BELIEVABILITY ASSESSMENT FOR FIGHTING GAME AI

Mola Bogdan Georgyy, Maxim Mozgovoy, Toru Ito, and Tatsuhiro Rikimaru
The University of Aizu
Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580 Japan
realgolg@gmail.com, {mozgovoy, s1210045, s1210063}@u-aizu.ac.jp

KEYWORDS

Turing test, cosine similarity, game AI, believability.

ABSTRACT

We describe two methods of analyzing human and AI play style patterns in an arcade fighting game. The first is the application of a Turing test to study game characters' behavior. The second is the calculation of a cosine similarity between "behavior fingerprints" consisting of sequences of individual actions or combo chains. The main goal of this study is aimed to find an approach that helps to determine the believability of game AI. Our experiments with Universal Fighting Engine environment and its built-in AI system demonstrated that both people and AI agents exhibit different play styles, and AI agents are virtually indistinguishable from human-controlled characters.

INTRODUCTION

Most types of computer games implement AI logic in some form. According to Dill, the purpose of game AI is to support a certain player experience (Dill, 2013)). AI plays the key role in supporting players' entertainment; too weak or too strong AI can reduce the overall quality of a game. In certain game genres an AI system is supposed to imitate human behavior. This ability is especially important if AI controls human-like characters or replaces real players.

One of the problems of designing human-like AI is to find actual criteria of human-likeness that can be used to distinguish characters controlled by human players and by AI. One of the possible solutions is to adapt *Turing test* (Turing, 1950) for game AI assessment, i.e., to rely on human evaluation of the believability (Livingstone, 2006), (Gorman, Thurau, Bauckhage, & Humphrys, 2006). Alternatively, it is possible to test human-likeness by comparing behavior patterns of an AI system with those of human players, as shown in (Tencé & Buche, 2008) and (Mozgovoy & Umarov, 2010).

We discuss both approaches for analyzing play styles and believability of game characters in an arcade fighting game. The first approach is a Turing test-inspired series of player evaluations performed by people watching pre-recorded game clips. The aim of this approach is to investigate human ability to discern individual players and to separate AI-controlled and human-controlled characters. The second approach uses an automated evaluation algorithm that builds a "behavior fingerprint" for each game character. The fingerprints are then compared to reveal similarities and differences between the players and between human- and AI-controlled characters. The experiments show that each player in our game possesses recognizable behavior traits, but separating humans from AI agents is difficult.

In our experiments, we use a publicly available fighting game engine called Universal Fighting Engine (UFE). During game sessions, players can operate game characters by controlling six attack buttons and four direction keys. In addition, the players can make game characters perform special actions such as fireball and uppercut by using key combos.

UFE contains a built-in AI system called *Fuzzy AI Add-on* that uses fuzzy logic to evaluate the information of the scene and calculate the desirability of each given action, translating the AI decisions directly into user input. In the experiments we use three AI-controlled characters, based on different Fuzzy AI settings: 1) very easy; 2) normal; 3) impossible.

TURING TEST FOR PLAY STYLE ANALYSIS

To verify whether human players show unique play styles and whether human players are distinguishable from AI-controlled characters, we prepared two types of a Turing test.

Matching game clips test

We asked a group of testers to watch five game clips, each showing a match between a player A-E and a random opponent. Players A-C were controlled by three different persons, while Player D and Player E were controlled by the Fuzzy AI system set to a very easy and normal modes respectively. Next, we asked the testers to watch five more clips showing the same players A-E playing against random opponents. Finally, the testers had to accomplish the following assignments:

1. To identify pairs of clips showing the same players A-E.
2. To identify whether each character A-E in the latter five clips is controlled by a person or by an AI system.

A tester gets one point for each correct pair or answer, and we perform the experiment twice. Therefore, the best possible score is 10 for each question, and the total number of clips each tester has to watch is 20.

Grouping game clips test

For this test, we prepared 15 clips showing each of the players A-E fighting against a random opponent three times. Players A-C were controlled by three different persons, while Player D and Player E were controlled by the Fuzzy AI system set to normal and impossible modes respectively. We have showed these clips to the testers, and asked them to:

1. Group together three clips of to the same player A-E.

2. Clasify the players into the "human" and "AI" groups.

The tester gets two points for each correct group of three clips belonging to the same character and one point for an incomplete group of two correct and one wrong clips. Therefore, one may score up to 10 points in this task.

When the tester correctly marks a group of three AI-controlled charcters as "AI group", we add three points to the score. When the marked group contains only two AI-controlled characters, we add two points. Therefore, the tester can score up to 15 points in the second assignment.

Turing test results

We carried out the "Matching game clips test" with the help of 10 testers. All of them are male students, 21-22 years of age, having vastly different experience with fighting games, ranging from "no experience" to "over 100 hours".

Our experiments show that the testers possess different guessing abilities: three testers scored 7-8 points, four testers scored 5-6 points, and three testers scored only 2-4 points. Interestingly, the outcomes of the assignments seem not to be related. For example, tester 1 scored well on the first ("matching-1") assignment, but performed poorly on the second ("matching-2") assignment. Similarly, tester 6 got a high score for the second assingment, but showed average result on the first assignment. Furthermore, the lack of experience in playing fighting games did not significantly affect the results. For example, tester 8 indicated that he has no experience of playing fighting games, and yet he got a high score in the first assignment. The ability of people to identify distinct play styles in a fighting game becomes apparent in comparison with the random guessing algorithm that provides "baseline" scores, obtained by running the algorithm 200 times (Table 1). It is also clear that the abilities of individual testers are highly dispersed.

Table 1: Matching (M) and Grouping (G) test scores

	Human Evaluation				Random Guess			
	M1	M2	G1	G2	M1	M2	G1	G2
Average score	5.3	5.0	4.7	8.6	1.9	4.8	2.1	7.5
Standard deviation	2.0	2.4	2.5	0.9	1.4	1.6	1.2	1.9

The results of the second ("Grouping game clips") Turing test were obtained with the help of 9 testers from the same initial group of testers. These results are generally consistent with the first "Matching game clips" test. Again results of the individual test assignments seem not to be related. For example, tester 6 scored poor in the first assignment, but was able to get a high score in the second assignment. The results further prove that game experience does not help: the testers 6 and 8 indicated that they have "over 100 hours of play" experience, but still scored poorly in the first assignment.

To summarize the results of the Turing tests, we may note that on average people consistently beat random guessing algorithm in play style-related assignments: 5.3 points vs. 1.9 points in the first test, 4.7 points vs. 2.1 points in the second test. However, people performed only marginally better than the random algorithm in the task of identifying AI players.

AUTOMATIC IDENTIFICATION OF PLAY STYLES

We also analyzed player similarity in Universal Fighting Engine by comparing behavioral fingerprints, obtained with two different methods. The first method represents fingerprints as vectors of probabilities of individual actions in a certain player's game log. The second method represents fingerprints as matrices of probabilities of two consecutive actions in the game log. The obtained fingerprints are compared with cosine similarity measure (Nguyen & Bai, 2010). In order to compare matrices, we first convert them into vectors by rewriting matrix elements row after row.

The experiment was performed as follows.

1. We organized a tornament for five human players A-E and three AI opponents, controlled by Fuzzy AI with different difficulty settings (Ve: very easy, No: normal, and Im: impossible).
2. These players played three matches against each possible opponent, and the game logs were recorded.
3. We used game logs to calculate behavior fingerprints of the game characters, and compared the fingerprints against each other. To evaluate the consistency of behavior of the same characters in different matches, we compared their fingerprints obtained on different game logs and averaged the results.

UFE implements 33 actions, so an individual actions-based fingerprint consists of 33 elements. Similarly, a fingerprint obtained on two-action combos, consists of 33×33 elements.

Results of cosine similarity analysis

The Table 2 (lower half) shows player style similarities calculated using a cosine similarity value for vectors of probabilities of individual actions. In general, we can see high similarity scores between the fingerprints of the same player, and much lower similarity between the fingerprints of distinct players. The only exception is the pair C-D, having a higher similarity score than than D-D.

The upper half of the table shows player cosine similarity values obtained for the combo chains-based fingerprints. These values are comparable to the ones shown in the lower half, so the method based on the combo chains gives no significant improvements. The similarity of distinct players' fingerprints is ≈0.5 on average, while the similarity of different fingerprints of the same player is ≈0.8 on average.

The idea to calculate characters' fingerprints on the basis of action combos was motivated by the suggestion that such fingerprints would include more data and thus supposedly would imrove the results (i.e., different players will get lower similarity scores, while the same player in different matches will get a higher score). However, it turned out that the difference obtained using these two methods is marginal.

Furthermore, there is no significant difference between the results obtained for human- and AI-controlled characters, which reinforces the observations made during our Turing tests: it seems that separating human players from AI players is indeed difficult.

Table 2: Cosine similarity values
(lower half: individual actions; upper half: combo chains)

A	0.81 0.80	0.45	0.31	0.34	0.47	0.44	0.44	0.44
B	0.40	0.84 0.78	0.36	0.55	0.65	0.71	0.43	0.69
C	0.20	0.28	0.85 0.96	0.66	0.60	0.31	0.43	0.36
D	0.25	0.38	0.78	0.68 0.73	0.61	0.52	0.50	0.53
E	0.49	0.60	0.70	0.62	0.73 0.74	0.55	0.50	0.58
Ve	0.46	0.55	0.41	0.50	0.53	0.90 0.73	0.50	0.70
No	0.45	0.27	0.54	0.54	0.45	0.60	0.79 0.91	0.59
Im	0.48	0.51	0.45	0.55	0.50	0.73	0.81	0.79 0.85
	A	**B**	**C**	**D**	**E**	**Ve**	**No**	**Im**

CONCLUSION

It was interesting for us to apply two different approaches in this study: automatic and human assessment. The main feature of a Turing test is direct involvement of target users. They judge game AI subjectively and ofen inaccurately. However, human perception is the ultimate judge of the resulting quality of game atmosphere and character believability. One of the main challenges for a successful Turing test is to engage enough testers to get adequate results of evaluation (however, there are no established recommended group sizes). A major downside of a Turing test is caused by the limits of typical human abilities, and may lead to incorrect results. In particular, tiredness that occurs if the amount of video clips is big enough may distort judgement. Next, the testers cannot keep track of a large number of agents and still make correct decisions. Thus, it becomes impossible to evaluate large sets of agents and lengthy game sessions. Therefore, a Turing test has low sclability. However, we have to emphasize that the analysis of human impressions is the only direct way to evaluate perceived believability and play style similarity.

The cosine similarity method has its own advantages. It provides the same stable reliability for any number of agents to be evaluated. The main disadvantage of automatic methods lies in their indirect and unreliable way to imitate human perception. With this approach there is always a chance to miss gameplay elements, ignored by the used algorithm, or to treat as significant certain details, ignored by people. Our experiments revealed that the characters in Universal Fighting Engine exhibit distinct play styles, distinguishable both with Turing tests and automatic assessment procedures. However, it is much harder to distinguish human- and AI-controlled characters. We cannot explain this observation reliably, but most probably it means that either Fuzzy AI is indeed human-like enough to be difficult to uncover, or most reasonable game strategies are relatively straightforward and thus leave little room for individual improvisation.

The results of automated play style analysis agree with the Turing tests thus proving that our evaluation algorithm is adequate for this task. We can suggest that automated assessment is inevitable for large number of game characters and long game sessions, but smaller-scale Turing tests are necessary to prove that the chosen method agrees with human evaluation. We hope that the present work will provide some insights into the nature of human behavior patterns in a fighting game, and will be helpful for futher development of human-like AI in this genre.

REFERENCES

Tencé, F., & Buche, C. (2008). Automatable Evaluation Method Oriented toward Behaviour Believability for Video Games. In *International Conference on Intelligent Games and Simulation*, pp. 39–43.

Mozgovoy, M., & Umarov, I. (2010). Building a believable agent for a 3D boxing simulation game. In *Second International Conference on Computer Research and Development*, pp. 46–50.

Nguyen, H. V., & Bai, L. (2010). Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision*, pp. 709–720.

Gorman, B., Thurau, C., Bauckhage, C., & Humphrys, M. (2006). Believability Testing and Bayesian Imitation in Interactive Computer Games. *Lecture Notes in Computer Science, 4095,* pp. 655–666.

Dill, K. (2013). What is Game AI? In S. Rabin (Ed.), *Game AI pro. Collected wisdom of game AI professionals*, pp. 3–10.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59,* pp. 433–460.

Livingstone, D. (2006). Turing's Test and Believable AI in Games. *Computers in Entertainment, 4*(1), pp. 6–18.