Evaluating Human-Consistent Behavior in a Real-time First-person Entertainment-based Artificial Environment

G. Michael Youngblood and Lawrence B. Holder

Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019 U.S.A.
{youngbld, holder}@cse.uta.edu

Abstract

A two-part method for objectively evaluating human and artificial players in the Quake II entertainment-based environment is presented. Evaluation is based on a collected set of twenty human player trials over a developed reference set of one hundred unique and increasingly difficult levels. The method consists of a calculated Time-Score performance measure and a *k*-means based clustering of player performance based on edit distances from derived player graphs. Understanding human and agent performance through this set of performance and clustering metrics, we tested this evaluation method utilizing our CAMS-DCA (Cognitive-based Agent Management System-D'Artagnan Cognitive Architecture) agents for human performance and consistency.

Introduction

Despite the increased level of research on the creation of Artificial Intelligence Platforms and game playing intelligent agents seeking to replace or imitate humans, validation techniques currently focus on human expert evaluations and opinions (Laird 2002). There is a need for more qualitative, consistent, and objective validation of agent performance, especially in the area of human-like or human-level performance evaluation—or what we more generally term as the level of human-consistency. We forward that the ability and level of human consistency of an intelligent agent should be based on an objective and qualitative analysis of its performance on a set of test scenarios against the performance of humans performing in the same environment under the same or similar constraints as the agent being evaluated.

Many research groups have focused on human-level performance and human-like behaviors for intelligent agents in electronic games and as Computer Generated Forces in simulation (Kaminka et al. 2002, Laird 2002, Laird 2001). To a further extent, some recent research has even focused on interchanging agents and humans in military simulation (Heinze et al. 2002).

In our own intelligent agent work involving a cogn itive-based agent system (CAMS-DCA) playing a defined subset of the electronic entertainment game Quake II (Youngblood 2002), we have developed a system that seeks to validate the human behavioral models of the agent by temporal qualitative analysis. This analysis compares sequences of steps in achieving goals and actions in time against those of a set of human players (Gonzalez 1999). The analysis utilizes two metrics for identifying agent performance in relation to human performance to provide a measure of the level of human-consistency of an agent. This approach could also be used for the evaluation of other humans as well.

This paper begins by discussing the established evalu ation environment and the human trials we conducted to acquire data. We then present our two evaluation metrics applied to this data and discuss the notions of human performance levels and human-consistency. Our findings from each of these metrics on our collected data and observations are discussed. We conclude with a summary discussion leading to future work.

Evaluation Environment

A test set of one hundred Quake II levels was created starting with simple one-room environments and culminating in five levels modified from freely available enthusiast-created levels that represent the typical difficulty of the original game. Each level was created to slowly introduce new concepts starting with very simple levels and gradually working up to the last five challenge levels, which involve some amount of problem solving to finish. The goal of each level was to find a compact disc (CD). All players were told not only to maximize their health by not receiving damage and collecting health points, but also to move through the levels as quickly as possible. Each level is independent, and the player state is reset upon entry into the next level. An example test level is shown in figure 1.

a)

b)

c)

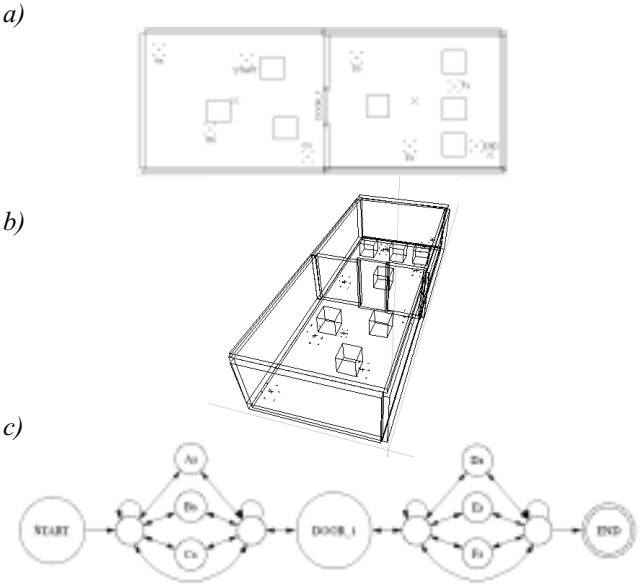

Figure 1. Quake II Example Test Level a) 2D view b) 3D view c) valid state transitions

Human Trial Data

We were fortunate to find twenty-four people to participate in our Quake II human trials. Of that group, only twenty completed a majority of the levels well enough to provide us with a data set. The roughly 17% that did not complete the levels cited extreme virtual reality sickness.

A large amount of data was collected which consisted of the name, position in three-dimensional space, and state information for every active object in the Quake II environment updated at 10 Hz. We needed to determine a way to represent a player's interaction in the Quake II environment for each level in a way that made it easy to compare players to other players. Observing the test set of levels we created, we noted that we could break down each level into a set of interaction feature points based on attaining or doing certain things in each level (e.g., pushing a button, picking up a health item, and so forth). These interaction feature points were all of the objects in each level that a player could interact with and that were tracked as individual objects by the game engine. Using this we could generate a graph composed of each level's interaction feature points as nodes, and the ability to move from one interaction feature point to another would create the edges as was physically possible in each level. For each level, the interaction feature points were identified and placed in a reference file, and a valid state transition diagram (e.g., figure 1.c) level was generated for additional validations.

One problem with the graph representation is that there is a definite issue of timing in games such as Quake II, and often the real difference between performances in a level is the time it takes different players to accomplish the same task. Gonzalez (1999) and Knauf et al. (2001) also note the importance of time in validation of human models. We capture time by weighting the edges a player traverses between interaction feature points with the time it had taken the player to travel that distance. Now we can abstract a player's performance in a level, removing the noise of motion through the environment, and capturing their approach of interaction—moving from one interaction feature point to another in a level while also capturing the time aspect of their performance. Figure 2 illustrates a captured player's actions in a level and the corresponding graph.

a)

b)

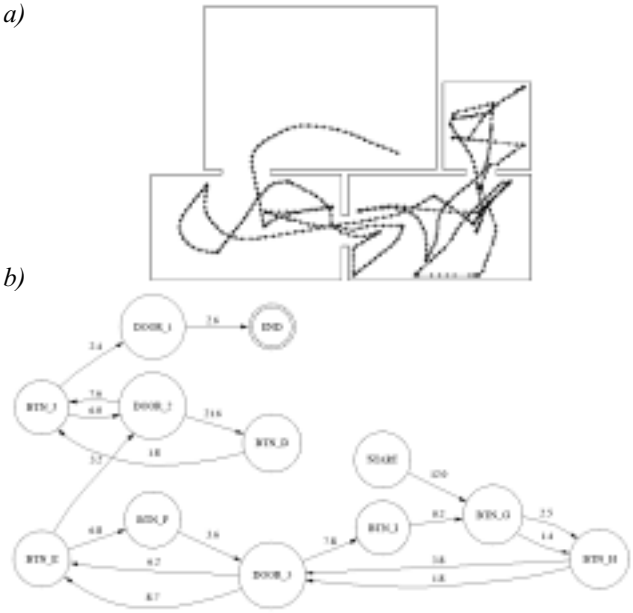

Figure 2. Player Trial a) actual recorded path and interactions b) resultant graph representation

Data Evaluation Metrics

In the course of evaluating the data we had collected, we developed two metrics for the evaluation of the data. The first metric produces a gross indicator of relative performance on each level based on the overall goals of maximizing the player's health and minimizing the time spent in each level. The notion of a Time-Score (TS) was used based on equation 1. Time in seconds was used with the metric having a scale of 0-200. The maximum possible health attainable for each level is known. However, the time-based metric facilitated the need to determine an ideal, yet unattainable, baseline time for completion for each level.

$$TS = \left(\left(\frac{health_{player}}{health_{\max}}\right)\cdot 100\right) + \left(\left(\frac{time_{\min}}{time_{player}}\right)\cdot 100\right) \qquad (1)$$

The second evaluation metric we designed uses the graph data generated from each level for each player and generates a matrix of edit distances between all players to serve as the distance function for use in *k*-means clustering (Friedman and Kandel 1999). By clustering player performance, we hoped to see clusters of players grouped by their relative skill level. If players cluster into their skill levels, then we would have a metric of player classification. More importantly, if we evaluate agent trial data with human trial data and the agents cluster into groups with human players, then we can assert that the agent played consistent with that group of humans, or that the agent behaved in a human-consistent manner for that level. If a player or agent constantly appears in the same cluster in a clustering over several levels, then we could classify the player or agent by its group associated performance.

We generated a matrix of edit distances between player graphs over all players for each level. The edit distance is defined as the minimum number of changes required to change one graph into another; where we define a change as the insertion or deletion of an edge, or an increase or decrease in time on an edge by 0.1 seconds. Each change carries an equal weight of one. This gives preference to time since it is the major factor of difference, but in Quake II time performance *is* the major differentiator between players. This heuristic could be easily adapted to place higher value on traversal than time or vice versa.

K-means clustering was performed for each level. We seeded our initial cluster centers with random members of our data set, and we iterated our clusters over all possible initial seed values.

We evaluated different clustering quality measures in order to produce the best achievable clusterings, creating several measures of clustering quality, but we only present the best observed one here as shown by equation 3. We exercise k from 2 to (n-2), where n is the number of trials by humans and/or agents, over the data set to ensure we find the best clustering in accordance with our clustering quality criterion function.

The clustering criterion function in equation 3 utilizes the distance measure, which is the distance between the j^{th} member of cluster i and the q^{th} member of cluster p, where the distance d represents the graph edit distance between two members.

We can define the mean intra-cluster distance of cluster i as follows, where n_i is the number of members in cluster i as shown in equation 2.

$$\overline{d}^{\,i}_{INTRA} = \frac{\sum_{j=1}^{n_i} \sum_{q=j+1}^{n_i} d(x_{ij}, x_{iq})}{\frac{n_i(n_i-1)}{2}} \qquad (2)$$

Then, the clustering criterion function is as follows in equation 3, where the minimum is taken over all possible assignments of members to the k clusters. This measures the normalized mean intra-cluster distance.

$$\text{Minimize:} \sum_{i=1}^{k} \frac{\overline{d}^{\,i}_{INTRA}}{n_i} \qquad (3)$$

The resultant clusters are used to group players by the similarity of their performance in the game.

Human-Consistency

The two metrics we employ for evaluation become the basis for our claims on performance and consistency. The Time-Score metric taken over a trial set of evaluated human players will create a data set that will allow for statistical analysis. For each level we compute the maximum and minimum values of the combined Time-Score metric over the trial set of human players. If another player falls within that range for a specific level we shall say that they were within the range of human performance for the level or, in other words, they completed that level within human performance levels.

We define human-consistency through the utilization of our clustering metric to be the property of being clustered in the same group as another human player. This means that there exists a strong enough similarity in time and traversal through a specific Quake II trial level that players are clustered into the same cluster. This is based on the clustering quality criterion, which utilizes the edit distance as a measure of graph similarity between graphs established by a set of known nodes defined by a level's interactive feature points. The edges are weighted by a player's traversal time in tenths of a second between interaction feature points. We argue that this provides enough granularity for comparison while removing noise and making a player's performance discretization computationally tractable. By this definition and through the clustering metric established, players that cluster with one or more known human players can be called human consistent, because they, at the least, played in a manner consistent with one or more human players. Human consistency by this definition must be bound to the Quake II environment until further research is conducted into applicability to other environments and domains. We believe that such a definition of human-consistency will transfer to other first person shooters and the genre of 3D character-based games in general, as well as military simulations and interactive simulation in 3D virtual worlds.

Our current research has led us to develop a cognitive-based agent model (the D'Artagnan Cognitive Architecture or DCA) that uses a multi-agent system (Cognitve-based Agent Management System or CAMS) as a framework in order to create an agent society to play in our modified Quake II environment. We are interested in creating an agent that plays like human players. It is this system to which we are applying the described evaluation methods.

Time-Score Metric Findings

In our examination of the data collected from our human user trials, we can see the Time-Score metric over all human players evaluated over all levels, and there is clearly a band of human activity for each level. This is illustrated in figure 3 by presenting a subset of the statistical values of the data set. In figure 3, we create a band of human normalcy in the metric by establishing minimum and maximum bounds of the observed humans.

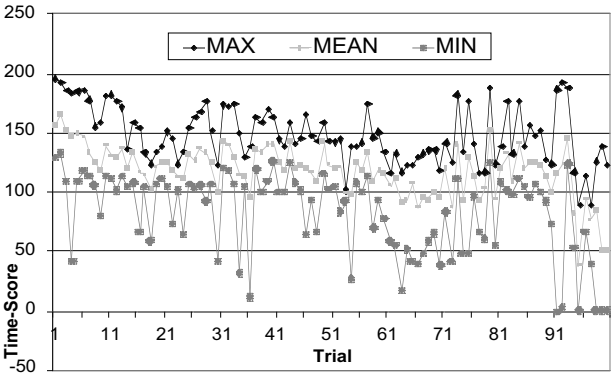

Figure 3. T-S Statistical Data from Human Player Trials

Human player trial data was observed to group in bands of similar performance across levels. It is understandable that performance in each individual level would group together, but after examining the performance data, we observed that there are bands of players that perform similarly across all levels.

Through Time-Score we have a rough measure of player performance and a means to generally classify human performance for each level. Players that do not perform in a human-consistent manner should fall outside of the statistical boundaries of our human player models.

Clustering Metric Findings

We now analyze the concept that players who play in a similar manner should collect into groups identified by their performance characteristics. Through our second evaluation metric, clustering, we will show the process of being able to group, or cluster in this case, human players based on their actual evaluated game performance at reaching interactive feature points present in each level. Utilizing a matrix of computed edit distances, derived from the low-level logging of player performance data processed into graphs and then compared, we apply the *k*-means clustering algorithm using the clustering criterion function shown in equation 3. Achieving a desired clustering of data is moreover a result of optimizing the clustering criterion function (Zhao and Karypis 2002). The trial clusterings were composed of clustering runs that cycled through all possible values of k from 2 to (n-2) and over all possible seed combinations so that we could find the best clustering in each level. This was computationally possible due to the small sample size (n=20) and a precomputed distance matrix.

Observations and Application

Understanding the data is key to understanding the results of clustering. We have taken data with a very high degree of dimensionality in the form of a graph and generated a distance metric between these graphs in this highly dimensional space. In some levels the space is so large between player graphs that, given the seed values based on the levels themselves, none of them appear to be close to any other. In this case the clusters become spread out. In other cases, the graphs do appear to be close and those can cluster together. In the third observed case, from some views that are so distant in space, all of the other players seem to be equally as close and so clump into one or more large clusters. Large clumps do not tell us much about the players in a level, and spread out clusters indicate that most of the data points are dissimilar. All players do not play the same, and especially not within the constraints of time being tracked at 10 Hz, so clumping should be avoided by the choice of a better clustering criterion or though level design. The data indicates that in the absence of complex geometry or situations in a level, time becomes the major distinguishing feature between players. The levels that produced clumps were simple in design, but because they often forced the player into a specific sequence of events and facilitated their rapid accomplishment, they actually caused the players to perform similarly.

More distinct player performance is observed in levels that increase complexity by presenting more choices in the path, less leading of the player, and more choices in time and traversal. The caveat for the designer is to not make them too complex or there will be too many possible paths, and the data will optimize on a spread clustering. In a small sample space of players and test levels of fair complexity, the individual player differences may not distribute into ideal clusters where $k = n/2$. The differences may be too disparate between players and will spread out. As the number of players increases, the clustering should become more distributed, or at least similar players should be correctly placed in the same clusters. The means to generate a representative set of human-consistent models in a virtual environment through the use of clustering algorithms can be provided by building a test environment with sufficient complexity, sufficient unique traversal paths and no player leading.

Generating a data set for comparison when there are an infinite number of combinations of choices in each level is difficult. Our human trial data set suffers from its small size, but is sufficient for clustering other similarly behaving humans. However, there may exist human players that may not be classified with the current set of players. What we present is an objective form of evaluating performance

against a particular set of humans and the means to test human-consistency within the boundaries of the known human models captured in the system.

We present the Time-Score metric as a means for identifying if a new player falls within the band of human performance. We also present a clustering metric for determining human-consistency and classification. Both metrics are based on a set of fixed procedures and the use of a trial human data set.

We previously presented our working definition of human-consistency as the clustering of player performance with groups of humans, and human performance consistency as performance within statistical norms—both based on human trial data. Using the presented Time-Score and Clustering metrics we have been able to evaluate our own agent system performance. Currently, the CAMS-DCA system was at its best able to complete 29 of the 100 test levels. The reflex-based agent completed 73.1% of its completed levels within human performance levels, and 15.4% were performed in a human-consistent manner. The combination reflex-random agent completed 69.0% of its completed levels within human performance levels, and 3.4% were performed in a human-consistent manner.

Conclusions

We hypothesized that actual or artificial player performance in first-person entertainment-based artificial environments can be evaluated objectively. As illustrated through the production of a Time-Score metric as an overall performance evaluation and a clustering metric for determination of human-consistency, we provided a means to evaluate the human-consistency of players in the Quake II environment. We first showed a general classification of system performance as compared to a defined set of human acceptable performance ranges and the ability to cluster a player's performance with those that played in a highly similar manner. This includes the evaluation of temporal performance as well as a spatial approach to sub-goal problem solving in those environments. The resulting clusters of players have a clear connection to performance in such environments and provide a means of classification based strictly on performance groups. These evaluations are conducted solely through numerical analysis in a defined process and do not contain any steps of subjective evaluation. Thus player performance in first-person entertainment-based artificial environments can be evaluated objectively.

We classify an agent's performance as being human-consistent by tracking their performance of movement and time between interaction feature points and comparing it in a clustering fashion to a set of human players under the same conditions. If an agent can be clustered in the same cluster, then we can infer that the agent played in a consistent manner as the other players in their cluster. Since those other players were human, we can say that the agent played in a human-consistent manner for that level.

Player performance can be represented to allow comparison between players of any type through the choice of a time-weighted graph representation between interactive feature points. This was chosen to reduce the size of information processing and ignore minor player positional noise and jitter.

There are still many more areas to be investigated. We need to continue work toward improved test levels and clustering criteria. We believe that this takes some important steps toward providing a means of objective player evaluation that may be useful across the modeling and simulation community.

References

Friedman, Menahem and Abraham Kandel. 1999. *Introduction to Pattern Recognition: Statistical, Structural, Neural, and Fuzzy Logic Approaches.* London: Imperial College Press.

Gonzalez, A. 1999. Validation of Human Behavioral Models. In *Proceedings of the 12th International Florida AI Research Society*, 489-493. Menlo Park, CA: AAAI Press.

Heinze, Clinton, Simon Goss, Torgny Josefsson, Kerry Bennett, Sam Waugh, Ian Lloyd, Graeme Murray, and John Oldfield. Interchanging Agents and Humans in Military Simulation. *AI Magazine* 23 (2): 37-47.

Kaminka, Gal A., Manuela M. Veloso, Steve Schaffer, Chris Sollitto, Rogelio Adobbati, Andrew N. Marshall, Andrew Scholer, and Shelia Tejada. 2002. GameBots: A Flexible Test Bed for Multiagent Team Research. *Communications of the ACM* 45 (1): 43-45.

Knauf, Rainer, Ilka Philippow, Avelino Gonzalez, and Klaus Jantke. 2001. The Character of Human Behavioral Representation and Its Impact on the Validation Issue. In *Proceedings of the 14th International Florida AI Research Society*, 635-639. Menlo Park: AAAI Press.

Laird, J. E. and M. van Lent. 2001. Human-level AI's Killer Application: Interactive Computer Games. *AI Magazine* 22(2):15-25.

Laird. J. E. 2001. Using a Computer Game to Develop Advanced AI. *Computer* 34(7):70-75.

Laird, John. 2002. Research in Human-Level AI Using Computer Games. *Communications of the ACM* 45 (1): 32-35.

Youngblood, G. Michael. 2002. Agent-based Simulated Cognitive Intelligence in a Real-time First-person Entertainment-based Artificial Environment. M.S. thesis, The University of Texas at Arlington.

Zhao, Ying and George Karypis. 2002. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. Technical Report, 02-022. Minneapolis: University of Minnesota.