Towards Basic Emotion Recognition using Players Body and Hands Pose in Virtual Reality Narrative Experiences

Gabriel Peñas and Federico Peinado

Department of Artificial Intelligence and Software Engineering
Universidad Complutense de Madrid
c/ Profesor José García Santesmases, 9. 28040, Madrid (Spain)
g.penas@ucm.es, email@federicopeinado.com
www.narratech.com

Abstract. Currently, players position recognition in most Virtual Reality applications is limited to the evident usage, like translating players avatar within the virtual environment, or using the view point at head height, without considering the posture at any time. In this article, we propose studying the player's body and hand expression, not only to recognize obvious interaction patterns but poses that, even without conscience, transmit information about the basic emotions of such player. That way, each time it is played, the result is altered without a conscious effort, the experience of interactive narrative resulting of computing generation depends on the input signals, which adds a layer of depth and enriches system decision making and conversation with non-player characters, for example. Our proposal is based on a system that relies on a system with a neural network which can recognize poses, according to the specialists, associated to basic human mood. After a simple calibration and a reasonable training, this system can be used, without the need of additional accessories, with the main Virtual Reality devices existing today. This article also discusses new paths of research and applications that arise around this system, many in the field of computer entertainment, but also in other areas such as therapy for patients with emotional and social communication problems and disorders.

Keywords: Human-Computer Interaction, Affective Computing, Feelings Analysis, Human Pose Recognition, Interactive Narrative, Dialog System.

1 Introduction

Noting the current applications of Virtual Reality [1] we can observe that there is still a great potential in the input signals information that has not been used, although it has been studied with other devices [2]. All virtual environments receive constant feedback about the positioning of players body, mainly his head and hands, thanks to head mounted displays (HMDs) and more sophisticated hand controllers (such as the HTC Vive and the Oculus Touch). This constant information flow is very valuable and allows us to translate subtle actions and movements from the player to the virtual world [3], providing a greater immersive interaction than conventional usage.

Even though may applications have used evident patterns to allow the player to communicate with the game (waving hand, nodding, thumbs up, swipe, etc.), we consider that it is possible to use all this data another way. Instead of recognizing symbols of a somewhat natural and intuitive language that the player uses to communicate, we will try to understand how such communication happens, even in unconscious terms. During gameplay sessions in a Virtual Reality experience, user always adopts a pose and this, in some way and according to nonverbal language experts [4], can reveal his mood against a situation that arises in a particular moment of the game. If we can detect them, new possibilities are opened to achieve a more meaningful and expressive interaction.

Our first approach is to create a simple system using a neural network [5] that can classify players poses. And once achieved, use that information to enrich interaction with the system, specifically by modifying the dialogs with non-player characters and redirecting all the narrative. Writing the conversations must consider those new variables, being as important the message as how is it said. Therefore, it is added a new layer of depth and realism to the experience, achieving greater credibility in the interaction by using the mood analysis.

2 Using Emotions in Video Games

Videogames have their own resources to provoke players emotions, a widely studied phenomenon, about which we can mention Proteus Paradox [6]. However, in almost all the current videogames emotions are not taken as an input parameter, his pose and gestures are completely ignored due to technological limitations that the medium drags since its origins. As it was not possible to know the mood of the player automatically, their emotions could be inferred, estimated, or even asked directly [7].

There are many games that have tried, with mixed success, to make use of players emotions to guide the narrative. For example, the *Mass Effect* series has elections in conversations that allows us to specifically distinguish the kind of answer that we want to give. This game can be included within the role-playing games (RPG) genre, characterized, inter alia, by giving the player the freedom to choose how to interpret the thoughts and emotions of his avatar, though often it is confused how they really feel and what they want to transmit.

Other games try to use more modern devices so the gameplay is affected by less voluntary actions from the player, for example IMMERSE Project [8] uses Microsoft Kinect to allow facial recognition, it uses a system to identify facial expressions associated with emotions and thus alter the games behaviour. It is also the case of Stifled [9] that uses the microphone to draw sound waves that allow us to see the virtual environment, these are generated by the voice or breathing that increases with your pace and intensity at the same time as players nerves, thus achieves a balance because if the player feels nervous he can see more around him which reduces his anxiety. The ability to apply user emotions to interaction with non-player characters and adapt their dialogues has also been studied before [10] but has not been developed or tested in practical examples.

Nowadays we have better ways to detect emotions, and the devices have greater computational power. On one hand, we have cameras with enough resolution that allow us to recognize facial expressions from the users. Some devices use those same cameras to track eye movement and position. There are also other devices like smart watches that can measure pulse and share the information in real-time. The information they provide is very useful, but their adoption is marginal within videogame core players.

However, thanks to the growth of Virtual Reality, the use of hand controllers is becoming a standard, and their position is tracked with precision as the HMD. For the HTC Vive, controllers are an essential part of the kit and the user must purchase them altogether. This integration is facilitating the adoption by the players gradually, which will allow the developers to have valuable information to work with, information that is simple to apply to our experience without being artificial or generate user rejection.

3 Reading Players Body Methodology

The goal of this research is top prove that we can create a system, based on a neural network, that classifies players body postures while he enjoys an interactive narration experience, all making use of existing and broadly use technologies. To achieve the objective, we have developed an experiment based in a brief interactive sequence where a conversation with a non-player character is taking place. This dialog is fixed and has some branches that can be explored, guided by the pose adopted by the player in each of the states of the conversation.

The system requires of hardware component to track poses and show an immersive vision in any direction, and a suitable software. We have used the HTC Vive [1] consisting of a HMD and two hand controllers, each one with six degrees of freedom (DOF) that allows us tracking position and rotation at the same time. Our solution is compatible with other devices supported by the OpenVR[2] library, such as Oculus Rift and OSVR, if they provide the three required elements: HMD and one tracking control per hand. Note that our experience works while standing or seated because it only tracks head and hands, not legs, chest, or waist.

We only calculate poses when the user interacts in the conversation so erratic movements are not a problem. The tracking system used in the HTC Vice gives us a worst-case latency of 22ms with a relative error of 1.7cm [11].

The implementation has been done using Unity[3] game engine. This choice was done because its ease of use in prototyping, allowing a fast integration of Virtual Reality and a comfortable and effective programming in C# language. As said, we also used the OpenVR library, not only because it supports multiple devices, but also because it simplifies the procedure and it is becoming an industry standard. Implementation is flexible and can be expanded both in the number of poses and the complexity of them.

---

1 HTC Vive, https://www.vive.com/eu/, last access 2017/05/25.

2 OpenVR, https://github.com/ValveSoftware/openvr, last access 2017/05/25.

3 Unity 3D, https://unity3d.com, last access 2017/05/25.

System workflow is as following: first step before using it preparing input data from the devices. To do it we must perform a calibration where the user must perform in two poses as shown in Figure 1: natural standing and T.

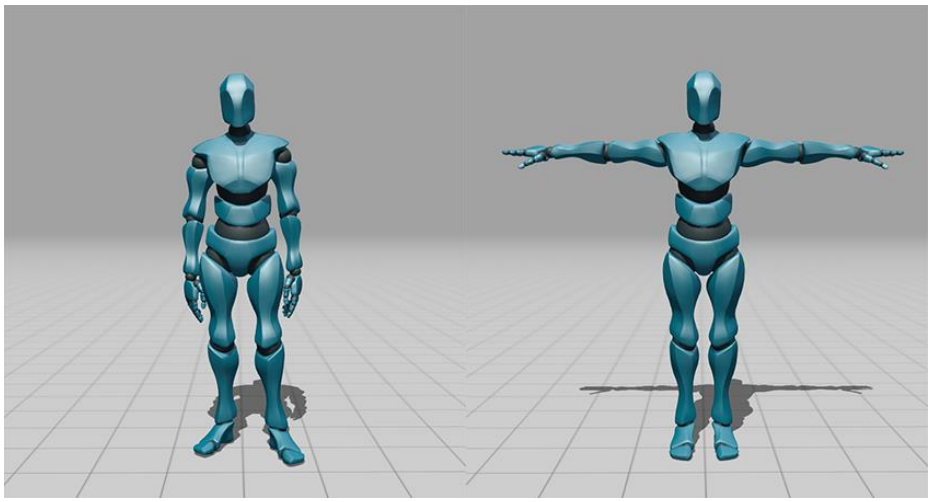

Fig. 1. Calibration poses: natural and T.

This way we can get the natural position of the head and the distance of both controllers to the HMD. We use this distance to normalize the data before it is sent to the system, thus allowing us to acquire some independence from the user body proportions.

For the pose classification, we have a neural network previously trained with some positive example cases. Obviously, the greater the number of there, greater will be the accuracy of the classifier. In this experiment, we have used a simple network with three layers:

- Input: 18 neurons (6 DOF * 3 devices).
- Intermediate: 30 neurons.
- Output: 3 neurons (one per pose).

We decided to use a neural network, instead of other methods, because we needed a fast system, ideal for real-time application as the experiment needs, although we sacrifice some recognition capacity as counter effect; the primary goal is testing the concept, not the implementation. If we vary the number of neurons in the middle layer or the number of the intermediate layers, we can achieve greater precision at the expense of higher resource use. The focus of this experiment was not obtaining the greatest results in detection and therefore only we only used one intermediate layer achieving good classification results and nice performance.

Figure 2 shows the poses we considered in the example scene, they are easily identifiable: neutral, aggressive, and defensive.

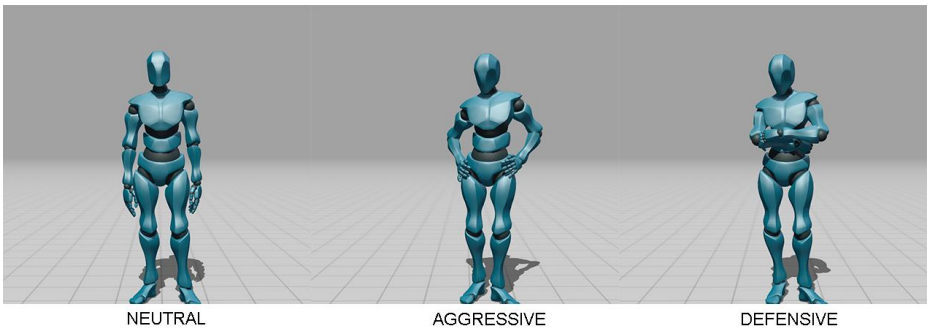

Fig. 2. Recognized poses.

These poses are generic and in our preliminary test do not seem to generate many problems while used by different users, all them coming from the same region and culture so they do them mostly the same way.

We wrote a small survey to get feedback from the test subjects, the questions are the following:

1. Genre. [1-Man/2-Woman]
2. Age. [0-N]
3. Education level. [1-Secondary/2-Bachelor/3-Degree/4-Master/5-PhD]
4. Number of experiences in VR simulations. [0-N]
5. Gaming level. [1-None/2-Casual/3-Hardcore/4-Competitive]
6. I enjoyed the experience. [1-6]
7. My attention was entirely on the experience. [1-6]
8. My perception was focused on the experience almost automatically. [1-6]
9. The environment was comfortable. [1-6]
10. I felt that the game was disorientating. [1-6]
11. I felt like I was a part of the game. [1-6]
12. The length was enough. [1-6]
13. The experience surprised me. [1-6]
14. I noticed that the NPC reacts to my poses. [1-6]
15. There is a high variety of poses. [1-6]

User takes control of an avatar that is talking with a non-player character having the capacity of expressing the six universal basic emotions (joy, sadness, rage, fear, surprise, and disgust) [12] plus a neutral one. A very recognizable facial expression manifests these emotions as shown in Figure 3, as well as a set of body animations that reinforce each emotion.

Fig. 3. Basic emotions.

During the conversation, which is shown in Figure 4 in a tree-like shape, it is possible to visit several branches, and, in this example, the player only has one sentence to give to the non-player character as answer to whatever is said in each node [13]. In fact, is the pose taken by the player what confirms the mood and guides the conversation towards a direction. This conversation is, therefore, very reduced, but allows us to make use of all the available poses and test if all the basic emotions implemented work how they should.

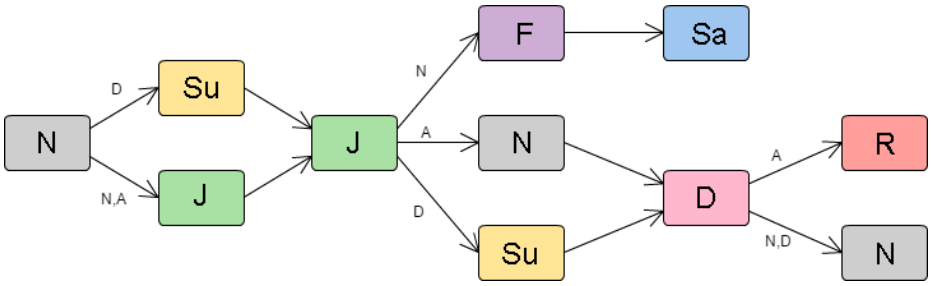

Fig. 4. Conversation tree diagram, each letter is a pose or emotion.

The system is highly parametrized in such way that we can make changes easily in the neural network, such as the number of neurons in each layer, and the number of layers. For this example, we trained the network each of the three poses using 25 positive examples per pose. Of these 25 examples per pose, we divided them in blocks of five done by 5 different people to give a greater variety to them. This way we have some variety in the training examples.

The main structure of the trainer and recognizer is simple to promote efficiency but works with great results.

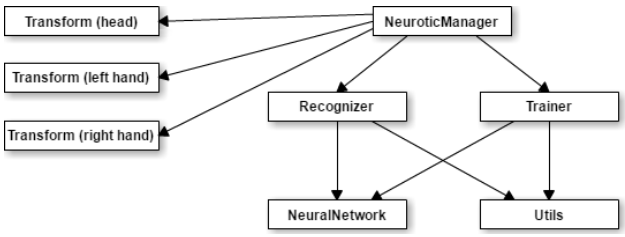

Fig. 5. System kernel.

Consists of a manager that deals with the control of the various states of the tool, their names are very explanatory as shown in Figure 5. The manager also deals with the tracking of the three elements (display and hands). The recognizer and the trainer manage control in a different way the neural network each one contains, while the first only applies input data to see the obtained result, the second generates iterations with all the examples of each pose to recalculate the weights between the neurons in the network. There is also a Utils class than mainly deals with the file read and writing.

Fig. 6. Scene where the conversation takes place.

4 Results and Discussion

We tested the experiment with 10 people, half men, half women, and all of them between 30 and 60 years and higher education. In terms of experience with Virtual Reality simulation environments it was none or negligible (a couple of sessions at most). Relation with videogames was more heterogeneous since there were some who have practically never played to regular players.

After trying the experiment with these users, we observed that there is an initial reaction of surprise once the subjects realize the ability the non-player character can vary is answers. Although they are not conscious during the test that it is due to their poses, as they were not informed of that, after explaining the workings they say it looked like the NPC reacts to their feelings as he was reading their minds.

On the other hand, it is noteworthy that the number of poses and nodes of the current conversation are still very few, even more considering that the subjects tend to go through the same branches of the conversation, visiting a reduced number of nodes. Generating a more complex conversation could improve replayability.

The negative part of the findings is that, being a reduced number of poses, the system is not able to recognize all the different types of mood the player can have, training more poses could provide very interesting information that now are completely ignored by the system. Also, as the test conversation is short, sessions pass too fast and the players do not develop an intellectual or sentimental attachment to the story.

All this information is obtained from the survey results shown in Table 1.

Table 1. Survey results, rows are the questions, columns are the answers.

Q/S	1	2	3	4	5	6	7	8	9	10
1	1	1	2	1	2	2	2	1	2	1
2	33	32	45	59	32	58	36	37	52	32
3	4	4	3	3	3	3	2	2	4	3
4	5	3	0	1	1	1	0	0	1	0
5	3	3	2	1	2	1	2	1	1	3
6	5	6	5	5	6	4	5	5	6	6
7	6	6	5	6	6	5	6	6	6	5
8	4	5	4	5	5	4	5	4	6	6
9	6	5	6	6	6	6	5	6	5	6
10	1	1	2	1	1	3	1	2	1	2
11	5	6	5	6	6	4	5	5	6	6
12	3	3	4	4	4	3	2	3	4	2
13	5	4	6	5	6	4	5	5	6	6
14	6	5	5	5	6	4	5	5	6	6
15	3	2	2	3	4	2	3	3	4	2

With this experiment, we achieved a new way to add a greater depth and immersion to the narrative experiences in Virtual Reality. The most remarkable achievement is the possibility of integration into more complex conversation or, directly, detect the user mood at any time to guide the story to new horizons.

This has direct implications in videogame design and allows developers to create new experiences reactive to the player and with a low latency response to the users' feelings. The fact that the system is simple does not make the experience design as simple. The number of different conversations that can be developed with this new input grows exponentially and requires high skill in the creation of interactive dialogues with quality and sense.

5 Conclusion

With a simple use of available tools, we can create conversations with greater meaning, more immersive and offering more emotional and complex content. We revealed also with this experiment a lack of interaction in videogames, particularly in Virtual Reality environments.

Interaction with systems is based merely on voluntary actions of the user, without taking care of subconscious elements as the emotions. If we use them, systems could adapt to the necessities of whoever is using them. Not only that, but we can also achieve a more natural connection in the communication between the users and the systems, a communication that can be more comfortable and productive.

We also make evident the extreme linearity of the videogame narrative. Breaking that barrier is complicated because the huge amount of content we must generate is complex and costly. It is not suitable for all games or players, but it could improve those with a strong narrative component.

We see many lines of further research:

- Adding devices with more complex expressive information, like data gloves, hand recognizers as Leap Motion[4] or full body motion tracking.
- Adding more varied devices that can give other parameters as the pulse, breathing, eye tracking and pupil dilation. This could provide more precision classifying the mood and detecting other poses that cannot be done only with the pose.
- Test other A.I. techniques that improve the system, either by having greater efficiency in real-time recognition, or able to recognize more complex elements.
- Implementing a gesture recognition that reinforces the data obtained from the poses. This can provide more subconscious information that can reveal contradictions between the pose the player is performing ant what he really feels.
- New methods of emotion guided narrative to develop deeper worlds that adapt every game session, even chatbots with emotion information in the conversations.

References

1. Anthes, C., Garcia-Hernandez, R. J., Wiedemann, M. & Kranzlmuller, D.: State of the art of virtual reality technology. IEEE Aerospace Conference, pp 1–19 (2016).
2. Manresa, C., Varona, J., Mas, R., & Perales, F. J.: Hand tracking and gesture recognition for human-computer interaction. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 5(3), pp 96-104 (2005).
3. Steuer, J.: Defining Virtual Reality: Dimensions Determining Telepresence. Journal of Communication 42(4), 73-93(1992).
4. Hermelin, B., O'connor, N.: Logico-affective States and Nonverbal Language, Communication Problems in Autism, pp 283-310, Springer US (1985).
5. Haykin, S., & Network, N. (2004): Neural Networks. A comprehensive foundation. Prentice Hall, United States (2004).
6. Yee, N.: The Proteus Paradox: How Online Games and Virtual Worlds Change Us—And How They Don't. 1st edn. Yale University Press (2014).
7. Bee, N., Prendinger, H., Nakasone, A., André, E., and Ishizuka, M.: AutoSelect: What You Want Is What You Get. Real-time processing of visual attention and affect. Proceedings International Tutorial and Research Workshop on Perception and Interactive Technologies. Springer, Kloster Irsee, Germany (2006).
8. Playabl.IA IMMERSE, http://www.playabl.ai/projects/, last access 2017/05/25.
9. Stifled, http://www.stifledgame.com/, last access 2017/05/25.
10. Gómez-Gauchía, H, Peinado, F.: Automatic Customization of Non-Player Characters Using Players Temperament. In: 3rd International Conference on Technologies for Interactive Digital Storytelling and Entertainment. Springer, Darmstadt, Germany (2006).
11. Niehorster, D. C., Li, L., & Lappe, M. (2017). The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research. I-Perception, 8(3) (2017).
12. Ekman, P: Basic Emotions. Handbook of cognition and emotion, pp. 45–60 (1999).
13. Dastani, M., Meyer, J.J.C.: Programming agents with emotions. In ECAI, pp. 215-219 (2006).

[4] Leap Motion, https://www.leapmotion.com/, last access 2017/05/25