Interacting with a Virtual Rap Dancer

Dennis Reidsma, Anton Nijholt, Rutger Rienks, and Hendri Hondorp

Human Media Interaction Group,
Center of Telematics and Information Technology, PO Box 217,
7500 AE Enschede, The Netherlands
{dennisr, anijholt, rienks, hendri}@ewi.utwente.nl
http://hmi.ewi.utwente.nl/

Abstract. This paper presents a virtual dancer that is able to dance to the beat of music coming in through the microphone and to motion beats detected in the video stream of a human dancer. In the current version its moves are generated from a lexicon that was derived manually from the analysis of the video clips of nine rap songs of different rappers. The system also allows for adaptation of the moves in the lexicon on the basis of style parameters.

1 Introduction

"Rapping is one of the elements of hip hop as well as the distinguishing feature of hip hop music; it is a form of rhyming lyrics spoken rhythmically over musical instruments, with a musical backdrop of sampling, scratching and mixing by DJs." (Wikipedia). Hip hop includes break dancing. Well-known are break dance moves as the six step or the head- and hand-spin. Rappers also move; they move to the rhythm and the lyrics, performing hand, arm and bodily gestures. But most of all there are lyrics with content and form that characterize rap.

Rap and hip hop have been studied by scholars. The usual viewpoints are ethnographic, cultural, social and geographic [1]. Time, place and cultural identity are considered to be important aspects that need to be studied and obviously, the lyrics invite to look at race issues and at the rage, the sexism and the dislike of authorities that make the message. The rap lyrics, and that brings us closer to hand and bodily gestures, also invite us to study the miming of meaning in words and phrases (iconicity) [2].

In this paper we look at a hardly studied phenomenon of rap performances: the series of gestures and bodily movements that are made by rappers while performing. It is not meant as a study of a rapper's movements and gestures from the point of views of the issues mentioned above. Our modest study consists of observing various rappers with the aim of distinguishing characteristic movements and regenerating them in an interactive and entertaining virtual rapper. Clearly, since the lyrics contain an enormous amount of violence, F-words, sexual references, obscenities and derogatory terms for women, one may expect that movements and gestures will reflect that. This is certainly true for iconic gestures that appear in series of gestures and movements. For example, hand signs can indicate gang (Crips, Brims, Bishop, etc.) or certain attitudes (f*ck you) and obscene hand gestures accompany the obscenities in the lyrics. How culture reflects in movements has not yet been studied. Rappers or rap-fans

M. Maybury et al. (Eds.): INTETAIN 2005, LNAI 3814, pp. 134 – 143, 2005.

make distinctions between East, North, West and South, but they as easily apply it to a continent, a country or a city.

As mentioned, we concentrated on the recognition of the characteristics of the sequences of gestures and movements rappers are making. Can we distinguish different styles, can we distinguish and characterize different rap gestures and body movements and can we model and animate them in a virtual rap dancer?

These questions have a much more general background. This background is about the design of creating autonomous embodied agent whose behavior is influenced with continuous real time input and feedback to what it perceives and is able to present through its different sensor and output channels. Its sensor channels may exist of video and audio channels, but we may also think of other, preferably non-obtrusive sensors, like chips, tags, and wearables. Hence, in such a situation, the main issues in the behavior of an embodied agent are the coordination of its behavior with the sensor input with appropriate timing and the selection and execution of the behavior in its own style. That is, there need to be fusion of information coming from different media sources to such an agent and there needs to be fission of information to be presented by an embodied agent.

Again, this is a preliminary study. We looked at the behavior of rappers. We extracted and analyzed their movements from video-clips. We imitated these movements, made modest attempts to distinguish styles, animated these movements, and designed and build a system that uses our findings in two different ways. Firstly, there is the design of a virtual dancer that moves along with the music of a performing rapper. This dancer, a simple VRML avatar, globally follows the music and retrieves its movements and gestures from a database. A slightly more sophisticated system has been obtained where the virtual rap dancer gets its input not from a real-time analyzed rap song, but rather from input from microphone and camera, that allows a human user to steer the rap movements of the virtual dancer. Hence, in this paper we present a virtual dancer that is able to dance to the beat of music coming in through a microphone and to motion beats detected in the video stream of a human dancer. In the current version the moves of the dances are generated from a lexicon that was derived manually from the analysis of the video clips of nine rap songs of different rappers. The system also allows for adaptation of the moves in the lexicon on the basis of style parameters.

2 Related Work

Above we gave a general characterization of our work. That is, we have an embodied agent that is able, through its sensors (audio, vision, keyboard, and maybe others) to sense the environment and possible interaction partners. Moreover, it has means to display its understanding of the communication situation, admittedly in a very limited way, that is, by displaying (nonverbal) rapping behavior in an embodied agent.

Embodied agent design and interaction design with embodied agents has become a well-established research field ([3,4,5]). Hence, when looking at related research we can take the view of humans communicating with embodied agents, where, in e-commerce application, the agents have to sell and demonstrate products, where the agents have to guide a visitor in a virtual or augmented reality environment, or where

the agent plays a role in a training, education, entertainment, or simulation environment. Our aim, to accept nonverbal input and to transform this input to nonverbal output, is not essentially different from much of ongoing recent research, but it certainly differs from regular research because of its emphasis on nonverbal input in the form of music, beats and physical movements, and its nonverbal output in the form of series of rap gestures and body movements in virtual reality.

Hence, here we are more interested in interactive systems that are able to provide various kinds of feedback to expressivity in music, dance, gestures and theatretical performance and that view has been chosen to look at related work. Dance, music and how to interact with computerized dance and music applications are the most important issues we address in the following subsection. In the section on conclusions and future work (Section 5) we look at other related work that we consider being important for the further development of our virtual rap dancer.

2.1 Music, Dance and Interaction: Some Previous Research

There have been many research and art projects where movements of dancers or players are captured by motion capture sensors or cameras. The movements, together with other input signals (e.g., speech, facial expressions, and haptic information) can be analyzed, manipulated and mapped on avatars, (semi-) autonomous agents, robots, or more abstract audio and visual representations. This mapping or re-generation can be done in real-time, allowing applications such as interactive theatre and virtual storytelling (see [6,7] for some pioneering work), or off-line, allowing more advanced graphics and animation, but less direct interactive applications. These latter applications are for example the simulation of traditional Japanese [8] or baroque dances [9] with virtual characters and the generation of new dances composed from extracted primitive dance movements and newly made or learned new choreographies. Attempts to identify emotions expressed in dance or act movements and choreography can also be made part of the mapping from dance to a re-generation.

There are also many examples of music interfaces. Music is used as input and analysis and interpretation allows for applications for education, entertainment and cultural heritage. Previously we have looked at recognizing and distinguishing percussion instruments and music visualization by an embodied performer [10]. The performer is a 3D animation of a drummer, playing along with a given piece of music, and automatically generated from this piece of music. The input for this virtual drummer consists of a sound wave that is analyzed to determine which parts of the percussion instruments are struck at what moments. The Standard MIDI File format is used to store the recognized notes. From this higher-level description of the music, the animation is generated. In an interactive version of this system we use a baton-based interface (using sensors on the tip of two drumsticks), as has been done by many others. Goto et al. [11] introduced an embodied agent that enables a drummer and a guitarist connected through Ethernet not only to musically interact with each other, but also, by observing this virtual character, through the animations of the character. Motion timing for the character comes from the drum, performed dance motions are chosen from the improvisations of the guitarist. A jam session system that allows a human guitarist to interplay with (non-embodied) virtual guitar players can be found in [12]. Different reaction models for human players can be obtained and imitated. An

example of a system that extract acoustical cues from an expressive music performance in order to map them on emotional states is CUEX (CUe EXtraction) [13]. This system has been interfaced with Greta, an embodied conversational agent in which the emotional states obtained from the music are transformed to smoothly changing facial expressions providing a performer with visual feedback [14]. In [15] a system is described that allows a musician to play a digital piano and sing into a microphone, from which musical features are extracted (pitch, specific chords), and responsive behavior in synthetic characters is displayed in real-time. A cognition layer of their system incorporates rules about the relationship between music and emotion. In the expression layer emotions are translated to parameters that guide the character's animations [16]. Well known research on extracting emotions from dance, body movements and gestures is performed by Camurri (see e.g., [17,18]).

3 Analysis of Rap Gesture Sequences

Various rap-video clips have been analyzed by our students[1], characteristic movements have been extracted and, in order to get more feeling for them, exercised. The following music video clips were selected for further study:

- Dr. Dre - Forgot about Dre (Westside)
- Jay Z - Big Pimpin' (Eastside)
- Xzibit - Paparazzi (Westside)
- Blackstreet & Dr. Dre - No diggity (Westside)
- Westside connection - Gangsta nation (Westside)
- Erick Sermon, Redman & Keith Murray - Rappers Delight (Eastside)
- LL Cool J - Headsprung
- KRS One - Hot (Eastside)

Frame by frame movements in these clips have been studied and positions of limbs and body of performers have been extracted. One of the questions that we hoped to answer is whether different bodily gesture styles can be distinguished in rap music and, if so, how we extract features to recognize them. It turned out that although rappers repeat their own movements and gestures, different rappers very much have their own style. One global observation, not based on many data, is that Westside rappers generally move more aggressive, faster and more angular, while Eastside rappers move more relaxed.

In Table 1 we mention the 14 rap movement sequences that have been distinguished in the clips and that have been selected for a database from which our virtual rapper is fed.

Table 1. Fourteen sequences of bodily gestures observed from rap video clips

Yo! Get down!	Bombastic	Bitching ho no!	Cross Again
Drive by	Hit your Cap	Hold me	Kris Kros
Magic Feet	.45	Lullaby	Wuzza
Yo Hallelujah	Wave, get down		

[1] D.F. van Vliet, W.J. Bos, R. Broekstra and J.W. Koelewijn.

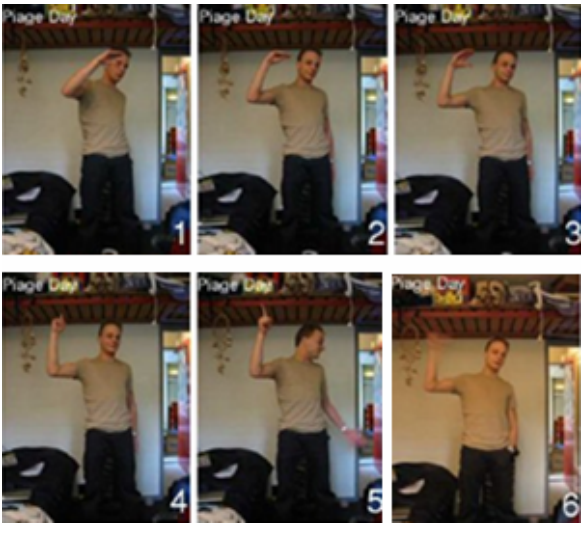

Fig. 1. Bitching ho no!

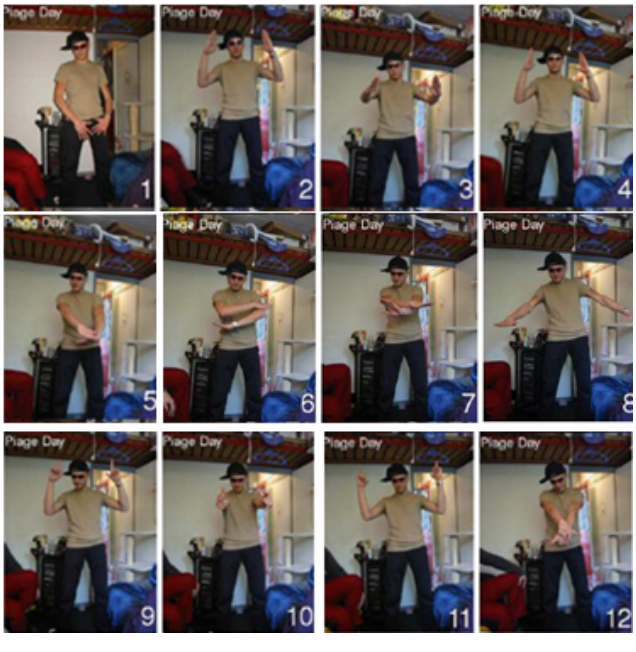

Fig. 2. Kris Kros

All movement sequences were studied, exercised and photographed, requiring, depending on the complexity of the movement, five to eighteen positions to be distinguished. In Figure 1 we illustrate the rather simple 'Bitching ho no!' movement sequence. Another, more complicated movement sequence ('Kris Kros') is

illustrated in Figure 2. Detailed verbal descriptions for these movement sequences have been made in order to allow corresponding animations of an avatar. For example, the Bitching ho no! sequence is a movement sequence where only the right arm is used. First the right hand is moving at head height from the right of the head to the head, making three duck quack movements, next the right hand index finger turns around on the right-hand side of the head. All sequences, with detailed information can be found in [19].

4 Architecture of the Virtual Rap Dancer

The (real time) architecture of the virtual rap dancer consists of several main parts. The sensor channels analyze incoming audio and video in order to detect beats in each separate channel to which the dancer can time its dance moves. The beat predictor module combines the different streams of detected beats, trying to merge beats that were detected in two different channels at the same time into one beat and trying to predict when a next beat is most likely to occur. This prediction is then used by the motion controller that will plan a next dance move in such a way that its focus point will coincide with the predicted next beat. The animation system finally will execute the planned movements after adapting them to some style parameters. The full architecture has already been implemented. The separate modules though are still in a first stage: each module will in the future be extended to achieve a more advanced system.

4.1 Analyzing and Combining Input

There are two types of modules that analyze the input (though of each type, a larger number of instances can be present in a running system). One takes an audio signal from the microphone; the other takes a video signal from a camera. Both modules attempt to recognize beats in the input, albeit in a very simple way in this version of the system. As soon as a module recognizes a beat, it will send out a BeatEvent. The audio system does this largely based on the energy of the audio signal. Clearly, vocal percussion is also possible and, similar as in the first (non-interactive) version of the system, any rap song can be given as input.

The video system tracks the face and hands of the person in the camera view (based on the Parlevision system[2] described in [20]), and will recognize beats based on the hands or face crossing (implicit) trigger-lines in the image. Two incoming beats that are too close together are assumed to be the same beat recognized by different sources. A simple beat-prediction algorithm takes the time-span between the previous beat (from any source) and this beat, and uses that to predict when a next beat is expected. In Figure 3 we show both the audio and video input and manipulation by the system.

[2] http://hmi.ewi.utwente.nl/showcase/parlevision/

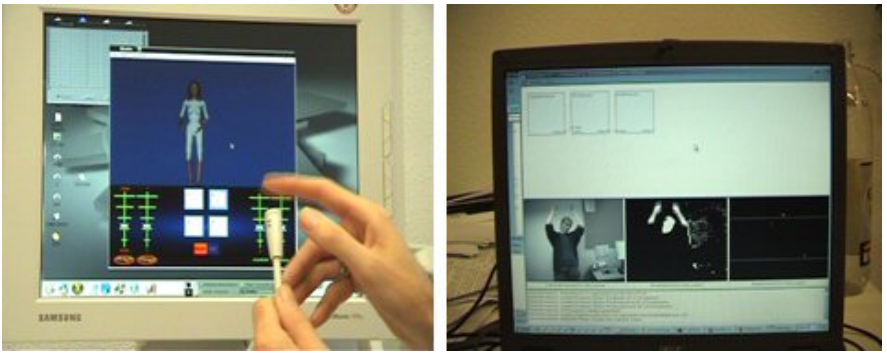

Fig. 3. Microphone tap input (*left*) and dance movement analysis (*right*) for beat detection

4.2 Generating Dance Moves from the Input and Some Style Parameters

The animation module uses a database of rap-dance moves which was constructed from the analysis of example dancers. A large number of rap-clip videos were analyzed. Standard recurring moves were stored in a database by extracting joint angles for key frames, plus some extra information such as the type of rap music that this moved was most often used in and some information about which key frames are to be performed on a (musical) beat (focus key frames).

The movement controller selects the next moves from the database to be executed based on the style parameter indicating what type of dance the user wants to see. These moves are then planned using the information about which key frames are to be aligned to musical beats, the incoming beat information, and the predicted next beat. If the beat predictor is able to successfully predict the next beat the movement planner will cause the dancer to be exactly at a focus key frame when that next beat occurs. Furthermore the dance moves are, before execution, modified with a style-adaptation. At the moment the only style adaptations that are actually implemented are one that

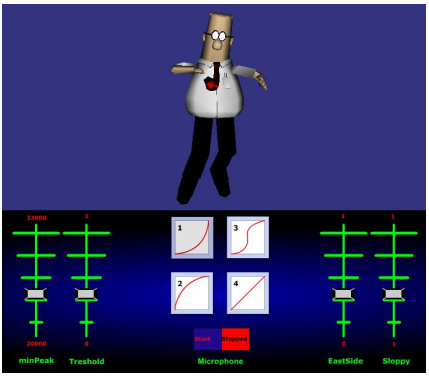

Fig. 4. Style settings for the virtual rap dancer

causes the dancer to dance sharper on the beat or more loosely around the beat and one that modifies the shape of the interpolation between key frames. In the future the system should also allow adaptation of more formation parameters.

The dance moves are animated using the animation package developed at HMI, used before in [21]. Summarizing, presently the user can interact with the system by dancing in front of the camera, giving music input to the microphone or changing the style settings (for now, move-selection parameters and sloppiness of execution). See Figure 4 for the interface with sliders.

5 Evaluation

Although the work described in this paper is sill in a very early stage, some evaluative remarks can already be made. In the first place, the beat-boxing interface using the microphone allows us to perform a (simplistic) qualitative evaluation of the beat recognition and the movement synchronization. Tapping on the microphone in a regular measure does indeed lead to the virtual dancer executing moves in that same measure. Tapping faster or slower causes the dancer, as one would expect for such a simple test, to move faster or slower in the same amount. When music is used as input, it is harder to give a good evaluation of the rhythmic quality of the movements. In general, strong beats are recognized better than subtle rhythmic movements. This means that music with a lot of drum and bass in it works better than, say, classical music.

6 Conclusions and Future Work

The architecture as described above has been fully implemented[3]. It works in real time, integrating microphone input, animation, and the video input (running on a separate computer). Future work will consist, among other things, of improved beat extraction, both from audio [22] and video [8], better beat prediction and fusion of different streams of beat extraction [23], looking at the relationship between joint angles and beats. Results from music retrieval, e.g. query by beatboxing [24], can be used to have the virtual rap dancer choose her movements based on vocal percussion.

The obvious next step, as should become clear from our discussion of related work, is to include more interaction between the virtual rap dancer and its human partner, by extracting motion primitives and looking at the expressiveness of bodily gestures (related to the rap) and make a translation to emotion primitives that impact the choice of bodily gestures, the hand movements and the expressiveness of the virtual rap dancer. While one the one hand the rapper can learn from its human partner and adapt its movements to those of the human partner, on the other hand, the virtual rap dancer can act as a teacher to its human partner. This would mean a great step closer to a real 'joint dancing experience'. The main challenge there would be to adapt the algorithms, developed for the very precise movement and posture capturing methods, to the much less precise motion and posture recognition from camera images. More inspiring creative ideas about possible interaction between rappers and a VJ can be found in a clip of video director Keith Schofield: "3 Feet Deep" of a rap performed by DJ Format.

[3] See http://hmi.ewi.utwente.nl/showcase/v-rapper/ for a live demonstration of the result.

Acknowledgements

The first version of the virtual rap dancer, which did not include interactivity, was made by four of our students (D.F. van Vliet, W.J. Bos, R. Broekstra and J.W. Koelewijn). They were also responsible for the analysis of the video clips and the identification of the 14 rap movement sequences. Ronald Poppe helped us with his image processing software to identify beats in the movements of the human dancer interacting with the rapper. Dennis Hofs and Hendri Hondorp took care of audio processing software that allows beatboxing.

References

1. T. Rose. Black Noise: Rap Music and Black Culture in Contemporary America. Reed Business Information, Inc. (1994)
2. Attolino, P. Iconicity in rap music: the challenge of an anti-language. Presentation at Fifth International Symposium Iconicity in Language and Literature, Kraków (2005)
3. Ruttkay, Zs., Pelachaud, C. From Brows to Trust. Evaluating embodied conversational agents. Kluwer Academic Publishers, Dordrecht Boston London (2004)
4. Payr, S., Trappl, R. (Eds.). Agent Culture. Human-Agent Interaction in a Multicultural World. Lawrence Erlbaum Associates, Mahwah London (2004)
5. Prendinger, H., Ishizuka, M. (Eds.) Life-Like Characters. Tools, Affective Functions, and Applications. Cognitive Technologies Series, Springer-Verlag Berlin Heidelberg New York (2004)
6. Tosa, N., Nakatsu, R. Emotion recognition-based interactive theatre –Romeo & Juliet in Hades -. Eurographics '99, M.A. Alberti, G. Gallo & I. Jelinek (Eds.) (1999)
7. Pinhanez, C., Bobick, A. Using computer vision to control a reactive graphics character in a theater play. Proceedings ICVS '99 (1999)
8. Shiratori, T., Nakazawa, A., Ikeuchi, K. Rhythmic motion analysis using motion capture and musical information. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (2003) 89–94
9. Bertolo, M., Maninetti, P., Marini, D. Baroque dance animation with virtual dancers. Eurographics '99, M.A. Alberti, G. Gallo & I. Jelinek (Eds.) (1999)
10. Kragtwijk, M., Nijholt, A., Zwiers, J. An animated virtual drummer. International Conference on Augmented, Virtual Environments and Three-dimensional Imaging (ICAV3D), V. Giagourta and M.G. Strintzis (eds.), Mykonos, Greece (2001) 319-322
11. Goto, M., Muraoka, Y. A Virtual Dancer "Cindy" - Interactive Performance of a Music-controlled CG Dancer, Proceedings of Lifelike Computer Characters '96 (1996) 65
12. Hamanaka, M., Goto, M., Asoh, H., Otsu, N. A learning-based jam session system that imitates a player's personality model. Proceedings International Joint Conference on Artificial Intelligence (2003) 51-58
13. Friberg, A., Schoonderwaldt, E., Juslin, P.N., Bresin, R. Automatic real-time extraction of musical expression. International Computer Music Conference - ICMC 2002, San Francisco International Computer Music Association (2002) 365-367
14. Mancini, M., Bresin, R., Pelachaud, C. From acoustic cues to expressive ECAs. 6th International Workshop on Gesture in Human-Computer Interaction and Simulation. Valoria, Université de Bretagne Sud, France (2005)
15. Taylor, R., Torres, D., & Boulanger, P. Using music to interact with a virtual character. International Conference on New Interfaces for Musical Expression (NIME05), Vancouver, BC, Canada (2005) 220-223

16. Taylor, R., Boulanger, P., Torres, D. Visualizing emotion in musical performance using a virtual character. 5th International Symposium on Smart Graphics, Germany (2005)
17. Camurri, A., Lagerlöf, I., Volpe, G. Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques. International Journal of Human-Computer Studies, 59(1-2) (2003) 213-225
18. Camurri A., Mazzarino B., Volpe, G. Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library. In A. Camurri, G. Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction, LNAI 2915, Springer-Verlag Berlin Heidelberg New York (2004)
19. Piage Day Project: Internal documentation. University of Twente (2004)
20. Poppe, R., Heylen, D., Nijholt, A., Poel, M. Towards real-time body pose estimation for presenters in meeting environments. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'2005) (2005)
21. Nijholt, A., Welbergen, H., Zwiers, J. Introducing an embodied virtual presenter agent in a virtual meeting room. In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2005) (2005) 579–584
22. Goto, M. An audio-based real-time beat tracking system for music with or without drum-sounds. Journal of New Music Research, 30 (2) (2001)
23. Kim, T-h., Park, S., Shin, S.Y. Rhythmic-motion synthesis based on motion-beat analysis. ACM Trans. Graph., 22 (3) (2003) 392–401
24. Kapur, A., Benning, M., Tzanetakis, G. Query by Beatboxing: Music Information Retrieval for the DJ. Proceedings of the International Conference on Music Information Retrieval, Barcelona, Spain (2004)