SceneMaker: Intelligent Multimodal Visualisation of Natural Language Scripts

Eva Hanser, Paul Mc Kevitt, Tom Lunney, and Joan Condell

School of Computing & Intelligent Systems
Faculty of Computing & Engineering
University of Ulster, Magee
Derry/Londonderry BT48 7JL
Northern Ireland
hanser-e@email.ulster.ac.uk,
{p.mckevitt,tf.lunney,j.condell}@ulster.ac.uk

Abstract. Producing plays, films or animations is a complex and expensive process involving various professionals and media. Our proposed software system, *SceneMaker*, aims to facilitate this creative process by automatically interpreting natural language film/play scripts and generating multimodal, animated scenes from them. During the generation of the story content, SceneMaker will give particular attention to emotional aspects and their reflection in fluency and manner of actions, body posture, facial expressions, speech, scene composition, timing, lighting, music and camera work. Related literature and software on Natural Language Processing, in particular textual affect sensing, affective embodied agents, visualisation of 3D scenes and digital cinematography are reviewed. In relation to other work, SceneMaker will present a genre-specific text-to-animation methodology which combines all relevant expressive modalities and is made accessible via web-based and mobile platforms. In conclusion, SceneMaker will enhance the communication of creative ideas providing quick pre-visualisations of scenes.

Key words: Natural Language Processing, Text Layout Analysis, Intelligent Multimodal Interfaces, Affective Agents, Genre Specification, Automatic 3D Visualisation, Affective Cinematography, SceneMaker

1 Introduction

The production of plays or movies is an expensive process involving planning, rehearsal time, actors and technical equipment for lighting, sound and special effects. It is also a creative act which requires experimentation, visualisation of ideas and their communication between everyone involved, e.g., playwrights, directors, actors, cameramen, orchestra, managers, costume and set designers. We are developing a software system, *SceneMaker*, which will assist in this production process. SceneMaker will provide a facility to test and pre-visualise scenes

before putting them into action. Users input a natural language (NL) text scene and automatically receive multimodal 3D visualisations. The objective is to give directors or animators a reasonable idea of what a scene will look like. The user can refine the automatically created output through a script and 3D editing interface, accessible over the internet and on mobile devices. Thus SceneMaker will be a collaborative tool for script writers, animators, directors and actors, sharing scenes online. Such technology could be applied in the training of those involved in scene production without having to utilise expensive actors and studios. Alternatively, it could be used for rapid visualisation of ideas and concepts in advertising agencies.

SceneMaker will extend an existing software prototype, CONFUCIUS [1], which provides automated conversion of single natural language sentences to multimodal 3D animation of characters' actions and camera placement. SceneMaker will focus on the precise representation of emotional expression in all modalities available for scene production and especially on most human-like modelling of body language and genre-sensitive art direction. To achieve this, SceneMaker will include new tools for text layout analysis of screenplays or play scripts, commonsense and affective knowledge bases for context understanding, affective reasoning and automatic genre specification. This research focuses on three research questions: How can emotional information be computationally recognised in screenplays and structured for visualisation purposes? How can emotional states be synchronised in presenting all relevant modalities? Can compelling, life-like and believable animations be achieved?

Section 2 of this paper gives an overview of current research on computational, multimodal and affective scene production. In section 3, the design of SceneMaker is discussed. SceneMaker is compared to related multimodal work in section 4 and Section 5 discusses conclusions and future work.

2 Background

Automatic and intelligent production of film/theatre scenes with characters expressing emotional states involves four development stages:

1. Detecting personality traits and emotions in the film/play script
2. Modelling affective 3D characters, their expressions and actions
3. Visualisation of scene environments according to emotional findings
4. Development of a multi-modal user interface and mobile application.

This section reviews state-of-the-art advances in these areas.

2.1 Detecting Personality and Emotions in Film/Play scripts

All modalities of human interaction express personality and emotional states namely voice, word choice, gestures, body posture and facial expression. In order to recognise emotions in text and to create life-like characters, psychological theories for emotion, mood, personality and social status are translated into

computable methods, e.g Ekman's 6 basic emotions [2], the Pleasure-Dominance-Arousal model (PAD) [3] with intensity values or the OCC model (Ortony-Clore-Collins) [4] with cognitive grounding and appraisal rules. Word choice is a useful indicator for the personality of a story character, their social situation, emotional state and attitude. Different approaches to textual affect sensing are able to recognise explicit affect words such as keyword spotting and lexical affinity [5], machine learning methods [6], hand-crafted rules and fuzzy logic systems [7] and statistical models [6]. Common knowledge based approaches [8, 9] and a cognitive inspired model [10] include emotional context evaluation of non-affective words and concepts. The unified writing style and strict formatting of screenplays and play scripts eases the machine parsing of scripts and facilitates the detection of semantic context information for visualisation. The prose of scene descriptions focuses on what is audible and visible. Through text layout analysis of capitalisation, indentation and parentheses, elements such as dialog, location, time, present actors, actions and sound cues can be visually recognised and directly mapped into XML-presentation [11].

2.2 Modelling Affective Embodied Agents

Research aiming to automatically model and animate virtual humans with natural expressions faces challenges not only in automatic 3D character manipulation/transformation, synchronisation of face expressions, e.g., lips and gestures with speech, path finding and collision detection, but furthermore in the refined sensitive execution of each action. The exact manner of an affective action depends on intensity, fluency, scale and timing and impacts on the viewer's interpretation of the behaviour. Various scripting languages specifically cater for the modelling of the detected emotions and affective behaviour characteristics.

Non-verbal behaviour of avatars is automatically modelled from conversational text with the Behaviour Expression Animation Toolkit (BEAT) [12]. Based on the analysis of linguistics and context of dialogue scripts appropriate Multimodal Presentation Mark-up Language (MPML) [13] annotations are automatically added to model speech synthesis, facial and body animations of 3D agents. SCREAM (Scripting Emotion-based Agent Minds) [14] is a web-based scripting tool for multiple characters which computes affective states based on the OCC-Model [4] of appraisal and intensity of emotions, as well as social context. ALMA [15] (a Layered Model of Affect) implements AffectML, an XML based modelling language which incorporates the concept of short-term emotions, medium-term moods and long-term personality profiles. The OCEAN personality model [16], Ekman's basic emotions [2] and a model of story character roles are combined through a fuzzy rule-based system [7] to decode the meaning of scene descriptions and to control the affective state and body language of the characters. The high-level control of affective characters in [7] maps personality and emotion output to graphics and animations. Postural values for the four main body areas, head, trunk, upper and lower limbs, manipulate the shape, geometry and motion of the character model based on animation techniques for believable characters [17] considering physical characteristics of space, timing,

velocity, position, weight and portion of the body. Embodied Conversational Agents (ECA) are capable of real-time face-to-face conversations with human users or other agents, generating and understanding NL and body movement. The virtual human, Max [18], engages museum visitors in small talk. Max listens while the users type their input, reasons about actions to take, has intention and goal plans, reacts emotionally and gives verbal and non-verbal feedback. Greta [19] is modelled as an expressive multimodal ECA. The Affective Presentation Markup Language (APML) defines her facial expressions, hand and arm gestures for different communicational functions and with varying degrees of expressivity (manner). The behaviours are synchronised to the duration of phonemes of speech.

Multimodal annotation coding of video or motion captured data specific to emotion collects data in publicly available facial expression or body gesture databases [20]. The captured animation data can be mapped to 3D models, which is useful for instructing characters precisely on how to perform desired actions.

2.3 Visualisation of 3D Scenes and Virtual Theatre

Visual and acoustic elements involved in composing a virtual story scene, the construction of the 3D environment or set, scene composition, automated cinematography and the effect of genre styles are addressed in complete text-to-visual systems such as SONAS [21], WordsEye [22], CONFUCIUS [1] and ScriptViz [25], and the scene directing system, CAMEO [27]. SONAS constructs a three-dimensional virtual town scenario according to the verbal descriptions of a human user. Besides information on the location, scene visualisation requires consideration of the positioning and interaction of actors and objects, the camera view, light sources and audio like background noises or music. WordsEye depicts non-animated 3D scenes with characters, objects, actions and environments. A database of graphical objects holds 3D models, their attributes, poses, kinematics and spatial relations in low-level specifications. In CONFUCIUS, multimodal 3D animations of single sentences are produced. 3D models perform actions, dialogues are synthesised and basic cinematic principles determine the camera placement.

Another modality, cinematography, can assist in conveying themes and moods in animations. Film techniques are automatically applied to existing animations in [23]. Reasoning about plot, theme, character actions, motivations and emotions, they follow cinematic rules which define appropriate placement and movement of camera, lighting, colour schemes and the pacing of shots. A high-level synchronized Expression Mark-up Language (EML) [24] integrates environmental expressions like cinematography, illumination and music as a new modality into the emotion synthesis of virtual humans. ScriptViz renders 3D scenes from NL screenplays immediately during the writing process, extracting verbs and adverbs to interpret events and states in sentences. The time and environment where a story takes place, the theme the story revolves around and the emotional tone of films, plays or literature classify different genres with distinguish-

able presentation styles. Commonly, genres are categorised into, e.g., action, comedy, drama, horror and romance. Genre is reflected in the detail of a production, exaggeration and fluency of movements, pace (shot length), lighting, colour and camerawork. These parameters are responsible for viewer perception, inferences and expectations and thus for an appropriate affective viewer impression. Cinematic principles in different genres are investigated in [26]. Dramas and romantic movies are slower paced with longer dialogues, whereas action movies have rapidly changing, shorter shot length. Comedies tend to be presented in a large spectrum of bright colours, whereas horror films adopt mostly darker hues. The automatic 3D animation production system, CAMEO, incorporates direction knowledge, like genre and cinematography, as computer algorithms and data to control camera, light, audio and character motions. A system which automatically recommends music based on emotion is proposed by [28]. Associations between emotions and music features in movies are discovered by extracting chords, rhythm and tempo of songs.

2.4 Multimodal Interfaces and Mobile Applications

Technological advances enable multimodal human-computer interaction in the mobile world. The system architecture and rendering can be placed on the mobile device itself or distributed from a server via wireless broadband networks. SmartKom Mobile [29] brings the multimodal system, SmartKom, onto mobile devices. The user interacts with a virtual character through dialogue. Supported modalities include language, gesture, facial expression and emotions carried through speech emphasis. Script writing tools assist the writing process of screenplays or play scripts, like the web-based FiveSprockets [30] or ScriptRight [31] for mobile devices. The Virtual Theatre Interface project [32] offers a web-based user interface to manipulate actors' positions on stage, lighting and audience view points.

The wide range of approaches, presented here, to modelling emotions, moods and personality aspects in virtual humans and scene environments along with first attempts to bring multi-modal agents onto mobile devices provide a sound basis for SceneMaker.

3 Design of SceneMaker

Going beyond the animation of explicit events, SceneMaker will use Natural Language Processing (NLP) methods for screenplays to automatically extract and visualise emotions, moods and film/play genre. The process will be tested with a software prototype, which will augment short 3D scenes with affective influences on the body language of actors and environmental expression, like illumination, timing, camera work, music and sound automatically directed according to the genre style.

3.1 SceneMaker Architecture

SceneMakers's architecture is shown in Fig. 1. The main component is the *scene production module* including modules for understanding, reasoning and multimodal visualisation situated on a server. The *understanding module* performs natural language processing and text layout analysis of the input text. The *reasoning module* interprets the context based on common, affective and cinematic knowledge bases, updates emotional states and creates plans for actions, their manners and the representation of the set environment. The *visualisation module* maps these plans to 3D animation data, selects appropriate 3D models from the graphics database, defines their body motion transitions, instructs speech synthesis, selects sound and music files from the audio database and assigns values to camera and lighting parameters. The visualisation module synchronises all modalities into an animation manuscript. The online user interface, available via computers and mobile devices, consists of two parts. The input module provides assistance for film and play script writing and editing and the output module renders the 3D scene according to the manuscript and allows manual scene editing to fine-tune the automatically created animations.

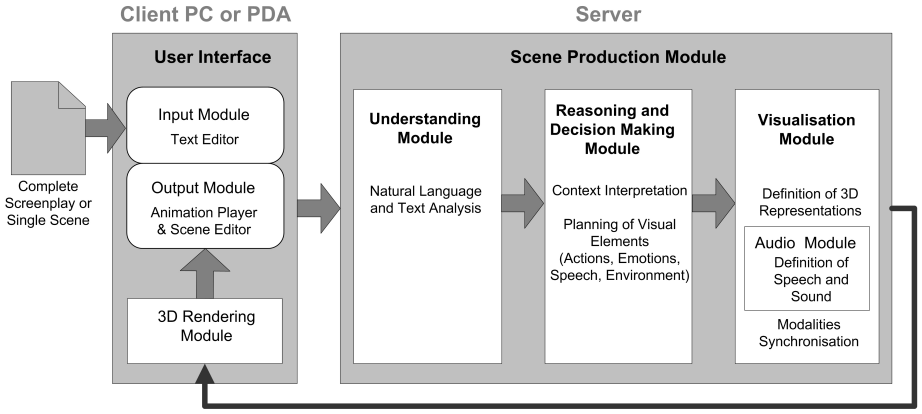

Fig. 1. SceneMaker architecture

3.2 Implementation of SceneMaker

Multimodal systems automatically mapping text to visuals face challenges in interpreting human language which is variable, ambiguous, imprecise and relies on common knowledge between the communicators. Enabling a machine to understand a natural language text involves feeding the machine with grammatical structures, semantic relations and visual descriptions to be able to match suitable graphics. Existing software tools fulfilling sub-tasks will be modified,

combined and extended for the implementation of SceneMaker. For the interpretation of the input scripts, SceneMaker will build upon the NLP module of CONFUCIUS [1], but a pre-processing tool will first decompose the layout structure of the input screenplay/play script. The syntactic knowledge base parses the input text and identifies grammatical word types, e.g., noun, verb, adjective or other, with the Connexor Part of Speech Tagger [33] and determines their relation in a sentence, e.g., subject, verb and object with Functional Dependency Grammars [34]. The Semantic knowledge base (WordNet [35] and LCS database [36]) and temporal language relations will be extended by an emotional knowledge base, e.g., WordNet-Affect [37], and context reasoning with ConceptNet [9] to enable an understanding of the deeper meaning of the context and emotions. In order to automatically recognise genre, SceneMaker will identify keyword co-occurences and term frequencies and determine the length of dialogues, sentences and scenes/shots. In SceneMaker, the visual knowledge of CONFUCIUS, such as object models and event models, will be related to emotional cues. CONFUCIUS' basic cinematic principles will be extended and classified into expressive and genre-specific categories. EML [24] appears to be a comprehensive XML-based scripting language to model expressive modalities including body language as well as cinematic annotations. Resources for 3D models are H-Anim models [38] which include geometric or physical, functional and spatial properties. For the speech generation from dialogue text, the speech synthesis module used in CONFUCIUS, FreeTTS [39], will be tested for its suitability in SceneMaker with regard to mobile applications and the effectiveness of emotional prosody. An automatic audio selection tool, as in [28], will be added for intelligent, affective selection of sound and music according the theme and mood of a scene.

Test scenarios will be developed based on screenplays of different genres and animation styles, e.g., drama films, which include precise descriptions of set layout and props versus comedy, which employs techniques of exaggeration for expression. The effectiveness and appeal of the scenes created in SceneMaker will be evaluated against hand-animated scenes and existing feature film scenes. The functionality and usability of SceneMaker's components and the GUI will be tested in cooperation with professional film directors, comparing the process of directing a scene traditionally with actors or with SceneMaker.

4 Relation to Other Work

Research implementing various aspects of modelling affective virtual actors, narrative systems and film-making applications relates to SceneMaker. CONFUCIUS [1] and ScriptViz [25] realise text-to-animation systems from natural language text input, but they do not enhance the visualisation through affective aspects, the agent's personality, emotional cognition or genre specific styling. Their animation is built from well-formed single sentences and does not consider the wider context. SceneMaker will allow the animation modelling of sentences, scenes or whole scripts. Single sentences require more reasoning about default

settings and more precision will be achieved from collecting context information from longer passages of text. SceneMaker will introduce text layout analysis to derive semantic content from the particular format of screenplays/play scripts. Emotion cognition and display will be related to commonsense knowledge. No previous storytelling system controls agent behaviour through integrating all of personality, social status, narrative roles and emotions. Only EML [24] combines multimodal character animation with film making practices based on an emotional model, but it does not consider personality types or genre. CAMEO [27] is the only system relating specific cinematic direction, for character animation, lighting and camera work, to the genre or theme of a given story, but genre types are explicitly selected by the user. SceneMaker will introduce a new approach to automatically recognise genre from script text with keyword co-occurrence, term frequency and calculation of dialogue and scene length. SceneMaker will bring all relevant techniques together to form a software system for believable affective computational animation production from NL scene scripts. SceneMaker will present a web/mobile based user interface for directors or animators to directly edit scenes.

5 Conclusion and Future Work

SceneMaker contributes to believability and artistic quality of automatically produced animated, multimedia scenes. The software system, SceneMaker, which automatically visualises affective expressions of screenplays, aims to advance knowledge in the areas of affective computing, digital storytelling and expressive multimodal systems. Existing systems solve partial aspects of NLP, emotion modelling and multimodal storytelling. Thereby, this research focuses on semantic interpretation of screenplays or play scripts, the computational processing of emotions, virtual agents with affective behaviour and expressive scene composition including emotion-based audio selection. In relation to other work, SceneMaker will incorporate an expressive model for multiple modalities, including prosody, body language, acoustics, illumination, staging and camera work. Emotions will be inferred from context. Genre types will be automatically derived from the scene scripts and influence the design style of the output animation. The 3D output will be editable on mobile devices. SceneMaker's mobile, web-based user interface will assist directors, drama students, writers and animators in the testing of their ideas. Accuracy of animation content, believability and effectiveness of expression and usability of the interface will be evaluated in empirical tests comparing manual animation, feature film scenes and real-life directing with SceneMaker. In conclusion, this research intends to automatically produce multimodal animations with heightened expressivity and visual quality from screenplay or play script input.

References

1. Ma, M.: Automatic Conversion of Natural Language to 3D Animation. PhD Thesis, School of Computing and Intelligent Systems, University of Ulster. (2006)

2. Ekman, P. and Rosenberg E. L.: What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system. Oxford University Press (1997)
3. Mehrabian, A.: Framework for a Comprehensive Description and Measurement of Emotional States. In: Genetic, Social, and General Psychology Monographs. Heldref Publishing, 121 (3), 339-361 (1995)
4. Ortony A., Clore G. L., and Collins A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, MA (1988)
5. Francisco, V., Hervás, R. and Gervás, P.: Two Different Approaches to Automated Mark Up of Emotions in Text. In: Research and development in intelligent systems XXIII: Proceedings of AI-2006. Springer, 101-114 (2006)
6. Strapparava, C. and Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing. SAC '08. ACM, New York, NY, 1556-1560 (2008)
7. Su, W-P., Pham, B., Wardhani, A.: Personality and Emotion-Based High-Level Control of Affective Story Characters. In: IEEE Transactions on Visualization and Computer Graphics, 13 (2), 281-293 (2007)
8. Liu, H., Lieberman, H., and Selker, T.: A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th International Conference on Intelligent User Interfaces. IUI '03. ACM, New York, 125-132 (2003)
9. Liu, H. and Singh, P.: ConceptNet: A practical commonsense reasoning toolkit. In: BT Technology Journal. Springer Netherlands, 22(4), 211-226 (2004)
10. Shaikh, M.A.M., Prendinger, H. and Ishizuka, M.: A Linguistic Interpretation of the OCC Emotion Model for Affect Sensing from Text. In: Affective Information Processing. Springer London, 45-73 (2009)
11. Choujaa, D. and Dulay, N.: Using screenplays as a source of context data. In: Proceeding of the 2nd ACM International Workshop on Story Representation, Mechanism and Context. SRMC '08. ACM, New York, 13-20 (2008)
12. Cassell, J., Vilhjálmsson, H. H., and Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and interactive Techniques SIGGRAPH '01. ACM, New York, 477-486 (2001)
13. Breitfuss, W., Prendinger, H., and Ishizuka, M.: Automated generation of non-verbal behavior for virtual embodied characters. In: Proceedings of the 9th International Conference on Multimodal Interfaces. ICMI '07. ACM, New York, 319-322 (2007)
14. Prendinger, H. and Ishizuka, M.: SCREAM: scripting emotion-based agent minds. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1. AAMAS '02. ACM, New York, 350-351 (2002)
15. Gebhard, P.: ALMA - Layered Model of Affect. In: Proceedings of the 4th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 05). Utrecht University, Netherlands. ACM, New York, 29-36. (2005)
16. De Raad, B.: The Big Five Personality Factors. In: The Psycholexical Approach to Personality. Hogrefe & Huber (2000)
17. Thomas, F. and Johnson O.: The Illusion of Life: Disney Animation. Abbeville Press/Hyperion. 47-69 (1981, reprint 1997)
18. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E. and Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans. In: Modeling Communication with Robots and Virtual Humans. Springer Berlin/Heidelberg. 18-37 (2008)

19. Pelachaud, C.: Multimodal expressive embodied conversational agents. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. MULTIMEDIA '05. ACM, New York, 683-689 (2005)
20. Gunes, H. and Piccardi, M.: A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior. In: 18th International Conference on Pattern Recognition, ICPR. IEEE Computer Society, Washington, 1, 1148-1153 (2006)
21. Kelleher, J., Doris T., Hussain, Q. and Nuallin, S.: SONAS: Multimodal, Multi-User Interaction with a Modelled Environment. In: Spatial Cognition. John Benjamins Publishing Company, Amsterdam, Netherlands, 171-184 (2000)
22. Coyne, B. and Sproat, R.: WordsEye: an automatic text-to-scene conversion system. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press, Los Angeles, 487-496 (2001)
23. Kennedy, K. and Mercer, R. E.: Planning animation cinematography and shot structure to communicate theme and mood. In: Proceedings of the 2nd International Symposium on Smart Graphics. SMARTGRAPH '02. ACM, New York, 24, 1-8 (2002)
24. De Melo, C. and Paiva, A.: Multimodal Expression in Virtual Humans. In: Computer Animation and Virtual Worlds 2006. John Wiley & Sons Ltd. 17 (3-4), 239-348 (2006)
25. Liu, Z. and Leung, K.: Script visualization (ScriptViz): a smart system that makes writing fun. In: Soft Computing, Springer Berlin/Heidelberg, 10, 1, 34-40 (2006)
26. Rasheed, Z., Sheikh, Y., Shah, M.: On the use of computable features for film classification. In: IEEE Transactions on Circuits and Systems for Video Technology. IEEE Circuits and Systems Society, 15(1), 52-64 (2005)
27. Shim, H. and Kang, B. G.: CAMEO - camera, audio and motion with emotion orchestration for immersive cinematography. In: Proceedings of the 2008 international Conference on Advances in Computer Entertainment Technology. ACE '08. ACM, New York, NY. 352, 115-118 (2008)
28. Kuo, F., Chiang, M., Shan, M., and Lee, S.: Emotion-based music recommendation by association discovery from film music. In: Proceedings of the 13th Annual ACM international Conference on Multimedia, MULTIMEDIA '05. ACM, New York, 507-510 (2005)
29. Wahlster, W.: Smartkom : Foundations of Multimodal Dialogue Systems. Springer Verlag (2006)
30. FiveSprockets, http://www.fivesprockets.com/fs-portal
31. ScriptRight, http://www.scriptright.com
32. Virtual Theatre Interface, http://accad.osu.edu/research/virtual_environment_htmls/virtual_theatre.htm
33. Connexor, http://www.connexor.eu/technology/machinese
34. Tesniere, L.: Elements de syntaxe structurale. Klincksieck, Paris (1959)
35. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press. Cambridge (1998)
36. Lexical Conceptual Structure Database, http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html
37. Strapparava, C. and Valitutti, A.: WordNet-Affect: an affective extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004. 4, 1083-1086 (2004)
38. Humanoid Animation Working Group, http://www.h-anim.org
39. FreeTTS 1.2 - A speech synthesizer written entirely in the JavaTM programming language: http://freetts.sourceforge.net/docs/index.php