SFPE EUROPE

Q2 2021 ISSUE 21 AN OFFICIAL PUBLICATION OF SFPE

How Can We Improve Realism in Crowd Simulations?

By: Martyn Amos and Jamie Webster, Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK.

In 1950, the mathematician Alan Turing essentially founded the field of artificial intelligence with a landmark paper [1], in which he proposed an "imitation game" for humans and computers. This protocol (which would subsequently become known, more popularly, as the "Turing test") avoided the pitfalls of asking the ambiguous question "Can machines think?", by focusing on whether or not a computer could plausibly pretend to be a human being during a dialogue with an interrogator. A human participant puts questions to both another person and a computer, and the job of the machine is to respond in such a way as to persuade the observer that it is actually the human. According to Turing, if the computer's responses are indistinguishable from those of a human over a pair of conversations, then, for all intents and purposes, we would be justified in ascribing to the machine a degree of "intelligence".

But how might the Turing test be relevant to fire safety engineering? Crowd simulations now play an integral role in performance-based fire safety design, and they are applied in a wide variety of domains, from evacuation planning and management to incident response and analysis. Although a number of commercial and open-source software tools exist to simulate crowd movement and behaviour, the underlying "engine" (which determines how simulated people move through space and interact with one another and their environment) is often based on microscopic movement models such as the Social Forces Model, or a variant thereof [2]. This generates crowd behaviours that are macroscopically valid, but which still occasionally lack realism (or "believability"). Although the overall outcome of the simulation (e.g., the time taken to evacuate a building) may be valid, any visualization of the moving crowd may lack subtle features that are present in real crowds (for example, sudden changes in direction or speed).

This "reality gap" may present a challenge in terms of the adoption of policies based on simulation visualisations; put simply, decision makers may not entirely trust the outputs of these models because they intuitively feel that they are somehow "unrealistic". In our recent work, we addressed this reality gap, using a Turing test model that asked human participants to

classify movies of real and simulated crowds. Our initial hypothesis was that real crowds possess features that allow human observers to distinguish them from simulations.

We began by confirming the existence of features or patterns that are specific to real crowds, and which are not present in simulated crowds. This meant that we could be confident that real crowds have one or more "signature" features that allow them to be distinguished from simulated crowds. We then followed up this work with a second study, which specifically *identified* those signature features. These findings immediately suggest a number of relatively straightforward modifications that may be made to crowd simulation packages in order to close the believability gap that exists between their outputs and reality.

First trial: establishing the existence of "signature" features of real crowds.

Using observations taken from the University of Edinburgh, we selected a number of clips of time series data on "real" crowd movement. We then extracted the features of these clips (entry/exit point distribution, start/end time distribution, etc.) and used this information to set up simulations of those specific scenarios. In this way, we generated six pairs of movies of "real"/"simulated" crowds, in which the "real" crowd was based on observations, and the "simulated" crowd was generated by the Vadere crowd simulation package. For each pair, each movie was rendered using a custom visualiser, so that they could not be distinguished by appearance alone (Figure 1). The task of our participants was to select, for each pair, what they thought was the *real* crowd.

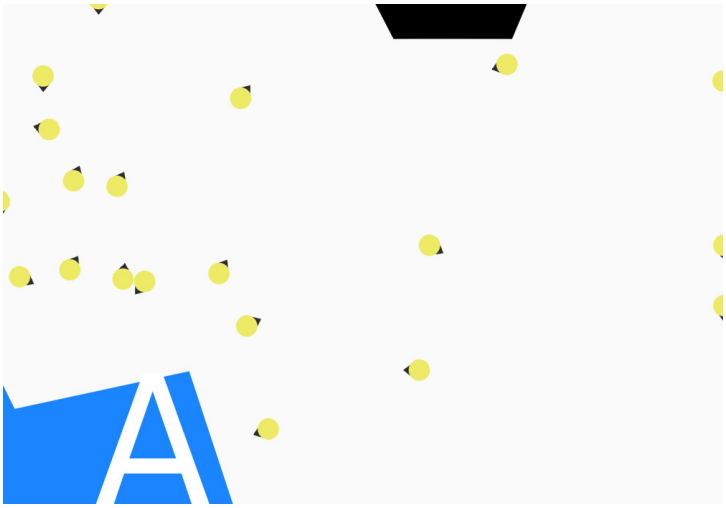

Figure 1. Rendering of a crowd clip. (Taken from [3]).

We recruited 384 students from Northumbria University for the initial identification trial. Our results were extremely surprising. The mean score was 1.6/6 (significantly worse than random guessing); the overall distribution of scores is shown in Figure 2, overlaid with the expected binomial distribution. A particularly striking result is that the most common score was zero. That is, a significant proportion of our participants were able to perfectly partition crowds into "real" or "simulated", but they were completely unable to tell which was which (i.e., classify them). A second version of the trial, in which we asked different participants to classify individual movies led to an improvement in performance, but still not enough to out-perform

random guessing. Our results [3], therefore, confirmed the *existence* of a set of features that allow humans to tell real crowds apart from simulated crowds (even if they are unable to ascribe them to the correct set). Our next task was to *identify* those specific features.

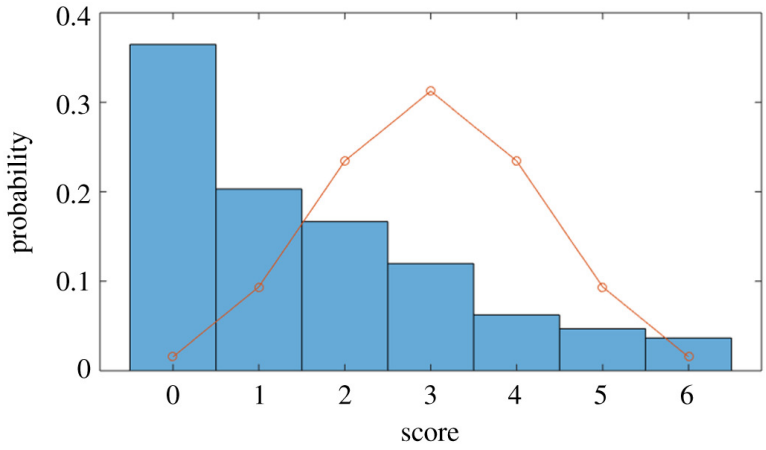

Figure 2. Identification trial: distribution of participant scores (the line represents the expected binomial distribution if individuals chose at random). (Taken from [3]).

Second trial: identifying specific "signature" features of real crowds.

Our follow-up study [4] was performed on-line, due to restrictions imposed by the COVID-19 pandemic. We recruited 232 participants via social media; the first test was a repetition of the classification test from the first trial (using different movies), which allowed us to assign each participant a "baseline" score of their ability to classify crowds. We then asked participants to sign up for the second test; 50 people in total responded (we offered a £10 Amazon gift card as an incentive). After an appropriate time had passed, we then performed a second classification test, but this time participants were first "trained" by showing them a series of movies that were explicitly described as being derived from real crowds. Our hypothesis was that participants would be able to detect features of real crowds from this training and apply this knowledge in the second classification test. The test is still available at www.martynamos.com/TTFC2

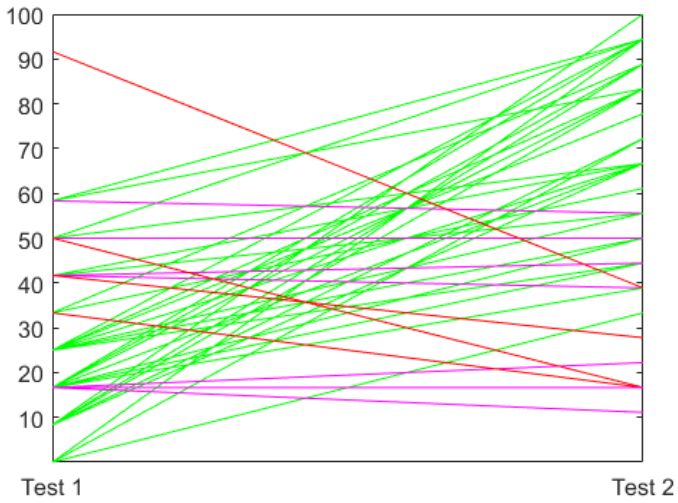

Figure 2. Slopegraph plot of changes in individual classification performance after training (50 individuals shown in total). Green lines show significant improvements, purple lines show small changes, and red lines show significant reductions in performance. (Taken from [4]).

Analysis of performances confirmed that the smaller second test group was representative of the larger first test group. We saw a *significant* uplift in classification performance (Figure 3), from 27% in the first test, to 60% in the second test. This confirmed our hypothesis that suitably trained individuals are able to improve their classification performance. We then used free text feedback supplied by participants to specifically identify the real crowd features that they found particularly helpful.

We performed textual analysis on the feedback to identify a number of common themes, and then plotted their frequency against the average improvement in classification performance. By focussing on features that both appeared relatively frequently and were mentioned by "high performers", we identified two main themes; that individuals in real crowds move more erratically/unpredictably, and that agents in simulations move much more smoothly. These observations are clearly complementary. A secondary feature we identified was collision avoidance; specifically, participants who performed badly assumed that individuals in real crowds would naturally avoid one another (we remind the reader that this test was carried out when strict social distancing rules were being applied, although these findings were replicated in the first trial, before the pandemic). The real dataset actually contained multiple instances of individuals coming into close proximity, and strict "bubbles" of avoidance were actually a feature of the simulation.

Conclusions and future work

Over two related studies, using more than 600 participants, we first confirmed the existence of "signature" features of real crowds that allow them to be distinguished from simulated crowds, and then specifically identified those features. We found that unpredictability in terms of individual trajectories is by far the best discriminator, with collision detection also providing useful indications.

Of course, the broader significance of these findings remains to be established, as (a) the dataset was taken from "routine" crowds moving through a space with a relatively simple topology, and (b) their applicability to evacuation scenarios should be further investigated. Nevertheless, our findings do suggest a number of straightforward modifications that could be made to commercial and scientific crowd simulation packages in order to improve their "believability". This will be the subject of our future work in this area (thus "closing the circle").

References

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

[2] Helbing, D. and Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, **51**(5):4282.

[3] Webster, J. and Amos, M**.** (2020). A Turing test for crowds. *Royal Society Open Science* **7**:200307. doi: 10.1098/rsos.200307.

[4] Webster, J. and Amos, M. (accepted). Identification of lifelike characteristics of human crowds through a classification task. To appear in Proc. *Conference on Artificial Life* (ALIFE2021), Prague, Czech Republic, Jul. 19-23 2021.