Context-dependent Action Interpretation in Interactive Storytelling Games

Chung-Lun Lu
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan 30013, R.O.C.
chunglunlu@gmail.com

Von-Wun Soo
Institute of Information Systems and Applications
National Tsing Hua University
Hsinchu, Taiwan 30013, R.O.C.
soo@cs.nthu.edu.tw

Abstract—**In this paper, a framework of context-dependent behavior interpretation in interactive storytelling system is proposed. A user can act as one of the role characters in a story to interact with other virtual characters in the system. We implemented two levels of action interpretation: activity and behavior. A Microsoft Kinect sensor is used to acquire and recognize user's activities in terms of the information of its body joints that will be trained by a pre-learned model. Then, with multiple-context modeling and the recognized activities, a dynamic Bayesian network is adopted to disambiguate user's behaviors in terms of his intentional and subgoal structure.**

Keywords-interactive storytelling; behabior interpretation

I. INTRODUCTION

In interactive storytelling, it's important for users to play some active roles and interact with the storytelling system. Past researches, for example [1], allow a user to play the role as one of the virtual characters in the cast; the user is projected into the virtual world by mixing the user image into the virtual world on a screen, and then the system interprets the user's actions based on the user's utterance and gestures. However, in storytelling, due to the variations of a plot, the same user's actions can imply different meanings under different contexts. Without proper background contexts, it is difficult to interpret the behavior of a virtual agent due to the ambiguities of the actions. So in this paper, we propose a system that could recognize the user's activities and interpret their underlying intention using the background context information in the virtual environment. The Microsoft Kinect sensor is used to capture the user's action images in the real world and so that a 3D virtual character in the virtual world can be controlled and directed by the user via "natural" interactions. Once the activities are captured and recognized and with the aid of multiple context models, a dynamic Bayesian network is used to interpret the user's action further in terms of the intentional and goal structure of the user which we call it the behavior level of action interpretation.

The remaining of this paper is organized as follows: Section 2 describes some related works of interactions in interactive storytelling games. Section 3 describes the presented scenario. The proposed architecture is described in Section 4, and the detailed context definitions are discussed in Section 5. The method of recognizing user's activities is described in Section 6 and Section 7 describes the method of interpreting user's behavior. Finally, we summarize the presented work in Section 8.

II. RELATED WORK

In 2001, Charles, Mead, and Cavazza [2] proposed a system that uses an unreasonable concept to interactive with the storytelling system, they called it user intervention. Instead of directly interacting with the virtual characters, users play a role like a god and hide important virtual objects, which the virtual character will look for. The missing of important virtual objects forces the system to re-plan in order for the virtual character to achieve his final goal. In 2004 [1], they improved the way of interaction; a user is allowed to play the role as one of the virtual characters in the cast. The user's actions are interpreted based on the user's utterance and gestures; however, the action interpretation in their work didn't take the variations of contexts into account. In 2010, Doirado and Martinho [3] proposed a system that allows a user to interact with the virtual world and a virtual observer, a virtual dog and detects the user intentions while the user moves; however, in their work, the system achieves the interpretation based on the distance measurement only. In this paper, we aim to interpret the user's actions based on higher level information, that is, the contexts.

III. DESCRIPTION OF THE SCENARIO

The scenario of the story presented in this paper is based on a famous Japanese detective comic. Phantom Thief Kaito, a well-known thief, using a fake name, Sot, to work as a butler in the house of Duke Edward. One day, Duke Edward found that one of his antique paintings was missing, so he hired Conan, a brilliant detective, to investigate this crime. After a few days, Conan found that there were three suspects involved, the butler Sot, the female servant Susan, and the guard Alex. The main theme of the story scenario is for the major character, the detective Conan, to investigate this criminal case and find out the actual thief.

Fig. 7 shows an example of the subgoal structure that may be used by the detective. The main goal of the detective is to find out the thief and the detective has two sub-goals, collect evidence information ("Collect Information" goal) and conclude the criminal case ("Conclude Case" goal). To collect sufficient evidence information, the detective needs to interrogate the three suspects ("Interrogate Suspects" goal) and examine the house ("Examine House" goal) to collect

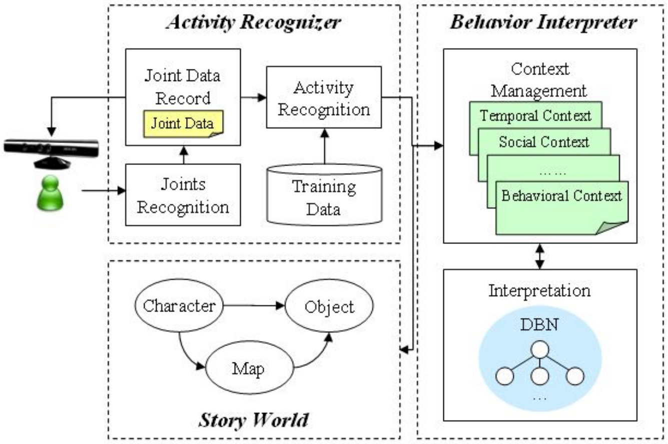

Figure 1. System architecture proposed in the paper.

objective evidence. In this paper, we assume the user play the role of Conan, the detective, and therefore, the user's activities will be captured via a camera and interpreted by the system.

IV. System Architecture

The architecture of the proposed interactive storytelling system is shown as Fig. 1. The system consists of three modules: the activity recognizer, the behavior interpreter, and the story world. The functions of each module are described as follows:

- Activity Recognizer

The activity recognizer generates the recognized activity for the behavior interpreter. The user activity is captured via Microsoft Kinect sensor, and then Microsoft Kinect SDK [4] is used to recognize the joints of user's body. The activity recognizer records the joint data until a pre-defined activity is detected and recognized. The output of this module, the recognized activity, is then sent to the behavior interpreter. In this paper a behavior is defined by an activity with semantic meaning in terms of goal and intention.

- Behavior Interpreter

The behavior interpreter manages the context change according to the story plot and the recognized activities and interprets the recognized activities based on the context information. The detailed interpretation scheme is discussed in Section 6.

- Story World

The story world in this paper is implemented using Unreal Development Kit (UDK) [8]. A virtual world map is implemented that has two rooms, a lobby and a small kitchen in the house of Duke Edward, shown as Fig. 2. The virtual characters are implemented using Autodesk Maya [5] and then imported into Unreal Development Kit, as shown in Fig. 3.

V. Context Definition

The contexts used in this work are classified into five categories: temporal context, spatial context, social context, emotional context, and behavioral context. Following is the detailed definition for each type of context.

- *Temporal context* and *spatial context*, including *state* and *location*, evolves according to the storyline. *State* is defined as the node in the hierarchical task network, as in Fig. 7, and *location* is the place where the detective is, a lobby or a small kitchen in the house of Duke Edward.
- *Social context* is defined by *role* and *relation. Role* is a set of pairs, {(*name*, *role*), …}, that denotes each character's social role and relations in a storyline; for an instance, {(Conan, detective), (Sot, butler)} means the Conan is a detective and Sot is a butler. *Relation* is a three-element set, {(*name*, *name'*, *relation*), that describes the social relation between characters; for an instance, {(Sot, Alex, colleague)} means Alex is Sot's colleague.
- *Emotional context denotes* the feeling state of characters; the *feeling* of a virtual character is indicated as a set of pairs, {(name, *feeling state*), …}, and the value of *feeling state* is one of the eight primary emotions proposed by R. Plutchick [7] plus the neutral emotion. The eight primary emotions are joy, sadness, fear, anger, surprise, anticipation, trust, and disgust. In this work, the emotional context is evolved while the system recognizes a user behavior successfully and its value is designed in the background story beforehand.
- *Behavioral context* identifies the activity of a human-played character. This context is composed of recognized activities and its target character or object of this activity, (*activity*, *target*). Fig. 4 shows an example of contexts.

Figure 2. A scenario scence of virtual world.

Figure 3. Virtual characters are imported into Unreal engine.

Spatial/Temporal Context:
- *location*: lobby
- *time:* evening 9 PM

Social Context:
- *role*: {(Conan, detective), (Sot, butler), (Susan, Servant), (Alex, guard)}
- *relation*: {(Conan, Sot, authority), (Conan, Susan, authority), (Conan, Alex, authority), (Sot, Susan, colleague), (Sot, Alex, colleague), (Susan, Alex, colleague)}

Emotional Context:
- *feeling*: {(Conan, neutral), (Sot, neutral), (Susan, neutral), (Alex, neutral)}

Behavior Context:
- *activity*: approaching
- *target*: Sot

Figure 4. An instance of a context.

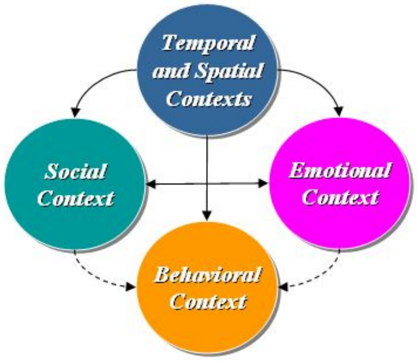

Figure 5. Relation among contexts.

Temporal and spatial contexts may be changed by the storyline and this change may result in different social and emotional contexts; for instance, once the detective knows that Sot is the thief who stole the painting, the main social role of Sot is changed from a butler to a thief. The change of main *social role* may cause *feeling* state to evolve; after the detective figures the crime out, the feeling state of the thief is changed to fear. Behavioral context differs from other contexts; it can be viewed as an independent one since its value is triggered by the user's activities.

Figure 5 shows the relations between different types of contexts. The changes of temporal or spatial context may trigger the changes of social, behavioral, or emotional context. The social context and the emotional context can influence each other. Under some situations, the behavioral context may be influenced by social or emotional context; however, in this paper, the behavioral context is triggered by the user's activities, so dotted line is used to represent that the change may not happen in this work.

VI. Recognizing User Activity

To recognize the user activity, three steps are necessary. First, body joints of a human need to be captured and recognized. Microsoft Kinect sensor is used to capture the body joints and NIUI is used to connect Kinect sensor and UDK; NIUI stands for OpenNI/Kinect API for UDK [6]. The result of connecting Kinect and UDK is shown as Fig. 6. We train the recognizer using machine learning techniques, for example, boosting or SVM. Finally, the model is used to recognize the activities and the recognized result is then sent to the behavior interpreter as an input of behavior context.

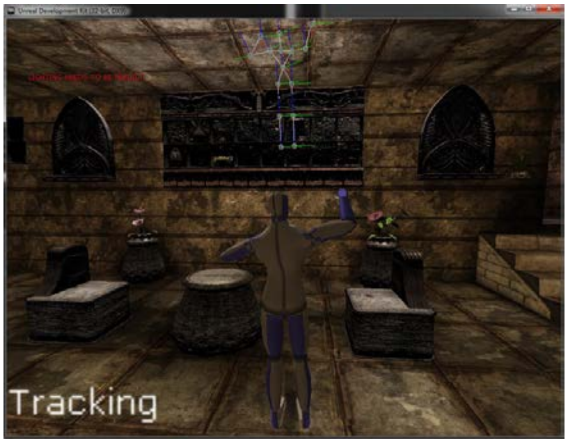

Figure 6. Result of connecting Microsoft Kinect and UDK.

VII. Interpreting User Behavior

The same user's activities can imply different intention under different background contexts. For example, if the detective approaches the butler at the investigating stage, it is more likely the detective wants to interrogate the suspects; if the detective approaches the butler at the case conclusion stage, it means the detective already knows the butler is the thief who stole the painting and he is trying to catch the butler.

Bayesian network has the advantage that each node's conditional probability distribution can easily be estimated, but the traditional Bayesian network cannot handle temporal data, that is, the new coming data cannot have any contribution to update the model. In interactive storytelling, contexts change over time, so in order to interpret the user behavior at anytime correctly, a dynamic Bayesian network is more appropriate than traditional Bayesian network. To model a dynamic Bayesian network, four parameters need to be defined:

- **Hidden state**, Q_t, specifies various context states we want to interpret; in our contexts, hidden states are defined by a set of context states. Based on the contexts, the system could interpret the user's activities in terms of the goal states of the detective defined in Fig. 7.
- **Transition model** is represented by the transition probability distribution functions (pdfs), $P(Q_t|Q_{t-1}, A_t)$; Q_t and A_t are the hidden state and action at time t respectively.
- **Observation model**, $P(Y_t|Q_t)$, specifies the dependency of the observation nodes according to the hidden nodes at time t; Y_t is the observations at

time *t*. In this work, observations are defined by a subset of contexts; that is, Y_t is the sensed contexts at time *t*.

- **Initial state distribution**, $P(Q_0)$, represents the probability distribution in the beginning of the story.

The dynamic Bayesian network is to find the probability distribution of the corresponding hidden states given a set of observations at time t, as in (1).

$$P(q_t \mid y_t) = r_t \Big/ \sum\nolimits_{q_t} r_t \ . \qquad (1)$$

where q_t is a set of t consecutive observations, y_t is a set of the corresponding hidden states, and the equation of r_t is in (2).

$$r_t = P(y_t \mid q_t)\cdot \sum\nolimits_{q_{t-1}} \big(P(q_t \mid q_{t-1}, a_t)\cdot P(q_{t-1})\big). \qquad (2)$$

The user behavior interpretation may face the problem of ambiguity; to reduce the ambiguity, the hidden node q_t with the highest probabilities in the dynamic Bayesian network is chosen as the most appropriate interpretation.

VIII. CONCLUSION AND FUTURE WORK

In this paper, a framework of context-dependent behavior interpretation in interactive storytelling system is proposed. This paper aims to propose a new way to interact with the storytelling system and adds the context-dependent behavior interpretation into interactive storytelling. Thanks to the release of the Microsoft Kinect, now it's convenient for researchers to implement an interactive system with body movement. Also, dynamic Bayesian network allows the changes of the plots to be triggered by a user who acts as one character of the cast and can update the probability distribution of states under various context change and thus support the reasoning backward and forward among hidden states, actions and observations along a sequence of activities; it supports the system in interpreting the user behavior at certain accuracy if the model parameter information is collected to some extent. Hence, the proposed system has high feasibility to provide a new interactive experience in interactive storytelling.

ACKNOWLEDGMENT

This research was supported in part by the National Science Council of Taiwan, R.O.C. under Grant NSC 99-2221-E-007-090-MY3.

REFERENCES

[1] M. Cavazza, F. Charles, S. J. Mead, O. Martin, X. Marichal, and A. Nandi, "Multimodal Acting in Mixed Reality Interactive Storytelling", IEEE Multimedia, July-Septemver 2004, Vol. 11, Issue 3, pp. 30-39.

[2] F. Charles, S.J. Mead, and M. Cavazza, "User Intervention in Virtual Interactive Storytelling", Virtual Reality International Conference, 2001.

[3] E. Doirado and C. Martinho, "I Mean It! Detecting User Intentions to Create Believable Behaviour for Virtual Agents in Games", Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems, 2010.

[4] Kinect for Windows SDK, http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/

[5] Maya, http://usa.autodesk.com/maya/

[6] OpenNI/Kinect API for UDK (NIUI), http://forums.epicgames.com/showthread.php?t=765636

[7] R. Plutchik, "The Nature of Emotions", American Scientist, 2001, pp. 244-251.

[8] Unreal Development Kit, http://www.udk.com/

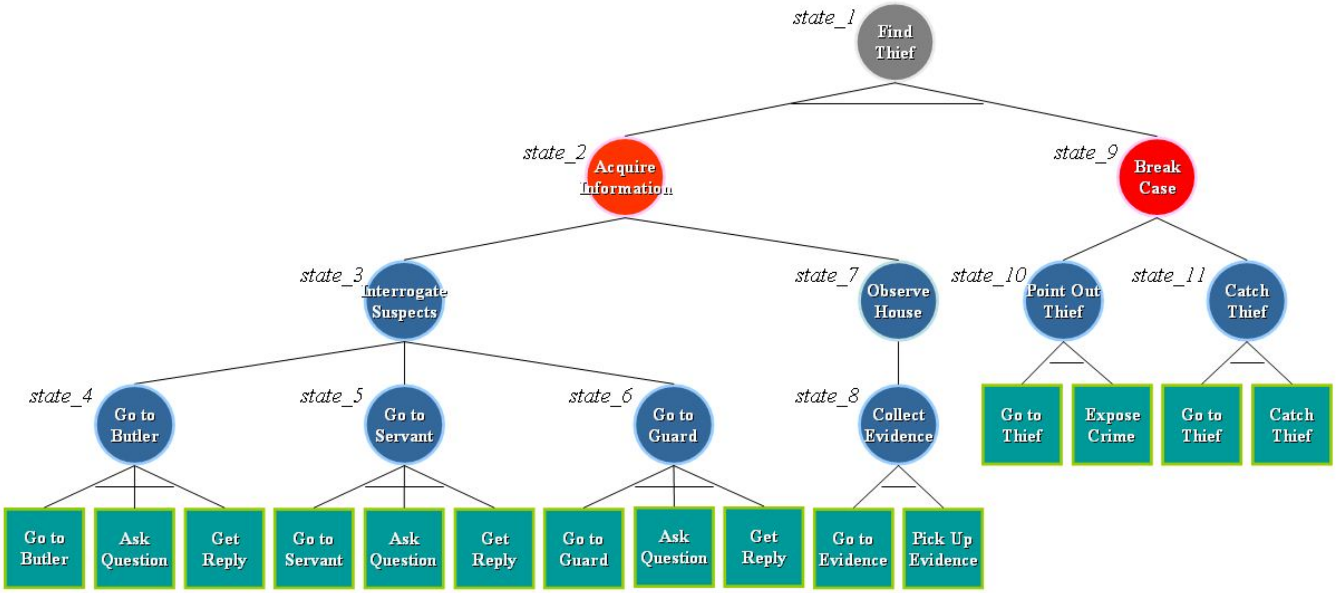

Figure 7. Example scenario of the detective storyline presented in this paper.