KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

BODILY NON-VERBAL INTERACTION WITH VIRTUAL CHARACTERS

Marco GILLIES[*a]

[a] *Department of Computing, Goldsmiths, University of London, UK*

ABSTRACT

Alongside spoken communication human conversation has a non-verbal component that conveys complex and subtle emotional and interpersonal information. This information is conveyed largely bodily with postures, gestures and facial expression. In order to capture the Kansei aspects of human interaction within a virtual environment, it is therefore vital to model this bodily interaction. This type of interaction is largely subconscious and therefore difficult to model explicitly. We therefore propose a data-driven learning approach to creating characters capable of non-verbal bodily interaction with humans.

Keywords: **Animation, Body Tracking, Non-verbal communication**

1. INTRODUCTION

Humans use their bodies in a highly expressive way during conversation, and animated characters that lack this form of non-verbal expression can seem stiff and unemotional. An important aspect of non-verbal expression is that people respond to each other's behavior and are highly attuned to picking up this type of response. This interaction, of course, includes verbal conversation, but it also includes non-verbal interaction, the use of the body to convey a range of social and emotional cues that are both subtle and complex. These cues, including gestures, posture and movement style, are vital to face-to-face interaction.

We propose that bodily non-verbal cues are a natural way of interaction with animated virtual characters[1]. Characters should be able to detect non-verbal cues in the behavior of a human and respond with appropriate cues of their own. These cues include gestures, posture and also other cues such as non-verbal aspects of speech (prosody). The cues used should be as close as possible to natural human cues that we use in our normal conversational interactions. This means that interfaces do not need to be learned, instead it is instinctive and often sub-conscious. If the character responds with sufficiently natural non-verbal cues then the human will respond

* **Corresponding author**: Marco Gillies, m.gillies@gold.ac.uk.

to them naturally and subconsciously as if they were a real person[1,2]. This creates a loop of non-verbal interaction that mimics real human interaction.

However, automatically generating this type of behavior is difficult as it is highly complex and subtle. This is an example of the general problem that the interactive behavior of a character is normally generated procedurally based on programmed rules and algorithms. It is difficult to capture subtle nuances of behavior in this way. Data driven techniques that are used for animation capture very well the nuances of an actor's performance. This paper applies data driven methods to creating characters capable of bodily non-verbal interaction. This involves both generating animated non-verbal behavior in the character and also responding to the speech and gestures of a human. We propose a two-layer model that separates learning the response model from generating realistic animation, and so can ensure that optimal techniques are used for both (figure 2). A Dynamic Bayesian Network is used to learn how the character responds to speech and gesture interaction. This model is then used to drive a motion graph that generates the animation. The character's movements and posture respond to emotional cues in the human's speech and movement.

2. BODILY INTERACTION

Much of human interaction is through speech. However, we should not forget that, in face to face conversation, this verbal communication is accompanied by other, non-verbal channels of communication. These are primarily bodily, postures, gestures and facial expressions, as well as non-verbal aspects of speech such as tone of voice. This non-verbal channel carries complex and subtle information and is produced and interpreted largely subconsciously, without most people even having a clear understanding of what non-verbal communication means. This information includes precisely those factors that are of interest to Kansei Engineering, emotional factors as well as relational and interpersonal factors. Importantly, non-verbal communication is also important for our evaluation of other people. It is therefore important to take non-verbal interaction into account when using a Kansei approach to applications with virtual characters. In fact it is vital, as most people will read non-verbal cues subconsciously and automatically. Even if no non-verbal cues are present, this will itself be interpreted as a cue (perhaps to a cold and stiff character), rather than as a lacking technical feature.

We therefore propose that bodily, non-verbal interaction is a vital aspect for any interactive virtual character. This should work both ways, characters should be capable of bodily expression, and people should be able to interact with characters bodily. The first involves animation and control algorithms that are expressive and that respond to people's behavior. This can be complex due to the subconscious nature of non-verbal behavior. We are largely unaware of the meanings we encode and interpret non-verbally and even at a scientific level they are not well understood. This means that we lack the information to design rules for controlling bodily interaction. For this reason, in this paper we propose using machine learning to discover the patterns implicit in data from human behavior, and use these patterns as a way of generating behavior. The subconscious nature of bodily interaction is also a reason for ensuring that people are able to interact bodily with characters. As people are unaware of how the produce non-verbal behavior it is very difficult, if not impossible, to explicitly control such behavior as would be needed if non-verbal cues had to be inputted using a traditional graphical user

interface. Body tracking interfaces make it possible to interact with characters using bodily movements. This allows people to interact naturally and expressively with a virtual character.

The remainder of this paper presents an initial prototype system for bodily interaction with a virtual character. It uses body tracking and voice analysis as input methods and the behavior of the character is learned from motion capture data.

3. LEARNING CONVERSATIONAL BEHAVIOR

We propose a method of learning a behavior controller from motion capture data. This method can create characters that interact in real-time with people, responding to a number of different user inputs. In this paper the characters respond to the persons voice, position and movement. The characters behavior is also affected by internal variables that can represent ”mental states” such as shyness or confusion. We achieve this by capturing data from a conversation between two people. One of the people is an actor playing the part of our character, and whose behavior is fully motion captured. This actor's conversational partner plays the part of the user interacting with the character. We record the voice of the conversational partner and position tracking data. The actor's behavior is captured in a number of different mental states or personalities such as capturing the actor being polite or rude. We then learn a model that relates the two. Another key element of the method is the separation between the animation and behavior layers. We use state of the art animation techniques to generate the character's movement and only use the learned model to control the animation layer. This makes the learning problem more tractable and generates higher quality animation, because tried and tested techniques may be employed. The method results in real-time prediction and realization of the behavior of a virtual character as a function of the behavior of a real tracked person. This is important because, for the first time, it provides the possibility of highly realistic, data driven interaction between real and virtual people.

This method was tested with a specific example dealing with response to emotionally charged interaction. It was an interaction between a customer and store clerk, the system aimed to detect aggressive behavior in the customer and have the virtual store clerk respond appropriately.

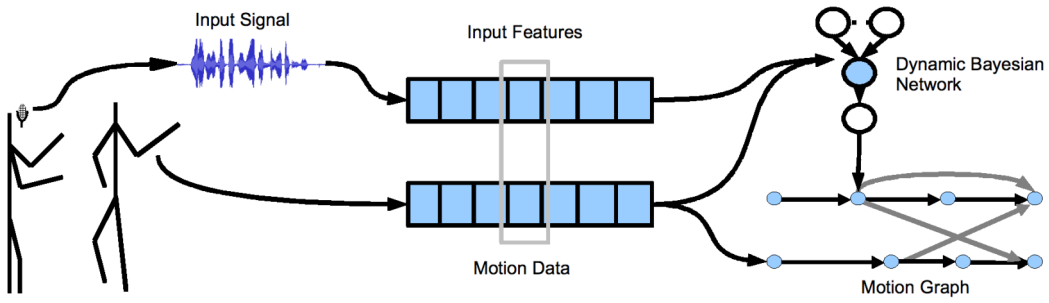

Figure 1 The process of capturing data and creating a virtual character

3.1. The Capture Process

The process of creating a character is illustrated in figure 1. We capture a conversation between two people, one playing the part of our character (referred to as “the actor”) and the

other the part of their conversational partner (referred to as “the conversational partner”). The actor is fully motion captured using an optical motion capture system. The aim is to capture their distinctive style of behavior. His or her behavior is captured in a number of different mental states. The conversational partner has their voice recorded and in some cases their position is tracked. From these two sets of data we create a model of both the style of movement of the motion captured person and their style of behavior, and how they respond to other people. This is possible through the use of state of the art machine learning techniques.

The capture scenario for our example involved an acted interaction between a male customer and a male store clerk. The shop assistant was motion captured while the customer had his voice recorded and head and hand position tracked. The customer was complaining and behaved in an aggressive way, shouting and moving in a threatening way. The actor playing the clerk was recorded responding in two different ways. The first was shy and submissive, the clerk was intimidated by the customer's behavior and responded in a fearful and submissive way. The second response was rude, the clerk paid little attention to the customer and when he did respond he responded in an aggressive way. The customer's voice was recorded and his head and hand were also tracked.

3.2. A Two Layer Learning Model

We use a two layer learning model shown in Figure 2. The lower layer is an animation model based on Motion Graphs [3,4,5], which is used to generate realistic motions. This model determines which animations clips can be played at a given time so as to ensure smooth motion. The higher level model is a parametric statistical model which selects one of these clips based on the input features.

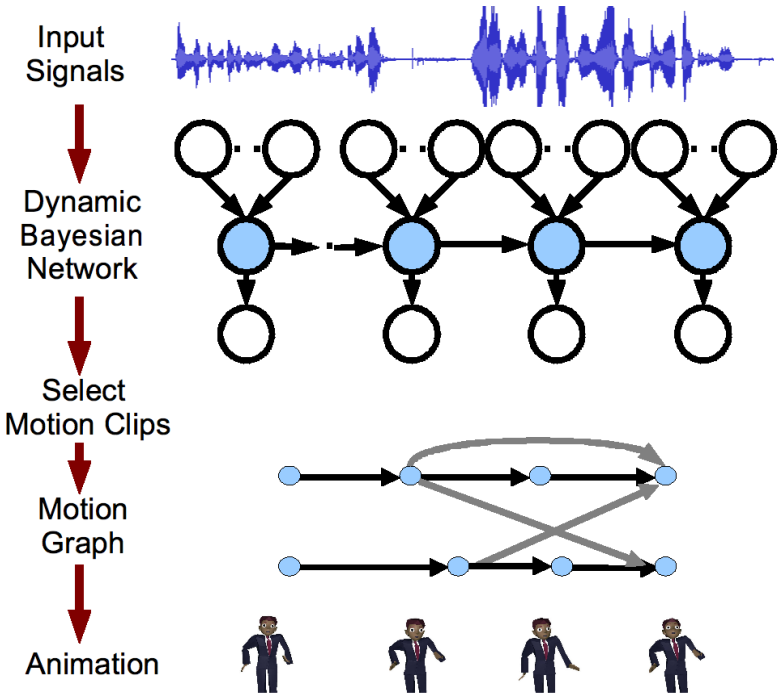

Figure 2 A two layer learning model

3.3. The High Level Model

The purpose of the high level model is to relate input parameters such as a real person's voice or the character's internal state with the output animation, selecting animation clips that are appropriate to the current set of input parameters. In order to do this we use a probabilistic model with which we can select motion clips based on their probability given the inputs, P(m|i). These probabilities are learned from data using Dynamic Bayesian Networks, which are a generalization of Bayesian Networks. A Bayesian Network is a directed acyclic graph structure consisting of nodes, which represent random variables and directed edges that represent dependence between variables (see Figure 3). More precisely node A is the parent of node B if there is an edge from A to B. Any node is statistically independent of all nodes other than its parents and descendents, given (conditioned on) its parents.

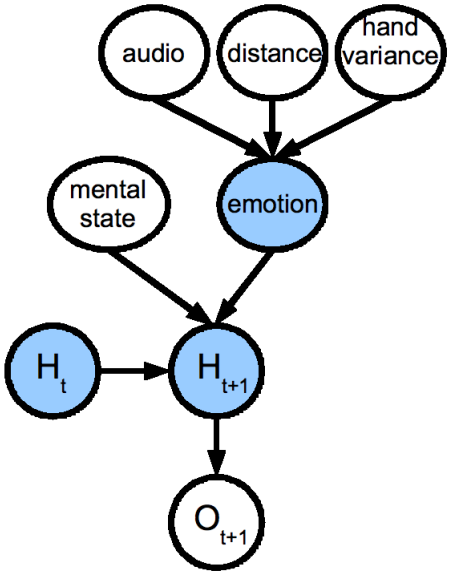

Figure 3 The Dynamic Bayesian Network used in our prototype. Observed nodes are shown in white and hidden nodes in blue. The value of hidden node H at time t+1 (H_{t+1}) depends on its own value in the previous time step (H_{t+1}).

Each edge is labeled with the conditional probability distribution of the child given the parent. Making the independence of variables explicit in the structure of the graph makes it possible to factor the full probability distribution of all variables into a number of smaller distributions relating the variables that have dependencies and thus enabling more efficient calculations. Bayesian Networks can be used to calculate probabilities within the network given some observed values of the variables. Some variables will be observed while others will be unobserved or hidden (H in Figure 3). The network can be used either to calculate the probabilities of the hidden variables, in order to estimate them, or to calculate the probability of a given observation. Dynamic Bayesian Networks (DBNs) [6] are a generalization of Bayesian Networks to sequences of data. Each step in the sequence is a set of values for the random variables. As, in a Bayesian Network, dependencies exist between the values of variables but variables can also depend on their own value, or values of other variables, in the previous step of the sequence (see Figure 3). Thus DBNs can model the evolution of variables over time. Early work by Ball and Breese[7] used Bayesian Networks for affect detection during

interactions with a virtual character. Pelachaud and Poggi [8] have used Dynamic Bayesian Networks for animated characters but they do not use machine learning, rather they use a priori probabilities as the parameters of their network. Brand and Herzman's Style Machines[9] can be regarded as a type of DBN and so this work can partly be thought of as a generalization of theirs. DBNs are closely related to Hidden Markov Models that have been used extensively for speech analysis and recently applied to non-verbal behaviour[10].

The fact that Dynamic Bayesian Networks can represent temporal sequences makes them very well suited to applications with motion data. For the current application the sequences consist of a number of frames of motion data, with each frame marked up with input features. The DBN topology used is shown in Figure 3. It contains a number of nodes for the input features. These features are combined together into a hidden node that represents their total effect (labeled “emotion” in figure 3). This node, together with a node representing the “mental state” of the character are parents of a second hidden node, which is the only node to depend on the previous time step. This hidden node provides a link between input and the animation. Because the hidden node depends on the previous time step it is able to represent the time varying aspects of the animation. It represents the current state of the animation, depending not only on the current position and posture of the character but also on how it depends on previous behavior. The hidden nodes value can be one of a number of different states of a motion, an example might be the different phases of a gesture. As the node is hidden, the exact meaning of these different states is learned directly from the data so as to optimize their ability to relate the inputs with the motions. Finally there is an output node, O which represents the motion data.

In our store clerk example, the input data recorded was the customers voice and head and hand tracking data. Each of these inputs can provide some indication as to whether the customer is angry Shouting can be picked up from the audio volume. The volume was discretized into three levels, level 0 was set to the sound level when the customer was not talking, 1 was the level when he was talking normally and 2 was the level when he was talking loudly or shouting. The position tracker can detect whether the customer has moved close to the clerk, a sign of aggression, by taking the distance between the two (discretized to two levels, far and close). Aggressive behavior is also associated with fast arm gestures, which can be detected with the hand tracker by taking the variance of the signal, discretized to two levels. However, none of these cues is a good predictor on its own, so our model combines them. A new hidden node, with two possible states, was introduced to represent the combined effect of these inputs. This node is called emotion in the diagram as it is intended to give an indication of the emotional state of the customer, however, its exact semantics are learned from the data. In addition a further input node represented the mental state of the clerk: shy or rude.

We represent motion data in the conventional way as a number of frames where each frame contains a vector translation and a rotation for the root, and a rotation each joint. Since there are 28 joints plus the root in out data set and each rotation consists of 3 parameters the data is 84 dimensional, however, there is a high degree of redundancy since there are strong correlations between the movements of different joints, and so the dimensionality can be greatly reduced. The first step is to a Principal Component Analysis which greatly reduces the dimensionality of the data (to between 5 and 10 depending on the data set). We then use vector quantization to reduce the data to a discrete variable. Vector quantization is an unsuperivised clustering method that finds an optimal discrete representation of a multi-dimensional data set.

3.4. The Low Level Model

As a low level animation model we use a Motion Graph[3,4,5]. This is a directed graph structure in which edges are motion clips and nodes are points are which transitions can be made smoothly between clips. Animation is generated by walking the graph, selecting an outgoing edge at each node. This edge is played and then a new edge is selected at its end node. We use the Dynamic Bayesian Network to select edges. All outgoing edges of a node are analyzed using the DBN. The probability of each edge, given the current input values, is evaluated. The edge with the highest probability is then selected.

Figure 4 Frames of the generated animation

3.5. Results

In order to produce the examples shown in this paper we made a desktop test system in which all three user inputs were triggered by the voice, if the user shouted the distance and hand movement nodes were activated. The frames from resulting animation are shown in Figure 4. The Figure 5 shows an example of a real-time interaction involving voice, head and hand tracking in an immersive projection environment.

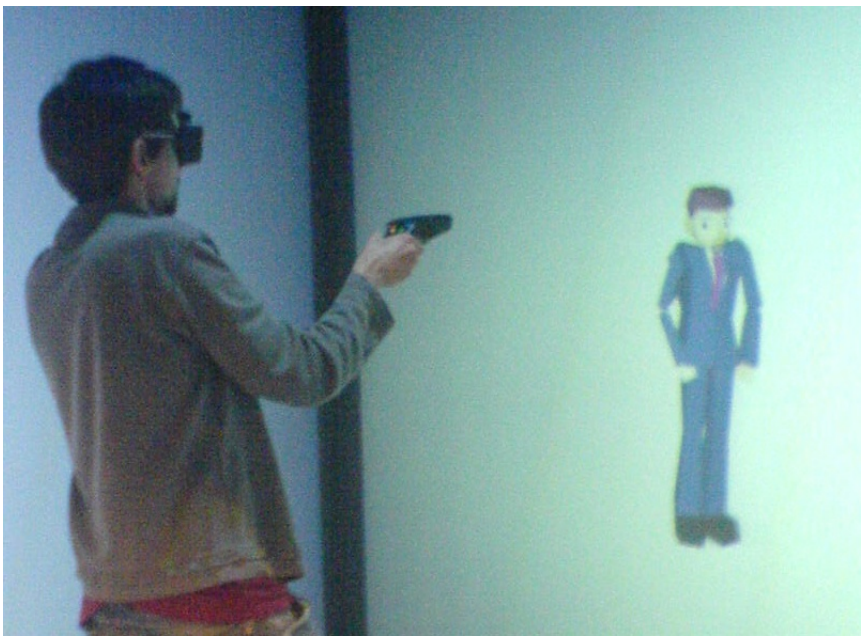

Figure 5 Bodily interaction with a virtual character in an immersive environment

4. CONCLUSION AND FURTHER WORK

In this paper we have proposed an approach to creating full body interaction with virtual character and have demonstrated a software framework that implements this. The next stage is to use this framework in practical applications as a real test of the validity of our method. This will bring a number of challenges. The first of these is the appropriate choice and modeling of input features. In the current example we have used *ad hoc* features based on voice and tracking. Further research will investigate in more detail what are appropriate feature to use and whether more complex methods are needed to models (for example Hidden Markov Models to extract actions). The use of more complex features also implies modifications to the high level model and in particular which DBN topologies should be used. How should inputs be combined, what hidden nodes are needed and which independence assumptions are valid.

REFERENCES

1. Vinayagamoorthy, V., Gillies M., Steed, A., Tanguy, E., Pan, X., Loscos, C., and Slater, M., Building Expression into Virtual Characters *In the proceeding of the Eurographics Conference State of the Art Reports* 2006
2. Pertaub D. P., Barker, C. and Slater, M., An experiment on public speaking anxiety in response to three different types of virtual audience, *An experiment on public speaking anxiety in response to three different types of virtual audience*, Vol. 11, No. 1, pp. 68-78, 2002.
3. Arikan O., and Forsyth, D. A., Interactive Motion Generation from Examples, *ACM Transactions on Graphics*, Vol. 21, No. 3, pp. 483-490, 2002.
4. Kovar, L., Gleicher, M., and Pighin, F., Motion Graphs, *ACM Transactions on Graphics*, Vol. 21, No. 3, pp. 473--482, 2002.
5. Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., and Pollard, N. S., Interactive Control of Avatars Animated With Human Motion Data, *ACM Transactions on Graphics*, Vol. 21, No. 3, pp. 491-500, 2002.
6. Murphy, K., *Dynamic bayesian networks: representation, inference and learning*, PhD Thesis, 2002.
7. Ball,G.,Breese,J.:EmotionandPersonalityinaConversationalAgent.In:Cassell,J.,Sullivan, J., Prevost, S., Churchill, E. (eds.): Embodied Conversational Agents. (2000)
8. Pelachaud, C., and Poggi, I., Interactive Subtleties of facial expressions in embodied agents, *Journal of Visualization and Computer Animation*, Vol. 13, pp. 287-300, 2002.
9. Brand, M., and Hertzmann, A., Style Machines *In the proceeding of ACM SIGGRAPH* pp 183-192 2000
10. Hofer, G., & Shimodaira, H. Automatic Head Motion Prediction from Speech Data. Interspeech, Antwerp, 2007.