal domains, and this may be due to a lack of trust in their outputs (caused, in turn, by a lack of realism). Specific concerns identified of relevance to the current paper include “unnatural motion paths”, so if we can go some way towards addressing this, then it may lead to increased acceptance and uptake of these techniques.

As we argue in (Webster & Amos, 2020), there is still a need for more realistic behavioural/movement models in crowd simulation, and “This is motivated by a widely-acknowledged need for crowd simulations to include more realistic features derived from individual and social psychology (such as group-level behaviours, indecision, etc.) (Lemercier & Auberlet, 2016; Seitz et al., 2017; Templeton et al., 2015), which are generally not included in software packages, and which give rise to rather unrealistic or “robotic” patterns of behaviour at the population level”.

Much work has already been done on making crowd simulations more realistic; here we highlight some representative contributions. (Lerner et al., 2007) describes the construction of a database of behavioural “motifs” which may be incorporated into an agent’s behaviour. (Peters & Ennis, 2009) used manual annotation of observations to extract information about group-level behaviours that were then incorporated into simulations (this study also included human trials of perception of realism). More recently, (Wei et al., 2018;

Yao et al., 2020) used machine learning to extract features of observed crowds, which were then incorporated into a crowd simulation, but neither study assessed whether or not these modifications actually made the overall crowd behaviour more realistic.

Fundamentally, what passes for realistic is inherently subjective. To our knowledge, until we performed this study no extensive work had been done on capturing the “essence” of what makes a crowd realistic *from the perspective of human observers*.

Our previous work (Webster & Amos, 2020) showed that crowd simulations that employ the most commonly-used movement model are valid (in terms of their outputs having the same statistical properties as observed crowds), but they still possess a “signature” that allows them to be distinguished from real crowds. Simply put, to human observers, simulated crowds are still perceived differently to real crowds. Importantly, though, we also found that although people are able to reliably *partition* crowds into real/simulated, *they are unable to tell which is which*. That is, individuals are able to separate crowd movies into two categories, but they are unable to reliably label the real crowds. We found that individuals tend to have an idealised view of the behaviour of real crowds, which is often at odds with reality. These findings confirm the observation that real and simulated crowds have different microscopic features that allow them to be partitioned, if not classified.

To summarise, our previous work established the *existence* of features that are present in real crowds but not in simulated crowds; the aim of the current paper is to *identify* those features. In (Webster & Amos, 2020) we argue that “Our results suggest a possible framework for establishing a minimal set of collective behaviours that should be integrated into the next generation of crowd simulation models.” Here, we use the “Turing test” classification task to identify that specific set of features that allow trained viewers to reliably *classify* (not just partition) real and simulated crowds. Our results show that classification performance over a population of observers increases significantly after an initial training phase, and that individuals are able to identify a core set of realistic behaviours that are present in real crowds, but which are absent in simulated crowds. This immediately suggests new features that must be incorporated into future crowd simulations if they are to be considered realistic.

3 Hypothesis

In a landmark paper (Turing, 1950), Alan Turing proposed a method to investigate what would become known as "artificial intelligence". Rather than directly answering the somewhat ambiguous question "Can machines think?", Turing preferred to reframe the issue in terms of an "imitation game", in which an interrogator engaged in conversation with two agents via "teletypes". One of the agents (A) is a man, and the other (B) a woman, and the interrogator's objective is to decide which is which by asking questions of both and assessing their responses. The task of A is to cause the interrogator to guess *incorrectly* (that is, persuade them that he is a woman), and the task of B is to "help" the interrogator to guess correctly, generally by giving truthful answers. We may, therefore, interpret the imitation game (commonly referred to as the "Turing test") more generally, with the role of A being played by an artificial system that seeks to persuade a human observer that it is the "genuine article", and B being played by an actual "real world" example of the system under study. Importantly, the test does not seek to establish the "truth" of A's outputs (that is, their validity), but simply whether or not A could be said to represent a reasonable facsimile of the system represented by B.

This conceptual framework has been proposed for biological modelling (Harel, 2005) and artificial life (Cronin et al., 2006) as a way of investigating the realistic properties of artificial systems. We previously used the same approach to investigate crowd simulations, basing our approach on a related Turing test for collective motion in fish (Herbert-Read et al., 2015). In (Webster & Amos, 2020), we describe the results of initial experiments, using a total of 540 in-person participants. The first set of trials presented individuals with a sequence of paired movies, using a side-by-side representation. In each pair, one of the movies represented the movement of a real crowd, and the other represented a computer simulation of the same scenario (the ordering was randomised). All observations were of the same physical space, and both movies were generated using the same custom rendering engine. For each pair (over six pairs in total), participants were asked to specify which of the pair they thought was the real crowd (that is, they had to *identify* the real crowd). For the second set of trials, participants were presented with the movies individually, and this time they were asked to *classify* each movie as either real or simulated.

We found that participants performed better when they were asked to *classify* crowds rather than having to choose between the two, but a striking feature of our results was that neither mode allowed participants to perform better than random guessing. A simplistic interpretation of this result could be that existing simulations are good enough to "pass" the crowd Turing test, as human observers are unable to distinguish between them, but here we emphasise that the imitation game, as originally described by Turing, requires the interrogator to be able to specify *which* agent is the man.

Strikingly, the most common score in the first trial was zero, meaning that a significant proportion of participants (36.46%) failed to identify a single real crowd. That is, their entire perception of what constitutes a real crowd was perfectly "flipped" compared to reality. This sizeable group of participants were able to perfectly partition movies into real or simulated, but were utterly unable to say which was which. This confirmed the existence of a set of real crowd behaviours (informally described by participants in terms of "standing around" and "moving with purpose") that allowed individuals to separate real from simulated, but which were incorrectly ascribed to the simulation as generating "unrealistic" crowd behaviour. Our conclusion was that participants had an idealised view of real crowd behaviour, and preferred to think that it was much less "messy" and unpredictable than observations would suggest.

Our hypothesis, therefore, is that participants in a crowd Turing test will improve their classification performance after being trained by viewing real crowds, as a result of being able to identify and ascribe *only to real crowds* the realistic features that are manifested in the training set.

4 Experimental Methods

Our protocol was largely modelled on that of (Webster & Amos, 2020), but limitations imposed by the COVID pandemic required us to perform our trials online, as opposed to in-person. We do not believe that this modification had any significant impact on our results; indeed, it actually allowed us to recruit a more diverse range of participants, rather than using only University students (which was a possible criticism of the original study).

We performed two sets of Turing test experiments; the first (Test 1) was an online-only repetition of the second (classification) test from (Webster & Amos, 2020), with entirely new participants. We attracted 232 participants, who were recruited via social media. This first test allowed us to assess the ability of each untrained participant to classify crowds as either real or simulated, thus assigning each one a baseline score. We allowed an appropriate period of time to pass (4 months) in order to ensure that the tests were independent (that is, any learning effects from the first test would not be carried over to the second). We then contacted every Test 1 participant who supplied an email address to invite them to participate in the follow-up Test 2 (they were each offered a 10 GBP gift card as an incentive); 50 participants accepted our invitation. Test 2 participants were then "trained" by asking them to first watch six rendered movies of crowds that were explicitly described as real. Participants then performed a second version of the classification task (as in Test 1), using a different set of real and simulated clips to those used previously (in order to avoid effects induced by familiarity with the clips).

Given that each participant had a known baseline score from Test 1, we were able to es-

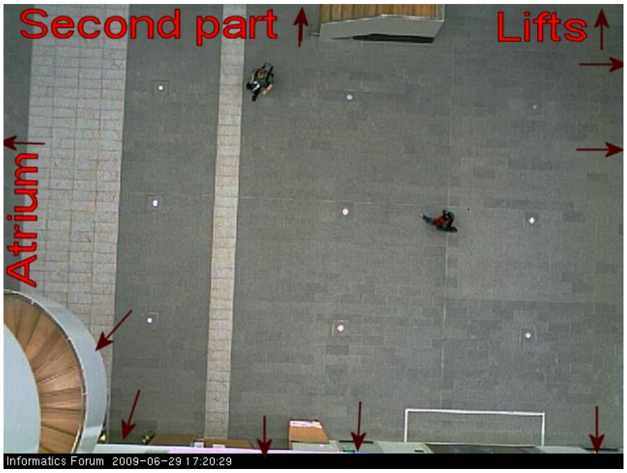

Figure 1: Single movie frame of the Edinburgh Informatics Forum, taken from (Majecka, 2009).

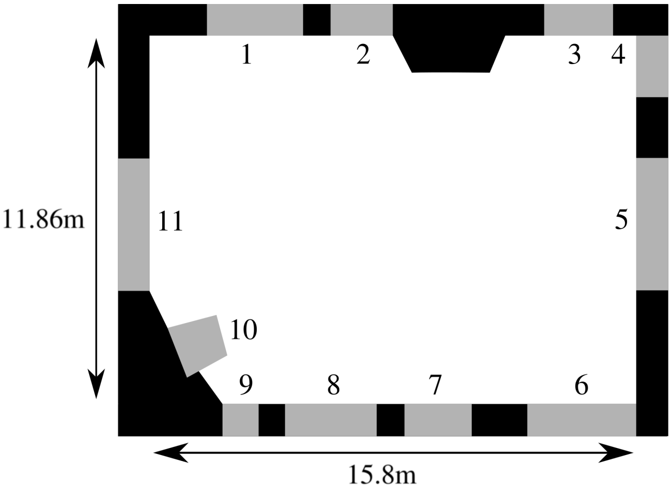

Figure 2: Diagram of Edinburgh Informatics Forum (ingress and egress points numbered), taken from (Webster & Amos, 2020)

tablish whether or not the training phase had a significant effect on classification ability. Participants were specifically asked to identify features that they thought allowed them to distinguish between real and simulated crowds.

Test 1 was performed at the end of June-start of July 2020, and Test 2 was performed in December 2020. Our trial protocol was approved by the Northumbria University Faculty of Engineering and Environment Ethics Committe, application number 24623. We now describe each component of the trial in more detail.

4.1 Pedestrian motion dataset

As we employed the same dataset used in our previous study, we take our description of it from (Webster & Amos, 2020). We used data on real pedestrians from the University of Edinburgh School of Informatics (Majecka, 2009). This public dataset, captured in 2010, contains over 299,000 individual trajectories corresponding to the movement of individuals through the School Forum, and is one of the largest open datasets of its type. It has been used in several studies of pedestrian movement and tracking; (Fernando et al., 2018) used the dataset to pre-train short and long term trajectory prediction models, proposing a "light-weight" sequential Generative Adversarial Network (GAN) architecture for person localisation, which "overcomes issues related to occlusions and noisy detections". In a case study on the Edinburgh Informatics forum, (Lovreglio et al., 2017) developed a "microscopic calibration procedure" for floor field cellular automaton models, comparing two floor field specifications to identify the best model for simulating pedestrians in the forum. However, this study was only concerned with individual trajectories, and did not consider the crowds as a collective. Finally, recurring activity patterns that "appear, peak, wane and disappear over time" were identified using non-parametric Bayesian methods which couple spatial and temporal patterns with "minimal" prior knowledge (H. Wang & O'Sullivan, 2016).

4.1.1 Environment

A photo of the Forum space is shown in Figure 1, and a diagram is shown in Figure 2. The Forum is rectangular in shape (measuring approximately 15.8 $\times$ 11.86 metres), has eleven ingress/egress points, and is generally clear of obstructions. Images were captured (9 per second) by a camera suspended 23m above the Forum floor, from which individual trajectories were extracted and made available (extraction was performed by the author of (Majecka, 2009)). We note that only the *trajectories* have been made publically available, and not the original video recordings, for ethical and practical reasons (these files require several terabytes of storage). Importantly, none of the individuals whose trajectories were captured were actively participating in movement studies; the trajectories, therefore, are as close to “natural” as possible (i.e., they have “behavioural ecological validity” (Lovreglio et al., 2017)).

4.1.2 Pedestrian dataset

The dataset is stored across a number of files, each file representing a day’s worth of crowd recordings. Each file stores a list of “sightings” over that period, where a sighting is defined as an individual entering (but not necessarily leaving) the frame (of course, individuals may also leave and then re-enter the frame, which would be interpreted as an entirely new sighting). Each row in the file therefore corresponds to a “sighting”. Every sighting during the time period covered by the file is assigned a unique “agent ID”, and the individual’s trajectory is stored as a list of 3-tuples of the form $< x, y, timestep >$. Each time step codes for one *frame* in the original footage (recorded at 9fps). (Majecka, 2009) note that “the sample rate can vary over short periods” due to errors with the capture program; however “since each captured frame is relatively independent of captured frames more than 10-20 seconds later”, this did not significantly impact on the quality of the resulting trajectories.

In what follows, we use the term “clip” to specifically refer to a time-limited sequence of trajectory data (whether taken from the Edinburgh dataset or from the output of a simulation), as opposed to a movie visualisation. We first wrote a script to convert a list of trajectories into a frame-by-frame representation of agent locations over time. This outputs co-ordinates for *all* of the visible agents at *each* time step, which is required for rendering the trajectories into videos, as well as for analysing the crowds at each point in time. We also wrote another script to essentially reverse this process (extracting individual trajectories from time step data), which is necessary for analysing certain features of individual trajectories in clips (both real and simulated).

4.1.3 Data cleaning

Occasionally lossy detection by the camera means that some trajectories have missing sections for several time steps; once rendered, these individuals temporarily disappear from the frame and then reappear. To address this, we automatically detected such situations and interpolated co-ordinates for the missing time steps when parsing the Edinburgh dataset. Each new co-ordinate is placed proportionally between the surrounding co-ordinates, depending on the number of missing time steps. As the Edinburgh data trajectories were recorded at 9 frames per second these additional co-ordinates prevent agents from disappearing in renders, but do not alter the overall shape of trajectories. Across the estimated 7.9 million coordinates in the dataset, a total of 230,046 trajectory time gaps were identified. Of these, 128,660 (55.93%) were made up of 1 frame and 49,794 (21.65%) were 2 frames in duration. The largest observed time gaps were 13 and 14 frames; however these were each only identified once, and were not present in the real crowd data clips used in this research. Approximately 99.20% of all identified time gaps were of 9 frames or fewer (approximately one second of camera tracking), and interpolation of these time gaps did not result in any observable issues. We also increased the number of frames per second of both sets of trajectories (real and simulated), from 9 to 72, by interpolating co-ordinates. This improved the “smoothness” of the trajectories once abstracted and rendered into video clips. This enables smooth video playback for the purpose of comparisons, but does not alter the shape of the trajectories, as the distance between co-ordinates is negligible. Figure 3 shows all co-ordinate trajectories in one crowd clip rendered to single images at both 9 and 72 frames per second.

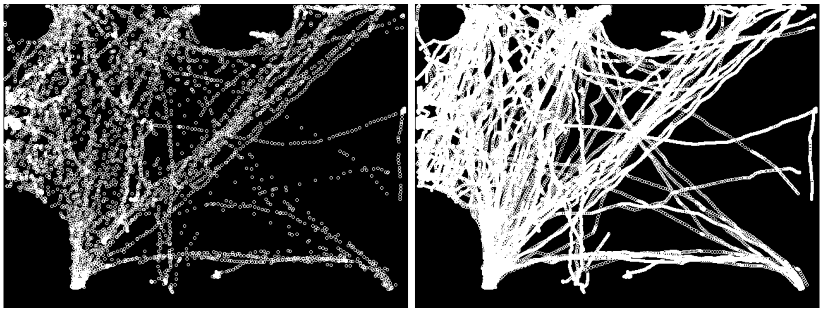

Figure 3: All trajectories in a crowd clip rendered to single images at 9 (left) and 72 (right) frames per second.

4.1.4 Visualisation

We wrote a utility to search the Edinburgh dataset and extract clips of a specific duration containing a specific number of individuals. Both simulated and real individuals were rendered in a uniform fashion, using a tool coded in Java. This allowed us to produce "top down" visualisations of both real and simulated clips that were identical in appearance, with individuals represented as filled circles, and headings depicted by an arrow (see Figure 4). Stationary agents in real crowd clips appear to "flick" their headings rapidly due to inaccurate camera detection, so headings are only rendered when an agent is in motion.

The use of abstract, simplified shapes, and a top-down, two-dimensional presentation is relatively common in crowd studies (N. Bode et al., 2015; Singh et al., 2009; Smith et al., 2009; Wagner & Agrawal, 2014; W. L. Wang et al., 2017; Zhang et al., 2019), although three-dimensional representations are also used (Loscos et al., 2003; Luo et al., 2008; Moussaïd et al., 2016; Pelechano et al., 2007).

As in (Webster & Amos, 2020), we decided against using realistic body shape rendering and 3D views, as initial tests suggested that such a presentation scheme (using animated

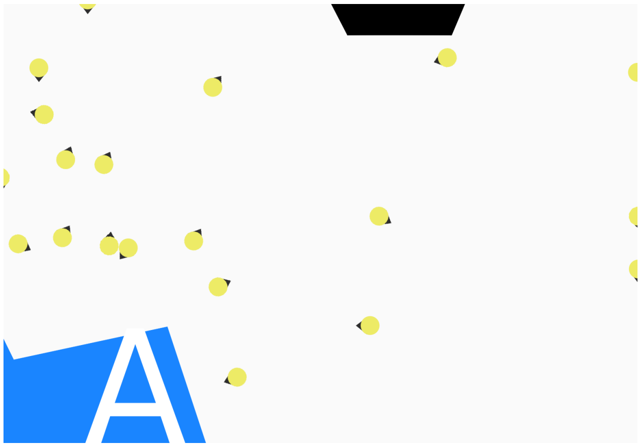

Figure 4: Example rendering of a crowd scene, taken from (Webster & Amos, 2020).

avatars) would actually distract viewers from the main aim of the experiment, which was to look for *patterns of behaviour* in the crowd. Additionally, at least one study has shown that crowds that are viewed from the top-down are perceived as being just as realistic as those viewed from eye-level (Ennis et al., 2011).

The simulated crowd trajectories were converted into the same format as the real crowds for rendering. Each time step has a corresponding set of co-ordinates representing a real or simulated person in the Edinburgh forum, as well as their heading. At every time step in a clip, our rendering tool generates a PNG image, and the sequence was then combined into a video. The staircase represented in blue is an obstacle which simulated agents avoid, and the staircase represented in black is an egress point located slightly inside the forum.

4.1.5 Clip analysis

For each clip, we extracted the route choice distribution and the entry time distribution for all individuals. This allowed us to initialise our simulations with the same distributions, ensuring that the runs closely matched the macroscopic properties of the real-world observations (while leaving room for the microscopic differences in which we are interested). In a later Section, we show heatmaps of the entry and exit distributions of the real crowd clips.

After rendering real crowd clips from the Edinburgh dataset for the first time, we saw a clear difference in the maximum velocity and acceleration of agents in several clips, with some agents moving unnaturally quickly. This was attributed to the variability in camera capture rate discussed earlier. To adjust for this variability, we calculated the average velocity of individuals in each clip, and used this to scale the clip's length (by modifying the video playback speed), thus normalizing the velocity of individuals relative to expected walking speed (Bohannon, 1997).

4.2 Simulation construction

Each test required participants to classify a number of clips of pedestrian movement as either real or simulated. We began by selecting, at random, a number of clips (30s duration) from the Edinburgh dataset, and extracting information about the number of individuals visible and the entry/exit point distribution. This information was then used to "seed" a simulation. In this way, we obtained both real and simulated versions of the same scenario; the real version was a rendered version of the actual observations, and the simulated version was a rendered version of the output of the model.

In order to model the scenarios captured in each real Edinburgh clip, we simulated pedestrian movement using the Vadere package (Zönnchen et al., 2020). This is an open-source package, which means that (unlike commercial software) its movement models are open to inspection. Importantly, it also allows for easy exporting of simulated pedestrian trajectories, which is necessary for rendering.

A crucial component of the simulation is the *crowd motion model*. This defines the rules of interaction between individuals (e.g., avoidance), and between individuals and their environment (e.g., repulsion from walls and physical obstacles), as well as route choice behaviour and differential walking speed. Many different crowd motion models exist (Duives et al., 2013), but perhaps the most commonly-used type is based on *social forces*. Helbing and Molnar's social force model (SFM) (Helbing & Molnar, 1995) is a microscopic, continuous model which uses "attractive" and "repulsive" force fields between individuals (and between individuals and their environment) to guide movement.

We selected the SFM as the baseline model for our simulations, as (1) it is very well-established and available for use in most open-source crowd simulation software, (2) "optimal" parameters have been refined over time, and (3) it is "recommended for pedestrian crowd movement research" following the thorough review by (Duives et al., 2013). We also compared the SFM with the Gradient Navigation Model (GNM) (Dutra et al., 2017), in order to avoid potential bias imposed by only using one motion model. The GNM is available as a default model type in Vadere, and we found that GNM simulation outputs have similar statistical properties to SFM outputs.

Table 1: Vadere simulation model parameters for SFM/GNM.

Parameter	SFM Value	GNM Value
ODE Solver	Dormand-Prince	Dormand-Prince
Pedestrian body potential	2.72	2.72
Pedestrian recognition distance	0.3	0.8
Obstacle body potential	20.1	20.1
Obstacle repulsion strength	0.25	0.25
Pedestrian radius (m)	0.2	0.2
Pedestrian speed distribution mean (m/s)	1.4	1.4
Pedestrian minimum speed (m/s)	0.4	0.4
Pedestrian maximum speed (m/s)	3.2	3.2
Pedestrian acceleration (m/s)	2	2
Pedestrian search radius (m)	1	1

For all simulations, we use the pre-supplied Vadere templates for the SFM/GNM, with default attributes and parameters (listed in Table 1). We note that all default parameter values are the same across both models, with the exception of "Pedestrian recognition distance" (0.3 for SFM, and 0.8 for GNM), but we do not believe this had any significant impact on our results.

Vadere stores its simulation input files in JSON format, and these files specify the topography of the simulation space and initial spawn parameters for each agent (or group of agents). This makes it possible to write a script which generates a JSON file for each simulation, including the Edinburgh forum topography, as well as a JSON object for each agent to be simulated. We ran each simulation in Vadere using the new simulation input files, and then imported each resulting file of crowd trajectories into MATLAB to be processed.

In Test 1 we used only the SFM movement model; in Test 2, we divided the simulations between the SFM and the GNM, in order to test whether different movement models have unique movement "signatures".

As discussed in (Webster & Amos, 2020), we added small amounts of noise to the simulated trajectories in order to replicate noise in the real crowd data. Typically, in crowd videos, shoulder "swaying" can account for perceived side-to-side movement of pedestrians; however, the Edinburgh individuals were detected by an overhead camera running at 9fps (placed too high to detect shoulder sway). However, occasionally faulty detection caused very short-term errors in the extracted trajectories. Once rendered, this caused individuals to appear to rapidly "flick" between two headings. As we had no reliable way to quantify the (by inspection, small) amount of noise in the trajectories, we adjusted this by eye until the apparent noise in the simulated data matched the noise level observed in the real data. At any time step, a simulated agent has a 15% chance of temporarily "flicking" their heading by a randomly selected value up to 45 degrees (without changing their trajectory). The inclusion of noise in simulations has been shown to replicate real behaviour in animal models (N. W. Bode et al., 2010) whilst "preserving emergent behaviours of previous models". In this case, the noise added to simulated trajectories only served to replicate faulty detection artefacts in the data, without altering the overall trajectories of the agents.

4.3 Simulation validation

It is important to ensure that simulations (regardless of the movement model) produce outputs that are valid, so we first calculated several statistical properties for a set of simulations and the Edinburgh observations on which they were based.

As in (Webster & Amos, 2020), we used two metrics (Herbert-Read et al., 2015); *polarization* and *nearest neighbour distance* (NND). The first metric is particularly useful for describing the existence of large groups who might be moving together along the same heading (e.g., leaving a lecture room and moving together towards an exit), while the second metric is used for estimating overall crowd density. Although these metrics have tended to be used in "swarming" models (e.g., of birds or fish) in which agents are supplied with local information about other agents in their vicinity, they have recently also been used effectively to assess a model of collective behaviour based purely on vision, which is perhaps better aligned to our current model (Bastien & Romanczuk, 2020).

Polarisation measures the level of "order" in a crowd, in terms of the heading alignment of members. Polarisation is zero when the crowd is completely disordered (everyone is pointing in a different direction), and has a maximum value of 1 when all members of the crowd have the same heading:

$$\varphi = \frac{1}{N} \left| \sum_{i=1}^{N} \mathsf{exp}(\iota \theta_i) \right| , \tag{1}$$

where N is the number of individuals in the frame, ι is the imaginary unit, and θ_i is the heading of each individual.

Nearest-neighbour distance (NND) measures the level of "clustering" in a crowd. The average NND for a single "frame" (derived from either the real dataset or the simulation) is calculated from the sum of nearest-neighbour distances of all N individuals:

$$\nu = \frac{1}{N}\sum_{i=1}^{N} d_i, \tag{2}$$

where d$_i$ is the nearest neighbor distance between point i and the closest individual in the frame, as calculated by the standard distance formula,

$$d_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{3}$$

We selected 20 random Edinburgh clips with varying crowd sizes, and then simulated each scenario 20 times with each movement model. Results are presented in Figure 5; these confirm that both movement models produce high-level outputs that are comparable to the real-world scenarios, and that there are no significant differences between the outputs of each movement model.

4.4 Classification tests

For both tests, we constructed a web-based application[1] which presented users with an information screen, asked them to click to confirm their consent to participate, and then presented participants with a randomised sequence of movies. For each movie, participants were asked to click either a "Real" or "Simulated" button, according to their own perception and opinion. At the end of the sequence, users were asked in a free text box to supply short notes on any features that they thought allowed them to identify the real crowd, to specify their level of expertise in crowd science ("High", "Medium" or "Low"), and to supply their email address (this was used as a participant ID to allow for tracking across the two tests). Once the user submitted their information, their responses were stored on the server, and they were told how many real crowds they had correctly identified (this may have inadvertently helped with recruitment, as some particularly high-scoring participants

[1] Available at http://www.martynamos.org/TTFC2/

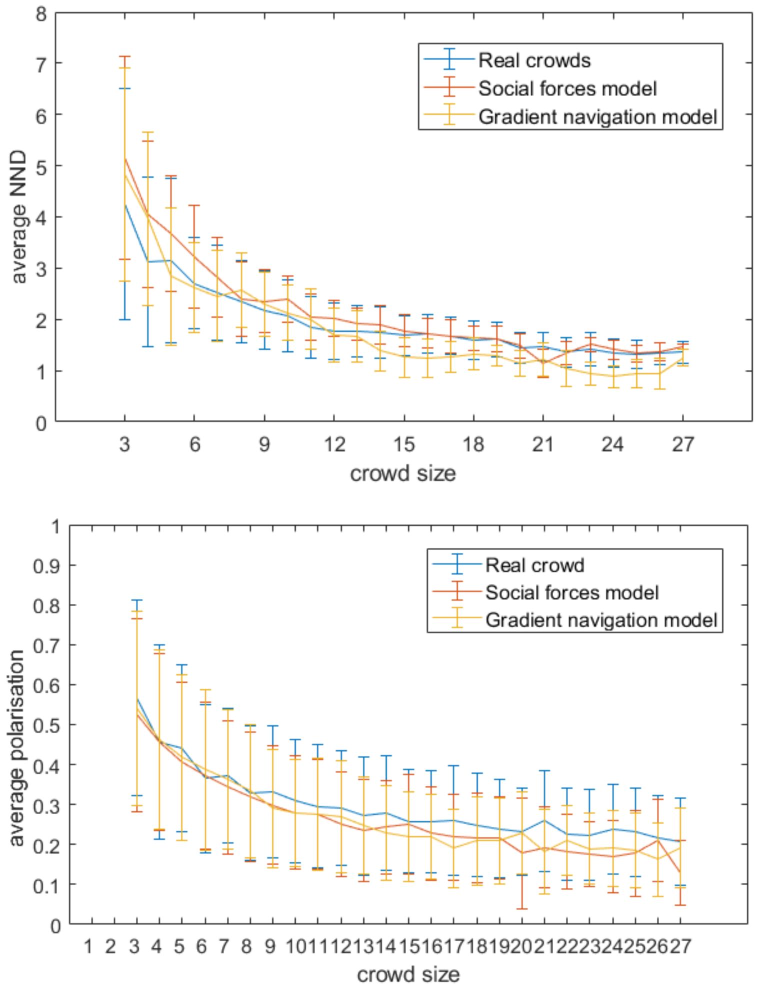

Figure 5: Movement models/real crowd statistical comparisons: Nearest Neighbour Distance (NND) (top) and polarisation (bottom) as a function of crowd size. The outputs of both movement models have properties that are close to those of the real crowds.

shared screenshots of their success on social media...)

4.4.1 Test 1

This was the “baseline” test to give each participant an initial score of their ability to classify movies as either real or simulated. We showed participants a sequence of 12 movies, 6 of which were based on real trajectories, and 6 of which were generated using the SFM-based simulation of that scenario. Each movie was 30s in duration (in all cases, participants were free to choose “early”, before the end of the movie, and move on to the next one).

For each real clip, the total number of individuals observed and average entry time interval is shown in Table 2 (the simulations were set up to reflect these). We present heatmap visualisations of the route choice distribution for each clip in Figure 6. The forum has 11 ingress points, and the 12th row and column represent individuals who start or end their observed trajectories *inside* the forum space.

Clip	Number of individuals	Mean entry time interval (s)	Standard deviation (s)
1	194	0.34	0.22
2	149	0.46	0.26
3	112	0.67	0.38
4	104	0.62	0.34
5	150	0.48	0.24
6	125	0.55	0.33

Table 2: The total number of individuals observed and mean entry time interval of each clip from Test 1.

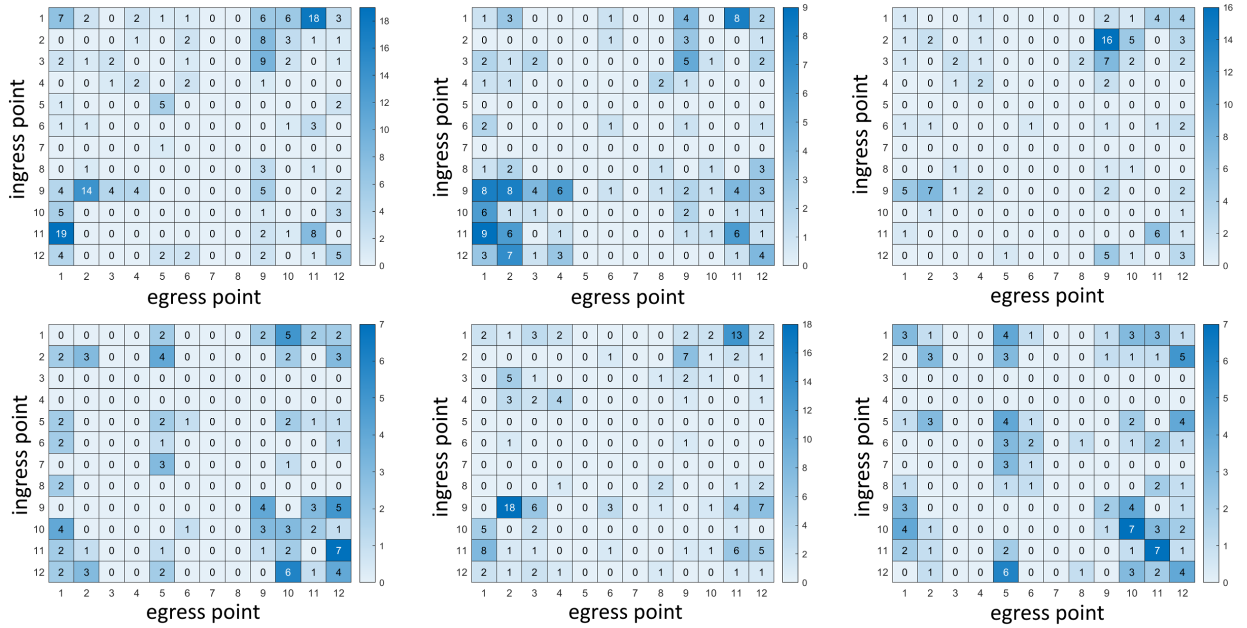

Figure 6: Heatmap representations of entry/exit point distributions for clips 1-3 (top) and 4-6 (bottom) from Test 1.

4.4.2 Test 2

We first required participants to undertake a training phase, in which they were shown 6 representative clips generated from Edinburgh observations. Participants were made

Clip	Number of individuals	Mean entry time interval (s)	Standard deviation (s)
1	149	0.49	0.27
2	122	0.54	0.28
3	132	0.47	0.26
4	162	0.38	0.24
5	144	0.39	0.26
6	133	0.47	0.47

Table 3: The total number of individuals observed and mean entry time interval of each clip from Test 2.

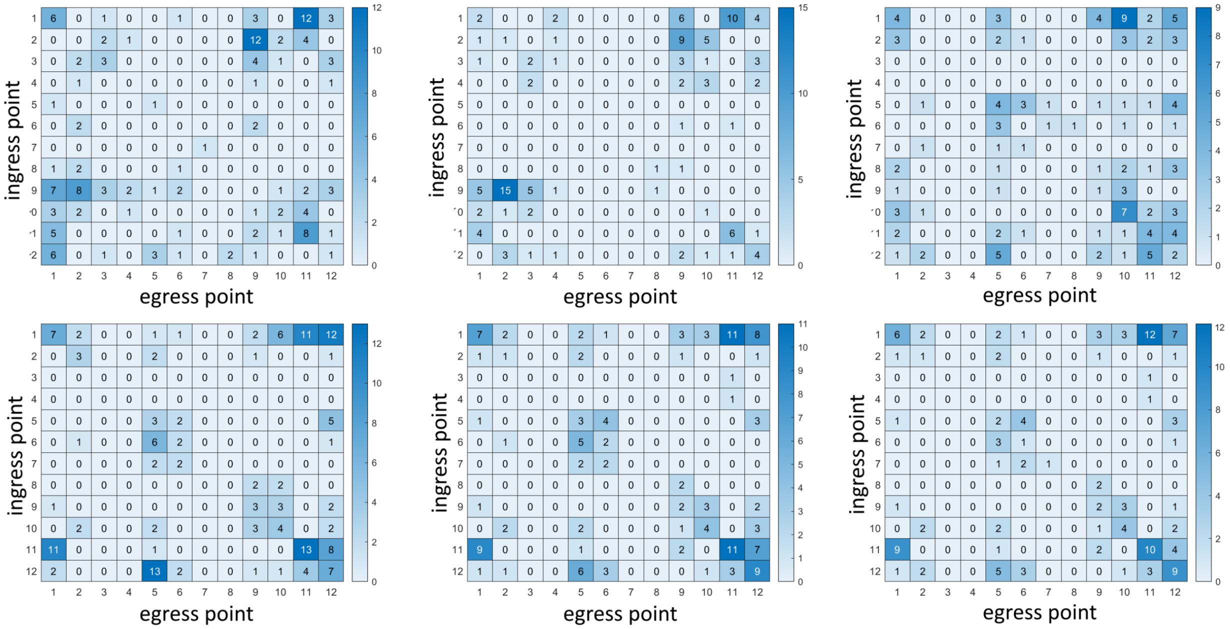

Figure 7: Heatmap representations of entry/exit point distributions for clips 1-3 (top) and 4-6 (bottom) from Test 2.

explicitly aware that they were watching real crowds. They were then shown 18 movies in total; 6 based on observations, 6 derived from SFM-based simulations, and 6 from GNM-based simulations.

For each real clip, the total number of individuals observed and average entry time interval is shown in Table 3 (again, the simulations were set up to reflect these). We present heatmap visualisations of the route choice distribution for each clip in Figure 7.

Set	**Test 1**	**s.d.**
$P_1 - P_2$	31.21%	20.19%
P_2	27%	19.31%

Table 4: Test 1 average scores for $P_1 - P_2$ and P_2. Scores are presented as “% correctly classified”, as the number of movies differed between tests. Analysis confirms that P_2 is representative.

5 Results

In this Section we present our trial results. In what follows, we adopt the following notation for participant groups; P_1 is the initial set of 232 participants who took Test 1 (to establish their baseline scores, with no training) and P_2 is the subset of 50 participants in P_1 who went on to take Test 2 (the new test that included a training phase to establish whether or not performance improves after viewing real crowd videos).

5.1 Classification accuracy

We first consider whether or not group P_2 is representative of the larger set of participants. In both Test 1 and Test 2, participants were scored according to their ability to correctly classify movies, and received 1 point for every correct classification. We calculate the average Test 1 scores for both $P_1 - P_2$ (that is, participants who only took Test 1) and P_2 (participants who took both Tests), and present them in Table 4 (scores are presented as % due to the fact that the number of movies differed between tests).

A Lilliefors test confirms that neither dataset is normally distributed, so we use a two-sided Wilcoxon rank sum test to confirm that data in $P_1 - P_2$ and P_2 are samples from continuous distributions with equal medians ($p = 0.0724$). We conclude, therefore, that P_2 is a representative group.

We then calculate the average Test 1 and Test 2 classification scores for P_2 *only*; these are shown in Table 5. This reveals a *significant* improvement in overall correct classification score after training (from 27% to 60%). In Trial 2, participants correctly identified SFM-

Test 1	s.d.	Test 2	s.d
27%	19.31%	60.22%	26.35%

Table 5: Test 1 and Test 2 average scores for P_2 only.

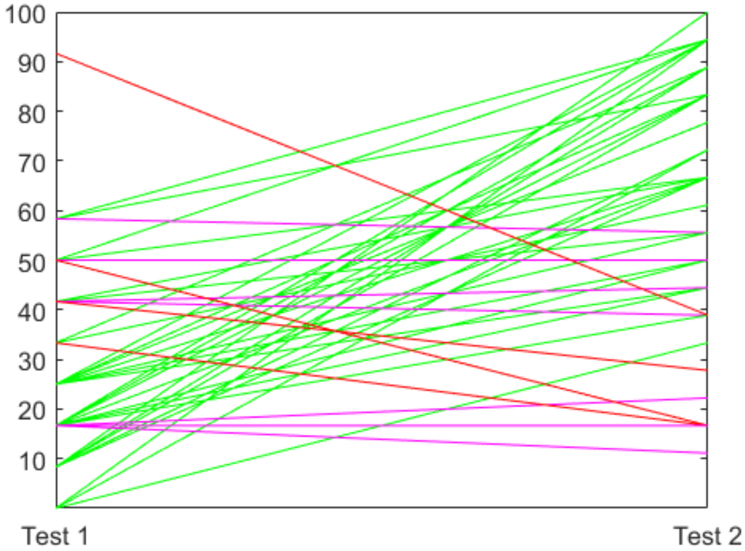

Figure 8: Slopegraph plot of changes in individual classification performance between Test 1 and Test 2 (50 individuals shown in total). Green lines show significant improvements, purple lines show small changes, and red lines show significant reductions in performance.

derived movies 63% of the time, and GNM-derived movies 59% of the time, so we cannot say that there exists a significant difference between the two models in terms of the overall characteristics of their outputs.

In Figure 8 we depict the individual changes in performance for the 50 members of P_2; visual inspection alone confirms that the vast majority of participants showed a marked improvement in classification performance after training. The average absolute change between Test 1 and Test 2 was 33.22%. If the participants had guessed at random in each test we would expect an average absolute change of 0%. A two-sided Wilcoxon signed rank test rejects the null hypothesis of a zero median in the distribution of average absolute change in our participant's test scores ($p < 0.001$). In Figure 9 we show the direction of improvement, confirming the bias towards an increase.

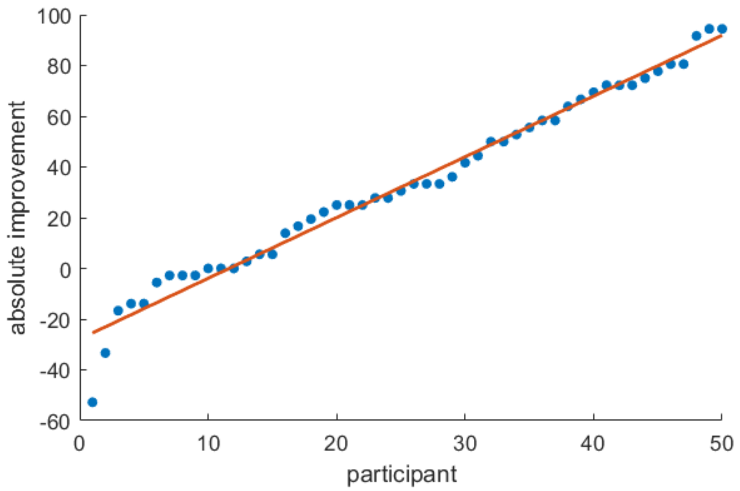

Figure 9: Trendline of absolute performance changes between Test 1 and Test 2 for P_2 participants.

These results confirm the first part of our hypothesis; that suitably trained individuals improve their classification performance after viewing movies of real crowds.

5.2 Narrative findings

We now move on to consider the free text supplied by members of P_2, and extract common themes that enable us to identify specific features of real and simulated crowds. We performed an initial version of this analysis in (Webster & Amos, 2020), but extracted only a small number of general themes, and did not correlate them with classification performance (as we do here). Our informal hypothesis is that participants who demonstrate significantly improved performance will correctly identify (in their free text responses) the characteristic features of both real and simulated crowds.

All 50 participants supplied feedback, so this provides useful additional context to explain the general uplift in performance. Given the relatively small amount of text, we performed manual thematic analysis to extract the predominant features highlighted in the supplied corpus. Each line of free text was broken down into thematic “atoms”, which were then semantically mapped onto over-arching themes. These are summarised in Table 6, par-

Real crowds	**Freq. %**
Heterogenous/diverse paths/speeds (R1)	9.21
Chaotic/unpredictable/erratic movement - rapid changes (R2)	21.05
Decisiveness/purposefulness - direct movement (R3)	6.56
Stop-start movement (R4)	7.89
Static individuals/groups (R5)	2.63
Groups/flocking/close proximity/collisions (R6)	7.89
Collision avoidance (R7)	5.26
Simulated crowds	**Freq. %**
Homogeneous behaviour (S1)	5.26
Rapid direction/speed changes (S2)	3.95
Goal-driven (S3)	3.95
Smooth/continuous movement (S4)	15.79
Clusters (S5)	1.32
Long interactions/collisions and close proximity (S6)	6.58
Collision avoidance (S7)	2.63

Table 6: Themes identified in narrative comments (labels given in brackets), and their observed frequencies. Related themes across “real” and “simulated” are numbered similarly, although there may not always be an *exact* correlation.

titioned into those features ascribed to real crowds, and those to simulated crowds. We also give the relative frequency of each feature/theme (a link to the full dataset is supplied at the end of the paper). We label each feature for ease of presentation/discussion.

We immediately notice two dominant features; R2 (*real* crowds exhibit chaotic or unpredictable movement, sometimes with rapid changes in speed/direction) accounted for 21% of thematic atoms, and S4 (*simulated* crowds show smooth/continuous movement) accounted for nearly 16% of all atoms. These observations are clearly complementary, in that (after training) observers believe that real crowds are more unpredictable than simulated crowds, which move more smoothly. The real dataset does include many examples of unpredictable/rapid changes in movement, where (we assume, not having access to the full video datasets) an individual is “dashing” across the space and adjusting their movements to avoid others, or where they double-back on themselves.

However, it is not sufficient to simply analyse the *frequency* of themes, since dominant features may not necessarily correlate with good classification performance in the partici-

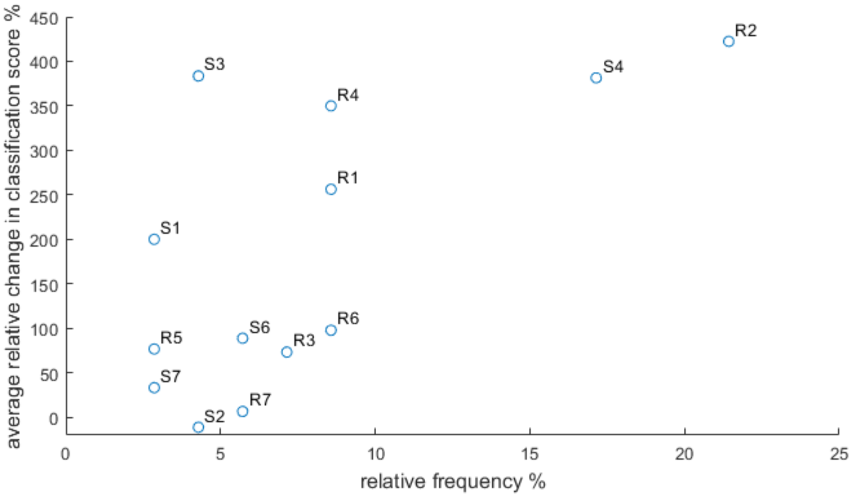

Figure 10: Thematic frequency versus average *relative* change in classification performance. The upper-right quadrant shows two themes (S4 and R2) which both appear frequently and which are correlated with significant positive relative change in classification performance in those participants who mention those themes.

pants who identify them. We also need to extract the features that have been identified *by the participants who perform best* (or who show the best relative improvement) in the classification task. We first consider *relative* changes in scores, and then look at the *absolute* changes, as each perspective yields insights.

In Figure 10 we plot each theme against both their frequency of mentions and the average relative change in classification performance of participants who specifically mention that theme. All scores are expressed in terms of the *percentage* of movies that were correctly classified, not the “raw” score (as previously stated, the number of movies differs between tests). For each participant, only where $score_1 > 0$, the relative change in $score$ is calculated by $((score_2 - score_1)/score_1 * 100)$. For example, a participant who scored 3/12 (25%) in Test 1 and 15/18 (83%) in Test 2 would have their relative change calculated as $((83 - 25)/25) * 100) = 232$%.

When calculating the average relative change, we discard 4 participants with a Test 1 score of zero, as the notion of relative change is not defined for a zero reference value (however, these participants are still included in the discussion of actual score differences, below).

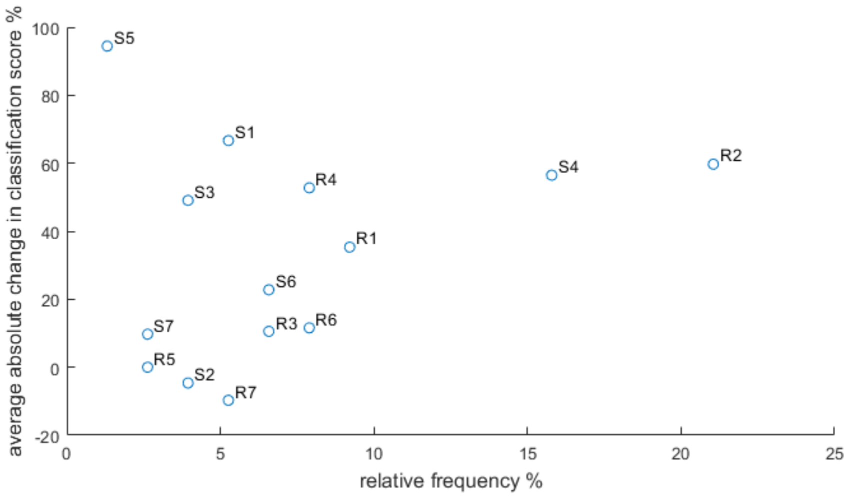

Figure 11: Thematic frequency versus average *absolute* change in classification performance. S2 and R7 are low -frequency themes that are nonetheless associated with reductions in classification performance.

We notice, from inspection, a cluster of themes that are relatively infrequently mentioned (< 10%), but which are associated with significant improvements in classification performance. However, we see that the two themes that are mentioned with frequency > 15% - S4 (smooth/continuous movement in simulated crowds) and R2 (unpredictable movement in real crowds) - are both also associated with performance improvements of around 400%. As noted earlier, these themes are complementary.

This finding is entirely consistent with our earlier informal narrative results (Webster & Amos, 2020), where participants who had “flipped” the real and simulated crowds believed that erratic movement was characteristic of “fake” (simulated) crowds, and that real crowds moved smoothly and predictably. After training on real crowds, however, the participants in this second trial correctly identified that real crowds are actually more noisy and unpredictable, and that overwhelmingly smooth, predictable trajectories are a characteristic of simulations.

We now consider *absolute* changes in classification score between tests. We see roughly the same clustering of labels as before (S5: presence of clusters in simulated crowds is an outlier, in that it was mentioned only by a single person, albeit one who saw a significant improvement in their classification score). Here we draw particular attention to the (albeit infrequently mentioned) themes that are correlated with *negative* shifts in performance. That is, the features that are mentioned by participants whose classification performance got worse after training. The two features to which this applies are S2 (rapid direction/speed changes in simulated crowds) and R7 (collision avoidance in real crowds).

Again, these findings are entirely consistent with both the current results and our previous study. If high-performing participants correctly spot that simulated crowds move smoothly, then it is entirely to be expected that low-performing participants will (incorrectly) ascribe S2 to them. Collision avoidance in real crowds (R7) is also specifically mentioned in our previous study; participants who performed badly assumed that individuals in real crowds would naturally avoid one another. As we observe in (Webster & Amos, 2020), "In reality, the opposite is true, as the real dataset contains multiple instances of individuals coming into close proximity. Moreover, the social forces model explicitly tries to keep individuals apart unless close proximity is unavoidable, so the behaviour (distance keeping) that participants attributed to real people was actually an in-built feature of the simulation."

However, we must approach these findings with a degree of caution, as it may be the case (for example) that the high-performing individuals are simply better learners, or some videos may be inherently easier (or more difficult) to classify. All we claim here is that there would appear to be a *correlation* between high classification performance and a small set of identifiable features of crowds. An investigation of the fundamental underlying process(es) is beyond the scope of the current paper, but may be performed in future work.

Based on these findings, we conclude that the primary feature of real crowds that allows trained individuals to correctly distinguish them from simulated crowds is their higher degree of unpredictability in terms of individual trajectories. A secondary feature is collision avoidance (specifically, proximity). Based on this work, our main suggestion (if what we seek is realistic believability in crowd simulations) is that models should include the facility to add a degree of unpredictability to the movement of individual agents (surprisingly, this feature is not generally provided). Models might also benefit from a relaxation of collision detection radii to allow for closer proximity of agents. In this way, we might easily replicate the appearance of at least some of the micro-level behaviours referenced by (Lerner et al., 2007).

6 Discussion and Conclusions

In this paper we report the results of a human trial to identify the "signature" characteristics of real crowds that allow them to be distinguished from simulated crowds. We find that unpredictability in terms of individual trajectories is by far the best discriminator, and proximity in collision detection is also relevant. We note some limitations of our study; the underlying crowd dataset is based on a relatively small physical space which is quite regular in nature, but we point out that it is actually much larger than the arenas used for artificial crowd experiments. Moreover, the observations have a higher level of ecological validity, as the recorded pedestrians were not consciously aware of being participants in an experiment. Our second test used a relatively small number of participants, but we have established that they were representative of a larger set. Finally, our findings are only applicable to "routine" crowds (that is, where people are going about their everyday business), and not to "emergency" or "evacuation" crowds, where behaviours will be very different.

However, there is still significant value in updating simulation of such routine crowds to render them more realistically, especially if important policy or design decisions are to be made based on how they are perceived. With this in mind, there may be value in training decision-makers who use such simulations as part of their process (in a manner similar to that performed in our Test 2), in order to ensure that they can first detect the characteristic features of real crowds (as opposed to making decisions based on flawed assumptions of how crowds behave). Fundamentally, the value of additional realism in crowd simulations may only be realised if end-users are able to *recognise* it.

This study has provided empirical evidence to support the inclusion of relatively straightforward modifications to any and all of the movement models underpinning both scientific and commercial crowd simulation packages. Importantly, the addition of noise to individual trajectories and the relaxation of collision detection radii are entirely generic updates, but ones that could significantly improve the believability of crowd simulations across a range of applications.

Future work may include the automatic detection of features of real crowds from larger and more complex datasets, consideration of the impact of changing movement model parameters, and the integration of identified features into commercial crowd simulation packages in order to test their impact on believability (thus "closing the circle").

7 Materials

All code (simulations and analysis scripts) and datasets generated are available at http://doi.org/10.6084/m9.figshare.c.5280902

Acknowledgements

JW was supported by a Ph.D. studentship from the Faculty of Engineering and Environment, Northumbria University. We thank Gerta Köster and her research team for useful discussions, and all of the trial participants for their contributions.

References

Adrian, J., Amos, M., Baratchi, M., et al. (2019). A glossary for research on human crowd dynamics. *Collective Dynamics*, *4*(A19), 1–13.

Aschwanden, G., Haegler, S., Bosché, F., Van Gool, L., & Schmitt, G. (2011). Empiric design evaluation in urban planning. *Automation in Construction*, *20*(3), 299–310.

Bastien, R., & Romanczuk, P. (2020). A model of collective behavior based purely on vision. *Science Advances*, *6*(6), eaay0792.

Bode, N. W., Franks, D., & Wood, J. (2010). Making noise: Emergent stochasticity in collective motion. *Journal of Theoretical Biology*, *267*(3), 292–299.

Bode, N., Kemloh Wagoum, A., & Codling, E. (2015). Information use by humans during dynamic route choice in virtual crowd evacuations. *Royal Society Open Science*, *2*(1).

Bohannon, R. (1997). Comfortable and maximum walking speed of adults aged 20—79 years: Reference values and determinants. *Age and Ageing*, *26*(1), 15–19.

Brambilla, M., Ferrante, E., Birattari, M., & Dorigo, M. (2013). Swarm robotics: A review from the swarm engineering perspective. *Swarm Intelligence*, *7*(1), 1–41.

Crociani, L., Lämmel, G., & Vizzari, G. (2016). Multi-scale simulation for crowd management: A case study in an urban scenario. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 147–162.

Cronin, L., Krasnogor, N., Davis, B., Alexander, C., Robertson, N., Steinke, J., Schroeder, S., Khlobystov, A., Cooper, G., Gardner, P., et al. (2006). The imitation game - a computational chemical approach to recognizing life. *Nature Biotechnology*, *24*(10), 1203.

Duives, D., Daamen, W., & Hoogendoorn, S. (2013). State-of-the-art crowd motion simulation models. *Transportation Research Part C: Emerging Technologies*, *37*, 193–209.

Dutra, T., Marques, R., Cavalcante-Neto, J., Vidal, C., & Pettré, J. (2017). Gradient-based steering for vision-based crowd simulation algorithms. *Computer Graphics Forum*, *36*, 337–348.

Ennis, C., Peters, C., & O'Sullivan, C. (2011). Perceptual effects of scene context and viewpoint for virtual pedestrian crowds. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 1–22.

Feng, T., Yu, L.-F., Yeung, S.-K., Yin, K., & Zhou, K. (2016). Crowd-driven mid-scale layout design. *ACM Transactions on Graphics*, *35*(4), 132–1.

Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2018). Tracking by prediction: A deep generative model for multi-person localisation and tracking. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1122–1132.

Fuchsberger, A., Tahmasbi, N., & Ricks, B. (2017). A framework for achieving realism in agent-based pedestrian crowd simulations. *AMCIS 2017 - America's Conference on Information Systems*.

Gloor, C. (2016). PedSim: Pedestrian crowd simulation [http://pedsim.silmaril.org].

Gorochowski, T. (2016). Agent-based modelling in synthetic biology. *Essays in Biochemistry*, *60*(4), 325–336.

Harding, P., Gwynne, S., & Amos, M. (2011). Mutual information for the detection of crush. *PloS ONE*, *6*(12), e28747.

Harel, D. (2005). A Turing-like test for biological modeling. *Nature Biotechnology*, *23*(4), 495.

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, *51*(5), 4282.

Herbert-Read, J., Romenskyy, M., & Sumpter, D. (2015). A Turing test for collective motion. *Biology Letters*, *11*, 20150674.

Kimura, T., Sekine, H., Sano, T., Takeichi, N., Yoshida, Y., & Watanabe, H. (2003). Pedestrian simulation system SimWalk. *Summaries of Technical Papers of the Annual Meeting of the Architectural Institute of Japan, E-1*, 915–916.

Klüpfel, H. (2007). The simulation of crowd dynamics at very large events - calibration, empirical data, and validation. *Pedestrian and Evacuation Dynamics (PED) 2005* (pp. 285–296). Springer.

Korhonen, T., Hostikka, S., Heliövaara, S., & Ehtamo, H. (2010). FDS+Evac: An agent based fire evacuation model. *Pedestrian and Evacuation Dynamics (PED) 2008* (pp. 109–120). Springer.

Lemercier, S., & Auberlet, J. (2016). Towards more behaviors in crowd simulation. *Computer Animation And Virtual Worlds*, *27*(1), 24–34.

Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. *Computer Graphics Forum*, *26*(3), 655–664.

Loscos, C., Marchal, D., & Meyer, A. (2003). Intuitive crowd behavior in dense urban environments using local laws. *Proceedings - Theory and Practice of Computer Graphics, TPCG 2003*, 122–129.

Lovreglio, R., Dias, C., Song, X., & Ballerini, L. (2017). Towards microscopic calibration of pedestrian simulation models using open trajectory datasets: The case study of the Edinburgh Informatics Forum. *Conference on Traffic and Granular Flow, Washington DC, USA*.

Luo, L., Zhou, S., Cai, W., Yoke, M., Low, H., Tian, F., Wang, Y., Xiao, X., & Chen, D. (2008). Agent-based human behavior modeling for crowd simulation. *Computer Animation And Virtual Worlds*, *19*(August), 271–281.

Mahmood, I., Haris, M., & Sarjoughian, H. (2017). Analyzing emergency evacuation strategies for mass gatherings using crowd simulation and analysis framework: Hajj scenario. *Proceedings of the 2017 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 231–240.

Majecka, B. (2009). *Statistical models of pedestrian behaviour in the forum* (Master's thesis). School of Informatics, University of Edinburgh.

Mckenzie, F., Petty, M., Kruszewski, P., Gaskins, R., Nguyen, Q.-A., Seevinck, J., & Weisel, E. (2008). Integrating crowd-behavior modeling into military simulation using game technology. *Simulation & Gaming*, *39*(1), 10–38.

Moussaïd, M., Kapadia, M., Thrash, T., Sumner, R., Gross, M., Helbing, D., & Hölscher, C. (2016). Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface*, *13*(122), 20160414.

Pelechano, N., Allbeck, J. M., & Badler, N. I. (2007). Controlling individual agents in high-density crowd simulation. *Proceedings of the 2007 ACM SIGGRAPHEurographics symposium on Computer animation*, *1*, 108.

Peters, C., & Ennis, C. (2009). Modeling groups of plausible virtual pedestrians. *IEEE Computer Graphics and Applications*, *29*(4), 54–63.

Pettré, J., Ondřej, J., Olivier, A.-H., Cretual, A., & Donikian, S. (2009). Experiment-based modeling, simulation and validation of interactions between virtual walkers. *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 189–198.

Pouw, C., Toschi, F., van Schadewijk, F., & Corbetta, A. (2020). Monitoring physical distancing for crowd management: Real-time trajectory and group analysis. *PloS ONE*, *15*(10), e0240963.

Pretorius, M., Gwynne, S., & Galea, E. (2015). Large crowd modelling: An analysis of the Duisburg Love Parade disaster. *Fire and Materials*, *39*(4), 301–322.

Reynolds, C. (1987). Flocks, herds and schools: A distributed behavioral model. *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, 25–34.

Ricks, B. (2013). *Improving Crowd Simulation with Optimal Acceleration Angles, Movement on 3D Surfaces, and Social Dynamics* (Doctoral dissertation). Brigham Young University.

Rivers, E., Jaynes, C., Kimball, A., & Morrow, E. (2014). Using case study data to validate 3D agent-based pedestrian simulation tool for building egress modeling. *Transportation Research Procedia*, *2*, 123–131.

Seer, S., Rudloff, C., Matyus, T., & Brändle, N. (2014). Validating social force based models with comprehensive real world motion data. *Transportation Research Procedia*, *2*, 724–732.

Seitz, M., Templeton, A., Drury, J., Köster, G., & Philippides, A. (2017). Parsimony versus reductionism: How can crowd psychology be introduced into computer simulation? *Review of General Psychology*, *21*(1), 95–102.

Singh, H., Arter, R., Dodd, L., Langston, P., Lester, E., & Drury, J. (2009). Modelling subgroup behaviour in crowd dynamics DEM simulation. *Applied Mathematical Modelling*, *33*(12), 4408–4423.

Smith, A., James, C., Jones, R., Langston, P., Lester, E., & Drury, J. (2009). Modelling contraflow in crowd dynamics DEM simulation. *Safety Science*, *47*(3), 395–404.

Templeton, A., Drury, J., & Philippides, A. (2015). From mindless masses to small groups : conceptualizing collective behavior in crowd modeling. *Review of General Psychology*, *19*(3), 215–229.

Thalmann, S., & Musse, S. (2013). *Crowd Simulation*. Springer.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.

Wagner, N., & Agrawal, V. (2014). An agent-based simulation system for concert venue crowd evacuation modeling in the presence of a fire disaster. *Expert Systems with Applications*, *41*(6), 2807–2815.

Wang, H., & O'Sullivan, C. (2016). Globally continuous and non-Markovian crowd activity analysis from videos. *European Conference on Computer Vision*, 527–544.

Wang, W. L., Lo, S. M., & Liu, S. B. (2017). A cognitive pedestrian behavior model for exploratory navigation: Visibility graph based heuristics approach. *Simulation Modelling Practice and Theory*, *77*, 350–366.

Webster, J., & Amos, M. (2020). A Turing test for crowds. *Royal Society Open Science*, *7*(200307).

Wei, X., Lu, W., Zhu, L., & Xing, W. (2018). Learning motion rules from real data: Neural network for crowd simulation. *Neurocomputing*, *310*, 125–134.

Yao, Z., Zhang, G., Lu, D., & Liu, H. (2020). Learning crowd behavior from real data: A residual network method for crowd simulation. *Neurocomputing*, *404*, 173–185.

Zhang, D., Zhu, H., Hostikka, S., & Qiu, S. (2019). Pedestrian dynamics in a heterogeneous bidirectional flow: Overtaking behaviour and lane formation. *Physica A: Statistical Mechanics and its Applications*, *525*, 72–84.

Zönnchen, B., Kleinmeier, B., & Köster, G. (2020). Vadere – a simulation framework to compare locomotion models. *Traffic and Granular Flow 2019* (pp. 331–337). Springer.