ABSTRACT

Creating agents which utilize natural verbal and non-verbal communication is an appropriate goal for many researchers involved in human-computer interaction. Using these types of agents enhances their capabilities as a communication tool for teaching humans inside a virtual environment. This paper describes how Herbert Clark’s theory of joint activities can be applied to agents, extending it to a domain in which communication is weighted towards the non-verbal, specifically body expressions. The major framework for this implementation is the joint action ladder, in which a communicative act is checked at several levels before it can be understood and responded to by the receiver of the act. A proposed environment has been created in the form of a basketball game, where a user can interact with his team through natural body movements. Through this type of interaction, the user has the potential to learn gestural expressions through observation as well as less explicit concepts such as the internal state of agents. Several avenues of research that can be followed through the use of the basketball game are also described.

Categories and Subject Descriptors
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—artificial, augmented, and virtual realities

1. INTRODUCTION

Together with verbal utterances, body expression is one of the most used modalities used in face-to-face communication between humans. However, research into this modality in the virtual world has been relatively sparse compared to that of verbal behavior and other non-verbal behavior such as facial expressions. The primary focus of this research is to create autonomous agents which are able to react to and express body expression behavior in a way which is natural for users. Applications of this work will be useful in many fields, and will provide insights to how agents with intelligent body expressions can be constructed.

Currently, there are a large number of methods to integrate body expression into intelligent agents including use of markup languages [13] and parameterization [10]. While these are sophisticated techniques, the agents which they are used for appear to be mainly used for one-to-one communication. Usually this communication focuses heavily on verbal components while non-verbal components are used in a supplementary manner. Furthermore, the agents do not consider personal space, which can provide contextual information to the user. This research tries to address these issues by viewing the problem using a different conceptual framework.

In regards to intelligent agents’ non-verbal behavior, one theory that has been somewhat overlooked is Herbert Clark’s theory of joint activities and joint actions [3]. In this theory, Clark argues that communication between people involves individuals trying to achieve goals by engaging in actions which require their co-ordination. Adequate communication cannot occur without engaging in these actions. While Clark’s work is referenced extensively in the social [6] and cognitive sciences [9], there has been little attempt to convert this theory into a suitable format for use in virtual environments. Some papers which have utilized this theory include [12] and [8], which gives an overview of the use of Clark’s theories applied to humans communicating online. However, these focus on verbal communication, as it is this domain in which Clark’s theories are tuned towards. The use of the theory as a basis for intelligent agents with body expression modality is one main goal of this work. It is intended to both provide virtual agent designers with a tool for generating natural body communication, and to discover insights on human-agent interaction with body expressions. Clark’s theories by themselves may not be adequate in their current form to fully explain body expression communication. Many of the examples given utilize verbal utterances, and these are no doubt important. The context of this research is slightly different, as verbal behavior is not the focus. Joint activity theory is ideal for describing face-to-face communication and situations where both verbal and non-verbal modalities can be used freely. An extension of the model might be necessary to handle contexts in which communication is weighted towards one modality. Discussion on issues surrounding this extension is given later in the paper.

Another avenue of study is an investigation into the effect of virtual gestures have on learning. This not only includes learning about the characteristics of the virtual agents, but...
also the semantics of body expressions themselves. For example, during a sports game, there are many different body expressions used by players, which can have specific meanings for their team or simply used to communicate an internal state. In team sports such as football, pointing can be used to order a team-mate into a certain position. Baseball catchers use special signals to indicate which type of pitch should be thrown. Through observation, the audience is generally able to comprehend the meaning of some of this behavior. For virtual characters, similar behaviors should be integrated to make communication smoother for the user. Through this, an analysis of how these behaviors affect learning can be identified.

The long-term goal is to integrate body expression and verbal communication by utilizing Clark’s theories with either or both modalities. In this paper, the current focus is on body expressions. Of course, much research has been designed to address the issue of synchronization with verbal utterances and gestural [1] [2]. However, in these cases the agents express a great amount of verbal communication compared to full body expressions, where gestures are used to supplement the verbal utterances. Agent communication utilizing a lower verbal/non-verbal modality ratio is less common, but these scenarios exist in the real world, such as in the aforementioned sporting events. This paper’s goal is to provide an overview of the framework used to create suitable agents by using Clark’s joint activities and actions. The next section provides an overview of the theoretical foundations of joint activity theory. Section 3 provides a concrete example of implementation of basketball, the target scenario of this paper. The proposed hardware is briefly described in Section 4 before general discussions and a conclusion of the paper.

2. JOINT ACTIVITY THEORY

It should be noted that in much of Clark’s work, the intent is to apply the framework to verbal, rather than non-verbal language use. The framework is also a much better fit - the sequential nature of verbal communication against potentially parallel non-verbal communication. For now, consideration should be on the original theory, while arguments for adaptation and extension will be made later in this paper.

The notion of language being comprised of joint actions was presented by Herbert Clark in his seminal work Using Language [3]. A brief summary of the theoretical foundations is given in this section. At the general level, people engage in communicative acts in order to achieve certain goals. These goals are not necessarily shared, such as people involved in playing chess. Such activities involving two or more people are termed joint activities and also contain roles for participants. Some aspects of joint activities, which will be focused on in this research, include joint actions, joint action ladders and adjacency pairs.

Common ground is a related feature of joint activities which deserves attention. It is this which can explain in some way how humans learn by engaging in communication. Common ground is essentially the joint knowledge or beliefs which exist between two people. While Clark’s examination of this phenomena is quite thorough, for brevity only the most necessary features will be highlighted. Common ground can be justified through evidence, in that it should be clear as to why an individual believes that they share some common knowledge with another. Furthermore, common ground occurs in special communities, such as the medical field, and is built up in a number of ways, such as through a shared experience. This concept is key for learning - each piece of common ground that is created between two people represents a new piece of knowledge.

2.1 Joint Actions

Clark asserts that joint activities are advanced through joint actions. These can be seen as the smallest unit of activity which requires some sort of co-ordination with another. For example, within the simple joint activity of shaking hands, a participant may engage in several joint actions, such as extending their hand, grasping the hand of the other person, shaking the hand, and releasing their grip. Clearly, a joint action involves the co-ordination of individual actions. Clark denotes a joint action with two participants A and B as:

\[
\text{joint}[\text{part}_1(A), \text{part}_2(B)]
\]

with \text{part}_1 representing a participant’s individual action \(x\). It is these joint actions which drive the undertaking and completion of joint activities. Joint actions may appear to be trivial, as we are able to execute them daily with little effort. However, they actually involve many different types of interactions. These interactions can be as simple as eye contact or a monosyllabic utterance.

2.2 The Joint Action Ladder

For use in a computational model, the most important structure proposed by Clark is that of the joint action ladder. This ladder consists of levels in which one joint action is performed by the participants. Each side of the ladder represents an individual action ladder. The participants must complete the joint actions at each level (starting from 1) to move onto the next level. A generalized joint action ladder is presented in Table 1.

Each of the joint actions at these levels has importance in the advancement of joint activities. A brief description of these levels is given below. In each of these descriptions, A is seen as the ‘speaker’ of the action, with B the addressee. While this terminology is more aligned with verbal utterances, the concept of sending and receiving can just as easily be applied to body expression behavior. In order to ground these theories in a concrete situation, one type of joint activity will be explained - that of A and B engaging in a handshake.

2.2.1 Behaviors

The lowest level of the joint action ladder is the executing of a behavior by A and the attending to the behavior by B. In other words, both participants must co-ordinate their actions so that they are at attention with each other. In the case of a handshake, this may arise as A begins to extend their hand towards B while B gazes at A. This joint action indicates that A has executed some expression which B acknowledges they should respond to.

2.2.2 Presentation

The second level involves presentation of a behavior, such as when A presents a signal which is to be identified by B. The delivery of the signal is the focus here, because A wishes to ensure that B understands what Clark terms the primary presentation of the signal. This component allows B to identify the signal presented. For example, when A
extends their arm towards B, the primary presentation is this signal.

2.2.3 Signal Meanings

Once attention has been established and A has presented a signal for B, the next joint action involves the meaning and recognition of signals. If B has knowledge of the meaning of A’s signal, he may then react appropriately. The types of signals and their interpretation is a vast field of study in its own right. Many signal interpretations are to do with the analysis of gesture and the work of researchers such as Kendon [4]. In the handshake example, B must recognize that the extension of the arm means that A is wanting to greet them by shaking their hand. For humans, the correct identification of this signal is largely based on handshaking being a cultural norm.

2.2.4 Joint Projects

Joint projects occupy the highest level of the joint action ladder. They occur when A proposes a joint project for B, and B accepts this joint project. In the handshake example, the joint project is the proposal given by A that B should greet them. The previous levels in the ladder consisted of A getting B’s attention, presenting the signal for B to shake their hand, and B identifying that signal as an offer to shake their hand. The uptake of this proposal by B indicates a joint project has occurred. Joint projects can be of any size. The example situation can be considered as a minimal joint project, however A and B introducing themselves (which includes handshaking) can be seen to be another larger joint project.

2.3 Adjacency Pairs

Clearly each level in a participant’s action ladder is not explicitly expressed in a sequential manner. For example, participant A extending their hand and participant B indicates both that B has understood the signal (level 3) and has undertaken the proposal (level 4). Clark explains the action-responses at these levels as adjacency pairs. The key property which differentiates adjacency pairs from a mere action-response mechanism is that they indicate the uptake of a joint project from A by B.

Adjacency pairs are the joint actions which are identifiable through body expressions. Like the handshake example, they are commonly used to show the completion of a joint project. They contain two expressive parts - an instigating action from A and a response from B. They are also evidence for what Clark terms joint construals. These construals are evidence that B understands the intended meaning of A’s action and will uptake the joint project.

3. MODELING BASKETBALL WITH JOINT ACTION THEORY

From the theories presented in Section 2, it is clear that there is much to consider, even for trivial body expressions. A generalized agent architecture will be described in order to create an agent model which draws from the theoretical frameworks. The major mechanism to be implemented is the joint action ladder. Through this, joint actions can be created which advance joint activities.

3.1 The Target Scenario

The scenario to be chosen for implementation carries great importance. Recall that the type of agent to be implemented will primarily use body expression over verbal utterances. It must also ensure that the user can observe and interpret this body expression as well as provide impetus for expressing this behavior. Furthermore, the scenario should have some grounding in reality. An unrealistic scenario may cause the user to feel as if they are merely taking part in a game world, where body expressions are either do not arise naturally, or worse, are not required at all.

In order to address these concerns, a simulation of a real world situation is being developed, that of a pickup basketball game. The game is so far being implemented using teams of three players to reduce complexity. It is assumed that most people are familiar with the objective and rules of basketball and have played the game using a computer, if not in real life. The user should be able to interact with their agent partners to create a situation of understanding and intuition which can be found in real-world sports teams.

The nature of basketball as a game makes it appropriate for analyzing body expression. Firstly, the game requires many forms of communication, not just one-to-one but one-to-many. These communications can come in a variety of forms. Verbal utterances may be used to direct players, however this presents a problem as these are also received by members of the opposing team who can then form a counter-strategy. Additionally, in a fast-paced sport such as basketball, complex strategies simply cannot be executed through utterances alone. What is more likely is players looking towards non-verbal cues in order to determine the next course of action. These clues can be the state of the game itself (i.e. spatial positions of players on the court), or signals performed by players in the game.

The type of signals used by players in basketball are varied. They can be as simple as looking towards a team-mate to indicate a pass, to a specific gesture indicating a certain play is to be executed. Within these extremes, there are many situational gestures which may be executed at various points in the game and ascribed meaning. It should also be noted that while simple signals such as passing are understandable to a majority of people, complex signals may be team-specific and therefore impossible for those not within the team to interpret. Initially, simple signals will be the focus.

Compared to commercial implementations of basketball,
the role of the user in the game differs greatly. In contemporary games the user can take control of any team-mate and is responsible for all passes and shooting behavior, while in this research the participant is restricted to their own character. Furthermore, the proposed view is first-person or third-person with a tracking camera behind the user’s avatar. Compare this with games where the player is able to view the entire court, which is unrealistic but provides greater spatial awareness.

Basketball, like many team sports, is itself a joint activity. Within the joint project of playing basketball, there are a number of sub-projects, including getting the ball into the opposition’s hoop and stopping the other team from scoring. At a more detailed level, joint projects include passing and catching the ball, or running certain plays. From now, the communicative act which will be focused on is that of simply passing the ball to a team-mate. Participant A is denoted as the passer, while participant B is the receiver. In this scenario the user can take the role of either participant.

3.2 Animated Behaviors as Adjacency Pairs

An adjacency pair for passing a ball is required. There needs to be an expression from A which indicates they are proposing a pass to B. Likewise, in response to this expression, B also needs to provide some public evidence that they have accepted A’s proposal of passing. These expressions are basically gestures generated by the graphics engine. Once these have occurred, the actual passing of the ball can occur. Theoretically any behavior can be used, but obviously gestures which are used in the real world are more appropriate. These selected signals are:

- **Participant A action**: Ball held in two hands at chest level, torso faced towards participant B.

- **Participant B response**: Both hands raised in front of chest, head facing participant A.

The end result of a completed joint project starts with A’s action and ends with B’s response. Of course, this is not the only adjacency pair required. The pass itself needs to be performed. In this case, a different adjacency pair could be as follows.

- **Participant A action**: Ball thrust by arms towards participant B.

- **Participant B response**: Ball is caught.

For clarity, assume that the joint project consists of the first adjacency pair. Although obviously important, communication within the latter pair is more weighted towards motor skills and individual co-ordination.

3.3 Modeling Joint Action Ladders

How can joint action ladders be implemented in intelligent agents? It is clear that in order to develop a similar model, each level of the joint action ladder must be accounted for. A joint action can be created at each level of the ladder and checked to see whether the next level in the ladder is satisfied. Recall that every joint action at a level lower than the current joint action must be completed before any processing of the current level can proceed. A suitable abstract structure for this behavior can be in the form of a state diagram. The diagram itself represents a specific action or joint project to be completed. Each state represents a level of the joint action ladder. For example, the first state for passing a ball is perception. After the initial movement by A, the test is whether or not B has noticed the movement. Transitions between states occur when these conditions are satisfied.

Using this kind of structure is ideal for showing the process flow between levels. Implementation is quite simple, and the transitions between each state can be made clear for any kind of joint activity. Furthermore, transitions to states in which joint actions are incomplete (described later in this section) can be introduced fairly easily. The process allows both proactive and reactive behavior from agents. Specific system activities will now be proposed for each level of the joint action ladder, as well as the type of information required for agents who utilize this model.

3.3.1 Activities at Each Level

The next issue is specifying exactly what occurs at each level of the ladder. In particular, which part of the system is responsible for ensuring whether a level has been fulfilled? In reality, humans do not make a systematic check of each of these levels, however this is possible in an agent system. Assume for now that the agent is in the role of the receiver of the pass, and take the example of the first level in the action ladder, that of perception. Clearly, all that is required for an agent is to perceive some sort of movement or behavior. Input sensors in a system make this process fairly trivial. Agent B’s internal mechanism for this action proceeds as follows. Firstly, B attends to A’s action. How does B show this attendance? One common way is to make eye contact with A, which can be shown by turning to face A. This action gives A confirmation that B has perceived them.

The second and third levels are where the signals are identified and ascribed meaning by the system. For this level, a recognition system needs to be applied in which gestures can be mapped to their abstract meanings. This is one of the most complex operations from a technical point of view. It is probably more important to establish a few gestures which are able to be recognized robustly, rather than attempting to cover all possible gesture types. In this stage, B checks their internal knowledge base to see whether they can determine A’s movement as having some meaning. Ideally, B recognizes A’s movement as being ‘looking for someone to pass to’. If B cannot recognize the movement exactly, they may make either a ‘best guess’ or create a new joint project to clarify. Obviously some sort of initial repository of gestures is required, comprising of at least a minimum gesture set. This minimum set is domain-specific and is required for an agent to be able to interact satisfactorily in a scenario. The actual content of this minimum gesture set is a target of this research.

At the final level, the agent must use the information gained from the previous levels to make a decision about their uptake of the joint project. Not only this, but they must indicate this to A. At this level, the agent’s belief system must be invoked to assess how uptake of the project will affect the completion of its goals. Once this decision has been made, the graphics engine provides body expressions as an output representation in the environment. In the passing situation, B has recognized A’s attention to pass. B checks against their internal beliefs, such as whether a pass is a valid action in the context of the game. If this is satisfied, the joint action and thus joint project is completed.
types is that they provide a method for identifying justifies their use over simple action-response mechanisms, through a refusal action.

Upon which some visual evidence will also have to be applied. On the other hand, B may opt to reject the joint project, through executing the second half of the adjacency pair.

Sequence

Figure 1: Basic state diagram of a joint action ladder sequence

through executing the second half of the adjacency pair. On the other hand, B may opt to reject the joint project, upon which some visual evidence will also have to be applied through a refusal action.

3.4 Incomplete Joint Actions

One powerful aspect of joint action ladders, and one which justifies their use over simple action-response mechanisms, is that they provide a method for identifying types of miscommunication from incomplete joint actions. If an agent can identify the source of this miscommunication, they may have some strategy to remedy the situation. Furthermore, these types of incompletions can be explicitly represented and modified by system designers. Even in the case of passing a basketball, there are several scenarios in which this seemingly basic joint activity can yield a breakdown of communication. By referring to the joint action ladder these may be identified.

Take the case of the first level of the joint action ladder, it might be possible that B is not focusing attention on A, so that A receives no signal that they can pass to B. At the second level, perhaps A does not execute their passing signal correctly, in which case B gives no reaction or is confused. The meaning of A’s passing signal might also not be understood by B. Finally, B might in fact understand A’s intention to pass, but refuse to co-operate for some reason. All these scenarios are where joint actions are not completed. While the outcomes may be similar, the reasons for the incompletions can be varied.

The reaction to these incompletions also varies. In some cases, A may decide to continue to try to pass to B even if B does not understand the signal. On the other hand, A may present additional signals to B to try and get them to understand their intention, perhaps through verbal means.

Another secondary joint project will then be created to complete this level of the joint action ladder. In other cases, A might discontinue the joint project altogether, as might be the case if B refuses to accept the pass. This may in fact be beneficial if, for example, it is likely that the pass will be intercepted by the opposing team. Examples of incomplete joint action scenarios are presented in Table 2.

A basic state diagram of the conceptual design of joint action ladders is given in Figure 1.

3.5 Common Ground in Basketball

Recall the brief discussion in Section 2 in which common ground was described. How can common ground be represented in agents? In particular, the common ground must be able to be modified an added to over the course of interactions, as well as being used as a decision-making tool. In this case, the gesture set and database for the second and third levels of the joint action ladder are an example of common ground representation. Agents may share the same knowledge in regards to how they interpret movements including passing, shooting, acknowledgment, refusal, and ordering a team-mate into a position. Furthermore, environmental common ground exists, such as how the ball bounces and the height of the hoop. Agents also hold shared awareness of how to play basketball and how to win the game.

What is more interesting is the common ground gained while communication occurs. Perhaps B refuses to accept A’s pass. The reason for this could be expanded on during further interaction, again adding to the common ground established between A and B. For example, B does not want to accept a pass when an opposing player is too close. The eventual outcome is that A has gained some new piece of information from B which can be added to common ground, namely ‘B does not want to be passed if an opposing player is too close’. As for the representation of this, it can easily be held in some form of database, and most importantly be accessible for future interactions with the same participant or members of the community that the participant belongs to. Other common ground to be learned is such information as how proficient a player is at shooting or passing. Part of the analysis also involves identifying the displays of common ground by the user during the game. Section 5 examines some of these issues further.

3.6 Extending the Basic Model

There are still limitations to converting Clark’s model into a computationally feasible structure, particularly the fact that the theory itself is more conducive to verbal rather than non-verbal communication. Firstly, the model requires that explicit adjacency pairs be used as evidence. In the case of body expression, there might be no recognizable public signal. For passing, players may not respond to indicate joint project uptake, but instinctively acknowledge their team-mate’s pass. A realistic model will have to account for this instinct, perhaps by including it as a parameter in the system, where the player does not wait for responses before executing joint projects.

Secondly, joint action ladders require four levels, although in the virtual world probably only three are feasible for any human-agent pair. At the second level, the agent can be programmed to always produce a correct signal, so there is no joint action required at this level. Similarly, there is little distinction between a user executing an incorrect signal and an agent not understanding it. As a result, these two levels should be combined. The model described previously uses this methodology.

The representation of common ground is also an issue, especially with how this information should be structured. For humans, common ground is simply remembered. In agents, common ground must not only be identified, but represented in a way which makes it relevant for future interactions. Common ground can be temporal (e.g. mutual attention) or persistent (e.g. knowledge of gesture meanings). A mechanism should exist so that agents differentiate between the two. Clark’s model can then be made more powerful for use in the virtual world.
4. AN IMMERSIVE ENVIRONMENT FOR LEARNING JOINT ACTIONS

The next issue is the design of the environment itself. Clearly it should contain sensors which can gather information about body expression, particularly the arm positions and body rotations. Furthermore, it should allow the user to interact with the virtual environment in a natural manner. For a basketball game, spatiality is also important. For this reason, a system has been constructed to achieve this task.

This system is named VISIE (Visual Interactive Spatially Immersive Environment). Full details of the system can be found in [7]. It is essentially a spatially immersive environment, with large displays completely surrounding the user. On these displays a virtual environment can be projected which immerses the user, preserving spatiality. The user is represented by an avatar, which can be viewed from either first or third-person perspective. For sensing the user’s input, several of Microsoft’s Kinect depth cameras are placed around the environment, allowing the user to rotate themselves and their avatar. These cameras measure the time taken to bounce light off an object and create a depth map. Trained decision forests then use this depth map to estimate the positions of several body parts such as shoulders, arms and legs, and thus create a fairly accurate skeletal representation of the user in real-time. Details of the algorithm can be found in [11].

The choice of the Kinect-based system was made to capture body modalities without the need for any peripherals or wear any markers, as in motion capture. Another possible solution is to use calibrated cameras to track movements. However, this was impractical with the use of the large displays. The Kinect system transmits data detailing the joint positions of the user, as well as their body orientation (the angle between their shoulder joints) and head orientation (the vector between the head and neck). This joint data is then processed by the main program and the rotations between the joints are calculated. Using this information, it is possible to manipulate an avatar which has its joints mapped to those recognized by Kinect.

This input mechanism is crucial for being able to successfully participate in the basketball game. A user may manipulate their avatar in real-time, and also allow for a gesture recognition system to be implemented. The basketball scenario is ideal in that it does not tend to rely on gestures which use smaller parts of the body, such as fingers. Kinect is so far unable to recognize these features robustly, although third-party software such as the FUBI framework [5] may offer a solution in the future. One issue is that the robustness of the multiple Kinect system is far from perfect, however development is ongoing to ensure that inconsistencies do not affect the activity of the user. In general, users are able to rotate and execute movements appropriately while using VISIE. Kinect also offers facial feature tracking, which may be used to provide avatars the ability to execute facial expressions. A small amount of supplementary verbal modality is also being considered, with a voice recognition system to be integrated in future iterations of VISIE.

The user must also navigate through the environment appropriately, much like a real basketball game. While achieving this is possible by using a pressure pad to recognize walking direction [7], this appears to be most useful for wide-area navigation. It may be more suitable for the user to simply indicate walking direction through transferring body weight as opposed to virtual walking. In this case, practicality would take precedence over naturalness.

A crucial aspect of VISIE is the spatiality it creates for the user. Ideally, the user should make decisions in the game much as they would in the real world, including those based on the locations of objects in the environment. If a traditional display were used, the need to rotate the world in order to view another direction would provide some inconsistency with regards to immersion. Furthermore, the mechanism for this rotation would add extra issues to the recognition of gestures. A full spatial environment addresses these concerns, allowing the user to simply to focus on executing gestures towards their target of interest.

5. DISCUSSION

The previous sections gave a theoretical framework of joint activities and how they can be implemented in a computational environment, and also a specific implementation with respect to a virtual basketball game. The next, and most important, issue to be discussed is what types of insights can be gained. This section will show that there are a large number of threads for research.

5.1 Analysis of Learning

Previously, the notion of using common ground as a means to analyze learning was touched upon. The argument was made that a correct use of a gesture as common ground gives public evidence that it has been learned. To clarify this, suppose that a virtual basketball team is aware of all the gestures needed to participate successfully. When a new member enters the team (in this case, the user), they are not aware of the types of gestures that are most appropriate. How does the user manage to learn what is the norm for this team? One way is through observation. For example, by observing how agents in the team catch and pass, appropriate gestures can be determined. These kinds of scenarios provide a simple form of analysis.

Another method of analysis is through interaction with the game itself. What further knowledge can be learned here, and how can this be proven? Since the user can now affect the environment, it presents an opportunity to modify the agent’s internal states and create some unique behaviors. Consider the situation where an agent repeatedly passes to a user yet they always miss the ball. The agent’s
new goal might be to refrain from passing the ball to the user unless absolutely necessary, in line with its belief that the user is not very good. The common ground representation, while not wholly explicit, is apparent to everyone playing the game. Whether or not the user can perceive and learn this information is a compelling phenomena.

The evidence that the user has learned this knowledge can come in several forms. Firstly, simply asking the user how they felt about the game may give some sort of proof. Another possibility is to analyze the behavior of the user during the game. There may be some evidence through body language that they have reacted to this knowledge. Perhaps they may try harder to redeem themselves to the agent, becoming more active during the game. Conversely, they may essentially give up and let the other agents participate more freely. Either way, this type of evidence will show that the user has learned something which was ‘taught’ to them by the agents’ body expressions.

5.3 Body Expression in the Virtual World

Above all else, this research attempts to create an agent which can interact using body expression. It makes sense to analyze body expression from the perspective of both agent and human. For agents, the interest lies in how accurately they can interact using body expression. It makes sense to analyze body expression from the perspective of both agent and human. For agents, the interest lies in how accurately they can recognize and react to behavior from the user. Accuracy can be measured through both the ratio of correct or incorrect communicative acts expressed by the agent as well as the number of breakdowns in communication which occur. The expressions which the agents use to repair communication breakdowns is also crucial, as this is a key part of the joint activity model.

On the other hand, human behavior in the virtual world is less predictable, but far more interesting as different situations unfold. One aspect of learning is how the user adapts to the behavior of the agent. Much like entering a new sports team, the user begins with little knowledge on the organization of the team and how communication occurs. As interactions occur, the user then begins to adjust to the team until eventually they are fully integrated and can communicate easily. This process in a virtual environment is the target phenomena. When the agent executes an unknown gesture, does the user guess their intention or do they communicate their confusion? Even stimulating the user to treat the agent as a social being is not a trivial task, so if the use of the joint activity model leads to more uptake of communication, it will show the benefit of this framework.

5.4 Other Learning Scenarios

So far, the discussion has dealt with the example of a basketball game. While this scenario is useful for displaying the value of the joint project framework, extension to other learning scenarios is far more interesting. Such examples include communication with people with autism or showing people from a different culture how they can communicate.

It is well-known that those with autism have difficulty in communication with others. Although this paper has little expertise on autism, it is assumed that communication breakdowns occur in one or more levels of the joint action ladder. The separation of joint actions levels is a powerful mechanism to analyze this. For example, perhaps autistic people can perceive the behavior of others but have trouble interpreting the behavior as an identifiable signal. By using virtual characters, it may be possible to pinpoint where the breakdowns in communication are most likely to occur, through agent parameterization. Experts are then able to use this information to better teach autistic people how to communicate smoothly, or even use virtual agents as teachers.

Another situation where correct body expression is called for is that of engaging in interaction with someone from a different culture. It is common for people arrive to a different culture and feel uncomfortable with how they should behave, particularly with gestural behaviors which are unknown to them. While not proven empirically, intuitively it would seem levels in the joint-project framework are culture-free. In this case, a system could be designed in which agents representing different cultures interact with the user with culture-specific gestures. It is expected that much communication occurs in the recognition level of the joint project framework. Agents utilizing this framework are thus able to identify the source of the problem and then repair communication accordingly while the user is allowed to become familiar with a cultural-specific style of interaction.

Of course, there are many more target applications for the joint project framework. The above scenarios and the basketball game demonstrate the wide range of fields in which these type of agents can be applied.

6. CONCLUSION AND FUTURE WORK

This paper described Herbert Clark’s theory of joint activities and how it can be implemented in a system to create agents which can exhibit natural communication. For this implementation, the work in progress was described, a basketball game which the user can play using their body movements. An example communication of passing a ball was described, which was created through implementing the theories. Through this environment, many types of analyses are able to be performed.

One issue is that Clark’s theories of joint activities are based on observations rather than empiricism. It is nearly impossible to determine whether users actually go through the stages of the action ladder when communicating. However, the use of agents explicitly using this process is still important. If this type of agent produces greater naturalness
of communication with the user, it provides further support for Clark’s model. Another potential obstacle is the environment itself. While efforts have been made to produce a completely immersive environment, navigation and body recognition issues can still occur. The goal is to reduce these so that they do not interfere with the user in a way which causes them to spend cognitive energy grappling with the system.

For a realistic agent, verbal information must be considered. Physical activities such as sports contain a large number of body expressions, but should be supplemented with verbal utterances. Finding a suitable synchronization technique for use with Clark’s model is one of the future challenges, which means synchronizing modalities in agents who are not simply face-to-face communicators.

Future work in this area includes further prototypes of the basketball game to make it suitable for experimentation. Rather than attempting a large-scale prototype, an incremental approach will be used which can still provide opportunities for experimentation. For example, even the process of human-agent communication when passing a virtual ball will yield novel results. This research has the potential to expand into many areas which ideally will lead to more sophisticated agents in the future.

7. REFERENCES


