Constructing an Intelligent Behavior Avatar in a Virtual World: A Self-Learning Model based on Reinforcement

Jui-Fa Chen+, Wei-Chuan Lin*, Hua-Sheng Bai+, Chia-Che Yang+, Hsiao-Chuan Chao+
+*Department of Information Engineering, TamKang University*
+*E-mail: alpha@mail.tku.edu.tw*
**Department of Information Technology, Tak-Ming College,*
**E-mail: wayne@takming.edu.tw*

Abstract

In this paper, a novel method for personal intelligent behavior avatar (IBA) is proposed to acquire autonomous behavior based on the interactions between user and smart objects in the virtual environment. In this method, the behavior decision model and the self-learning model are integrated by Bayesian Networks and reinforcement learning. The Bayesian Networks can treat interaction experiences using statistical processes, and the sureness of decision making is represented by certainty factors using stochastic reasoning. The reinforcement learning is implemented by learning experimentation or trial and error mechanisms to improve the performance of IBA through feedback. Therefore, the IBA makes a strategic decision that is approximated and appropriate to the user through the self-learning process by reinforcement learning. Finally, the feasibility of this method is investigated by imitating user's behavior and the results of self-learning process. The results of simulation show that the method is successful in imitating user's behavior and improving the performance of IBA..

Keywords: virtual environment, intelligent behavior avatar, Bayesian Networks, reinforcement learning.

1. Introduction

The intelligent agents have been used in many different domains in the past few years. Personal intelligent agent has become one of the main research interests of Artificial Intelligence community recently. How to construct the knowledge system of agent with personal characteristic is the most important issue. The final goals of this paper are listed as follows:

(1) Construct intelligent behavior avatar, which has personal characteristic, to imitate user's behavior efficiently and exist in a virtual world.
(2) Based on the user's behavior model, a behavior model is developed to improve the performance of intelligent behavior avatar through reinforcement learning.

In order to achieve these goals, it is necessary for the agent to acquire behavior intelligence such as the preferences of the user, a behavior strategy for the user, and the knowledge of each environment.

However in many researches, developers often give the agent innate knowledge [1] with the assumption that the agent acts in a well-known way. Hence, the intelligent behavior avatar (IBA) is proposed for simulating personal characteristic. In this method, the behavior decision model and the self-learning model are integrated using Bayesian Networks and reinforcement learning. Based on this method, the IBA makes a strategic decision that is approximated and appropriate to the user through the self-learning process by reinforcement learning.

The rest of the paper is organized as follows: Section 2 outlines the key related methodologies. Section 3 is the proposed system architecture. Section 4 is the experimental results. Section 5 is the conclusion and future work.

2. Related Methodologies

A. Bayesian Networks

Bayesian Networks (BN) [2] are directed acyclic graphs that are constructed by a set of variables coupled with a set of directed edges between variables. BN are very successful in reasoning between the variables via conditional probabilities. A typical BN is shown in Fig 2.1, and it consists of the following elements:

- A set of variables and a set of directed edges between variables.
- Each set contains a finite set of mutually exclusive states.
- The variables coupled with the directed edges construct a directed acyclic graph (DAG).
- Each node(A) with parents(B1,B2,…,Bn) has a conditional probability P(A|B1,…Bn).

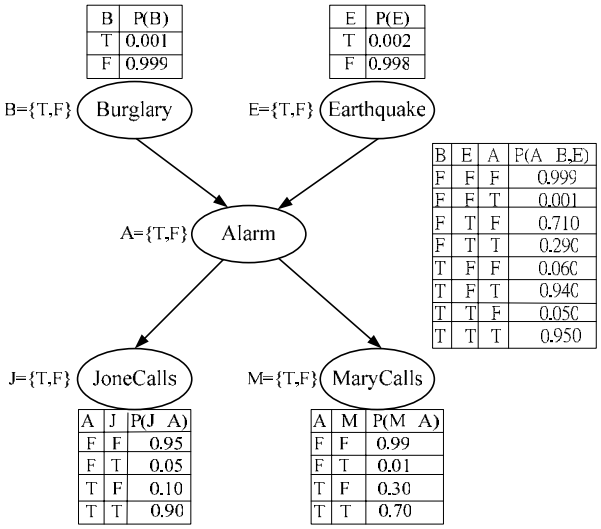

Fig 2.1 An example of BN

A BN serves as a model for a part of the virtual world, and the relations in the model reflect causal impacts between events [3] [4]. The reason for building the virtual world is to use to simulate the user for decision making. That is, probabilities provided by the network are used to support some kind of decision- making with personal characteristic [5].

The BN graph which is used to analyze the user's behavior model is listed as Fig 2.2.

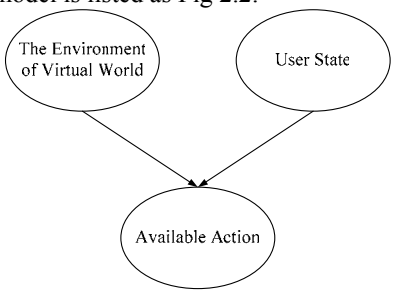

Fig 2.2 The BN graph of system

Due to the variables of virtual environment and user's states would influence the user's action; the BN is constructed as shown in Fig 2.2. The probability (preference) of each action is obtained from the conditional probability table which is generated by the "Available Action" node. The IBA uses the value of probability to make decisions.

B. Reinforcement learning

Reinforcement learning (RL) is a class of learning system. A learning system that learns by experimentation or trial-and-error mechanisms improves performance through feedback. The purpose of this section is to present the essential principles of RL and its relevance to the proposed learning algorithm [6] [7]. The proposed RL is based on learning from interaction within the environment. A basic reinforcement model is shown in Fig 2.3.

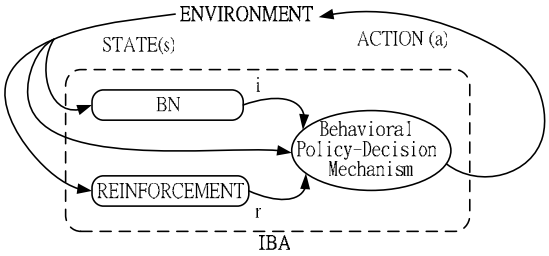

Fig.2.3 Basic Reinforcement Model

In the process of interaction, the IBA perceives the state(s) of the environment and receives an input(i) which is the probability provided by BN and the other input(r) which is the reward value provided by RL. The Behavioral Policy-Decision Mechanism of IBA calculates the weight of each action according to the input data and chooses the action with biggest weight to act.

3. System Architecture

As shown in Fig 3.1, there are two main modules in the proposed system architecture, the virtual environment and the IBA.

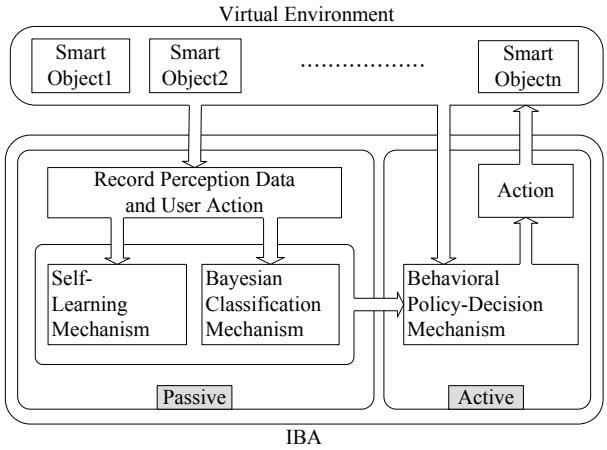

Fig 3.1 The system architecture

3.1 Virtual Environment

The virtual environment is a platform for IBA to execute and develop [8] [9]. There are two main kinds of smart objects (SO):

- Virtual human: The role which user or IBA plays in the virtual environment.
- Environment objects: Other objects in the virtual environment. Ex: weapon, health potion, money, etc...

The user's behavior model is analyzed according to the interaction information between Virtual human and Environment objects.

3.2 Intelligent Behavior Avatar (IBA)

When the user enters the virtual environment, the system configures an IBA to the user. The main purposes of IBA are imitating the user's behavior model, and improving the behavior performance by reinforcement learning. The IBA has two different executing durations such as passive and active duration. Passive duration is when the user is on-line, which the IBA is used to learn the behavior model of the user in the virtual environment. Active duration is when the user is off-line, which the IBA is used to simulating the user behavior to survive in the virtual environment. They are described as follows:

3.2.1 Passive IBA

- Duration: From a user on-line to the user off-line
- Mission: Record the perception data and the user's action to analyze the user behavior model. There are two mechanisms, that are used for procuring the mission in the passive duration, are explained as follows:

A. <u>Bayesian Classification Mechanism</u>

There are two steps to analyze the user's behavior model:

Step1:

Several major influence factors are considered and constructed several BNs according to each condition of SO occurrence. The probability (P_{BN}) of each action is computed from the conditional probability of each BN.

Step2:

The relation between actions and influence factors which have not been considered at step1 is analyzed to get the percentage value (α) of each action.

B. <u>Self-Learning Mechanism</u>

According to the major purpose of surviving in the virtual environment, the relation between the user's behavior model and the overall benefits is analyzed to adjust the weight of each action by reinforcement learning. The reward value (**RL**) of each action is obtained by reinforcement learning to feedback to simulate behavior of the user.

3.2.2 Active IBA

- Duration: From user off-line, and IBA activation to IBA termination.
- Mission: Survive in the virtual world according to the user's behavior model.

C. <u>Behavioral Policy-Decision Mechanism</u>

Several major influence factors are considered to construct BN, the probability should be adjusted by the other influence factors which have not been considered in BN. The formula is computed as follows:

Formula (3.1)

$$F_x(a_j) = \begin{cases} P_{BN}(a_j), & if\ x = 0 \\ (1-\alpha_x(a_j)) * F_{x-1}(a_j) + \alpha_x(a_j) * V_x(a_j), & if\ x > 0 \end{cases}$$

where the paramenters are explained as followed:

a_j : The number of available actions(j = 0,1,2,…,n) n. ∈ N

x : The number of influence factors which were not considered in BN(0,1,2,…,m) m ∈ *N*

$F_x(a_j)$: The weight of action(a_j).

$P_{BN}(a_j)$: The probability of action(a_j) from available action table.

$\alpha_x(a_j)$: The percentage value of action(a_j) for the xth unconsidered influence factor.

$V_x(a_j)$: The expected adjustment value of action(a_j) for the xth unconsidered influence factor

In formula (3.1), when x = 0, F0(aj)= PBN(aj). It means that all influence factors are considered, and the probability can be obtained from PBN. When x > 0, the probability should be adjusted by the other influence factors which have not been considered in BN, and the number of influence factors is m. The n is the n+1th action which means that there are n+1 actions needed to calculate.

After the formula is calculated, the IBA chooses the action (a_j) with the biggest $F_x(a_j)$ to act. If the IBA wants to improve the performance by reinforcement learning, the other formula is listed as follows:

Formula (3.2)

$$W(a_j) = F_x(a_j) + RL(a_j)$$

where the paramenters are explained as followed:

$RL(a_j)$: The reward value of action(a_j).

The IBA chooses the action (a_j) with the biggest $W(a_j)$ to act.

4. Experimental Results

4.1 Implementation

First, the kind of smart objects (SO) is defined in the virtual environment as shown in Table 4.1.

Table.4.1 The SO in virtual environment

SO Name	Attribute	Action(a_i)
Enemy	Grade, life	Attack (0)
Health Potion	Capacity	Eat (1)
Money	Occur or not	Get (2)

The rule in the virtual environment is that the player must beat the enemy successfully to gain experience for upgrade. In the process of attacking the enemy, the player's health status may decrease, so the player must eat health potion to maintain his health. If the player died or failed to attack, the experience will decrease.

The three mechanisms of IBA will be introduced in the following section.

4.1.1 Passive IBA

When the user is on-line in the virtual environment, the passive IBA records the perception data such as the environment variables, user states, and use actions …etc. in the "Record Table" and analyzes the user's behavior model. The entries of "Record Table" are shown in Table 4.2.

Table 4.2 The Record Table

Entry Name	Description	Value
Hp	The life of player	0~100
D	The grade of player	Integer
Eb	The life of enemy	0~100
E	The grade of enemy	Integer
Diff	E-D	Integer
H	The capacity of health potion	20, 100
M	The money exists(1) or not(0)	0, 1
De	The distance to enemy	1~10
Dh	The distance to health potion	1~10
Dm	The distance to money	1~10
Action	0:Attack, 1:Eat health potion, 2:Get money, 3:Ignore	0~3
Result	The result of attacking, 0:"Lose", 1:"Win"	0,1

A. **Bayesian Classification Mechanism**

Step 1: According to the occurrence condition of each SO, seven BNs are constructed without considering the "**distance**" factor. The variables of BN and seven BNs are constructed as in Table 4.3 and Fig 4.3.

Table 4.3 the variables of BNs

State (Hp)	0	1	2	3	4
The life of player	0～20	20～40	40～60	60～80	80～100

State (Eb)	0	1	2	3	4
The life of enemy	0～20	20～40	40～60	60～80	80～100

State (Diff)	-5	-4	-3	-2	-1	0	1	2	3	4	5
E-D	-5	-4	-3	-2	-1	0	1	2	3	4	5

State (Health potion)	1	2
Capacity	20	100

Step2: Analyze the relation between each action and the "distance" factor.

$\alpha(0)$ = P(De is smallest | Action(0)，Hp，Diff，Eb，H，M)
$\alpha(1)$ = P(Dh is smallest | Action(1)，Hp，Diff，Eb，H，M)
$\alpha(2)$ = P(Dm is smallest | Action(2)，Hp，Diff，Eb，H，M)

The value of α determines the influence of the "distance" factor on the user's behavior. Ex: When α is approximated to 1, the user often takes action on the closest object. When α is approximated to 0, user often takes no action on the closest object. When α is approximated to 0.5, the "distance" factor has little influence on the user.

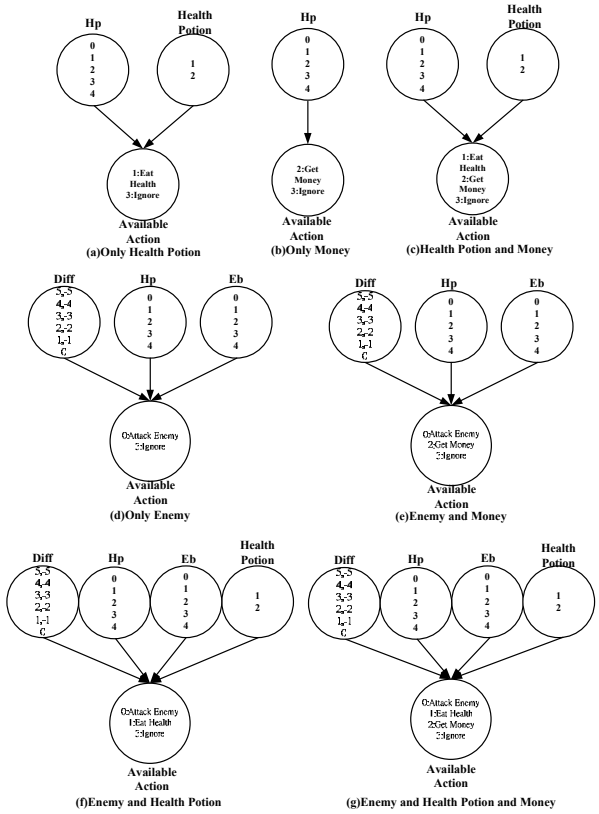

Fig 4.3 The seven BNs

B. **Self-Learning Mechanism**

The purpose of surviving in the virtual environment is beating the enemy successfully to gain experience for upgrade. Therefore, the "WIN" conditional probability (β) is calculated to adjust the "attack(0)" behavior and other actions remain unchanged.

β = P(Result(1)|Hp，Diff，Eb，Action(0))

The reward value of action(0) is computed as follows:

$RL(0) = 2 * \beta - 1 \ \ (-1 \le RL(0) \le 1)$

$$\begin{cases} if\ \beta > 0.5,\ then\ RL(0) > 0 & \text{//increase the desire to attack} \\ if\ \beta = 0.5,\ then\ RL(0) = 0 & \text{//remain unchanged} \\ if\ \beta < 0.5,\ then\ RL(0) < 0 & \text{//decrease the desire to attack} \end{cases}$$

The RL(1) = RL(2) = RL(3) =0

C. **Behavioral Policy-Decision Mechanism**

This mechanism is to calculate the weight of each action (a_j) and choose the action with biggest weight $W(a_j)$ to execute. The formula is listed as follows:

$$W(a_j) = F_x(a_j) + RL(a_j) = (1 - \alpha_x(a_j)) * P_{BN}(a_j) + \alpha_x(a_j) * V_x(a_j) + RL(a_j)$$

The value of $P_{BN}(a_j)$ and $\alpha(a_j)$ are obtained from Bayesian Classification Mechanism, and the value of $RL(a_j)$ is got from Self-Learning Mechanism. The $V(a_j)$ is the adjustment value of action (a_j) for ignoring the "distance" influence factor. To decrease the mileage of IBA for walking in the virtual environment, the concept of "fuzzification" is used. That is if the distance to the object(i) is smaller, the value of $V(a_j)$ is larger. If the distance to the object(i) is larger, the value of $V(a_j)$ is smaller. To maintain the equilibrium of $P_{BN}(a_j)$ and $V(a_j)$, the value of $V(a_j)$ is controlled between 0 and 1. The value of $V(a_j)$ is computed as follows:

$$V(a_i) = \frac{\text{MinDistanceToObject}(a_i, a_j, a_k)}{\text{DistanceToObject}(a_i)}$$

Ex: De=2, Dh=5, Dm=9, the
V(0)=2/2=1, V(1)=2/5=0.4 , V(2)=2/9=0.22

As the "Enemy" object is the nearest to IBA, the value of V(0) is the largest.

4.2 Simulation Results

There are two purposes of this simulation:

- Design IBA to imitate the user's behavior model efficiently.
- Design IBA to develop a better behavior model through reinforcement learning.

4.2.1 Imitate the user's behavior

As shown in Table 4.4, the user and IBA are given the same environment variables to choose actions and construct their own conditional probability table. While comparing the conditional probability table of the user and IBA, the difference between them is obvious. Only one compared result is shown in the conditional probability as in Fig 4.4.

Table 4.4 Enemy and Health Potion

				PLAYER BN f			***IBA BN f***		
HP	EB	DIFF	H	ATTACK	EATH	IGNORE	ATTACK	EATH	IGNORE
1	2	-2	20	0.4870	0.5130	0	0.5169	0.4831	0
1	3	-2	20	0.4951	0.5049	0	0.5298	0.4702	0
1	4	-2	20	0.5918	0.4082	0	0.4906	0.5094	0
1	0	-1	20	0.9046	0.0954	0	0.9064	0.0936	0
1	1	-1	20	0.9114	0.0886	0	0.9953	0.0047	0
1	2	-1	20	0.5240	0.4760	0	0.5093	0.4907	0
1	3	-1	20	0.5235	0.4765	0	0.5109	0.4891	0
1	4	-1	20	0.5387	0.4613	0	0.5455	0.4545	0
1	0	0	20	0.9861	0.0139	0	0.9817	0.0183	0
1	1	0	20	0.9883	0.0117	0	0.9855	0.0145	0

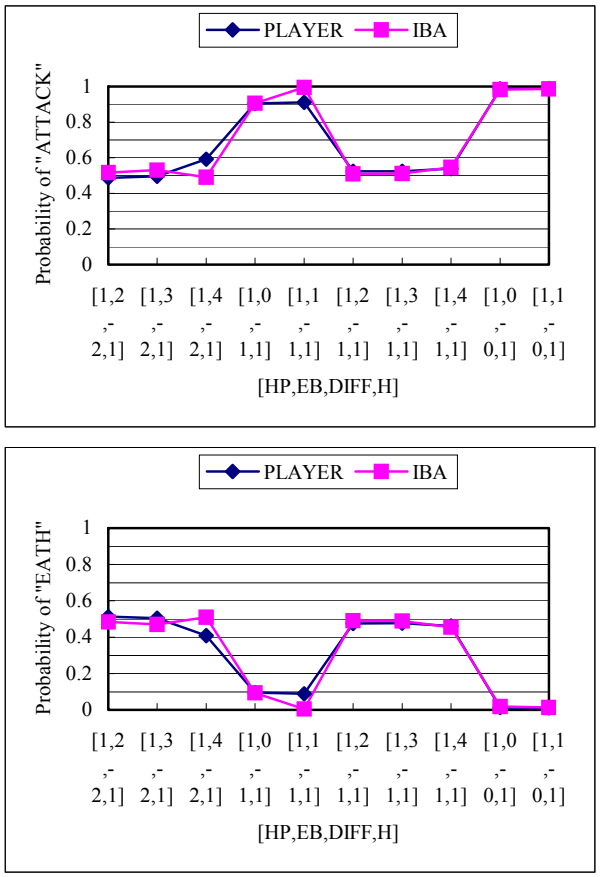

Fig 4.4 The comparison of probability

As shown in Table 4.4, the probability of IBA and the user are very similar. The range of inaccuracy is between 0.13 and -0.13. The reasons of causing inaccuracy are explained as follows:

1 The quantity of record table:
The fewer the record table is, the larger the inaccuracy will be.

2 The user's characteristic:
The more irregular the user's behavior is, the larger the inaccuracy will be.
3 The IBA Behavioral Policy-Decision Mechanism:
The preferences are adjusted by considering the "distance" factor. So it will cause inaccuracy.
4 The design of BN:
When the variables "Hp" and "Eb" are defined, the life range of each state is 20. If the range is larger, the inaccuracy is larger.

4.2.2 The performance of reinforcement learning

The same environment variables are given to the IBA and IBA_RL (with reinforcement learning) for choosing actions. The result is shown in Table 4.5.

Table 4.5 The result of IBA and IBA_RL

Environment variables	*10000*		*100000*		*600000*	
	IBA	**IBA_RL**	**IBA**	**IBA_RL**	**IBA**	**IBA_RL**
Grade	14	14	17	17	20	20
Experience	430440	463780	5534940	5882735	38948670	41309375
Number of attacking time	3862	3849	39036	38489	234352	231047
Number of Failure time	570	365	6066	3911	36438	23352
Failure percentage	14.76%	9.48%	15.54%	10.16%	15.55%	10.11%
Mileage	60170	68386	605502	688638	3536158	4140477

Compared IBA with IBA_RL, it can be seen the difference between them that are summary as follows:

- Experience: IBA_RL > IBA
- Failure percentage: IBA_RL< IBA
- Mileage: this result is focused on the "attack" behavior, and the mileage did not make more adjustment. Hence, it cannot guarantee that the value of mileage will be smaller by reinforcement learning.

From the experiment result, it can prove that the IBA can improve the performance by reinforcement learning.

5. Conclusions and Future Works

A novel method for personal IBA is proposed to simulate the user behavior in a virtual environment. In this method, the behavior decision model and the self-learning model are integrated using Bayesian Networks and reinforcement learning. Based on this method, the IBA makes a strategic decision not only approximated to the user, but also appropriate through the self-learning process by reinforcement learning.

The future works are listed as follows:

- The kinds of SO could be more complicated, and the IBA can cooperate with other players to finish the mission.
- The using of "Dead-Recon" algorithm to setup the threshold response of IBA to improve the performance of IBA.
- According to the user's behavior model, the range of RL could be set dynamically. In this way, the IBA makes a strategic decision not only approximated to the user, but also more appropriate through the self-learning process by reinforcement learning.

6. References

[1] Stuart J. Russell, Peter Norvig., "Artificial Intelligence: A Modern Approach (2nd Edition)", Prentice Hall, 2002.

[2] Finn V. Jensen, "Bayesian Networks and Decision Graphs", Springer-Verlag, New York, 2001.

[3] Richard E. Neapolitan., "Learning Bayesian Networks", Prentice Hall, 2003.

[4] T. lnamura, M. Inaba, and H. Inoue., "Integration Model of Learning Mechanism and Dialogue Strategy based on Stochastic Experience Representation using Bayesian Network," Proceedings of the 2000 IEEE International Workshop on Robot and Human Interactive Communication, Osaka, Japan, Sep., 2000, pp.247-252.

[5] F.Sahin, J.S.Bay, "Learning from experience using a decision- theoretic intelligent agent in multi-agent systems," Soft Computing in Industrial Applications, 2001. SMCia/01. Proceedings of the 2001 IEEE Mountain Workshop on , 25-27 June 2001,pp109 – 114.

[6] Richard S. Sutton, Andrew G. Barto.,"Reinforcement Learning," Cambridge, Mass. : MIT Press, 1998.

[7] S.Ramachandran, David S. Bree, "Learning action selection in autonomous agents," Thesis report, University of Manchester, 2001.

[8] Haw-Yun Hung, "Building an intelligent Behavior Avatar in a Virtual World," Thesis report, University of TamKang, Tansui, Taiwan, 2003.

[9] M. Kallmann, J. Monzani, A. Caicedo, and D. Thaimann, "ACE: A Platform for the Real Time Simulation of Virtual Human Agents," EGCAS'1100 - 11th Eurographics Workshop on Animation and Simulation, Interlaken, Switzerland, 2002, pp.1100.