SceneMaker: Automatic Visualisation of Screenplays

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. Our proposed software system, *SceneMaker*, aims to facilitate the production of plays, films or animations by automatically interpreting natural language film 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. 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 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., play writers, directors, actors, camera men, orchestra and set designers. We are developing a software system, *SceneMaker*, which will provide a facility to pre-visualise scenes. Users input a natural language (NL) script text 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. 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 character actions. 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. SceneMaker will include new tools for text layout analysis of screenplays, commonsense and affective knowledge bases for context understanding, affective reasoning and automatic genre specification. This work 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 the conclusion 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 script
2. Modelling affective 3D characters, their expressions and actions
3. Visualisation scene environments according to emotional findings
4. Intelligent storytelling interpreting the plot.

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

2.1 Detecting Personality and Emotions in Film scripts

All modalities of human interaction, namely voice, word choice, gestures, body posture and facial expression, express personality and emotional states. 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 models, 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. Different approaches to textual affect sensing are able to recognise explicit affect phrases such as keyword spotting and lexical affinity [5], machine learning methods [6], handcrafted 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 strict

formatting of screenplays eases the machine parsing of scripts and facilitates the detection of semantic context information for visualisation. Through text layout analysis of capitalisation, indentation and parentheses, elements such as dialog, location, time, present actors, actions and sound cues can be recognised and directly mapped into XML-presentation [11].

2.2 Modelling Affective Embodied Agents

Research aiming to automatically animate virtual humans with natural expressions faces challenges not only in automatic 3D character transformation, synchronisation of face and body expressions 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. Various XML based scripting languages specifically cater for the modelling of affective behaviour, e.g., the Behaviour Expression Animation Toolkit (BEAT) [12], the Multimodal Presentation Mark-up Language (MPML) [13], SCREAM (Scripting Emotion-based Agent Minds) [14] and AffectML [15]. The Personality & Emotion Engine [7], a fuzzy rule-based system, combines the OCEAN personality model [16], Ekman's basic emotions [2] and story character roles to control the affective state and body language of characters mapping emotions to postural values of four main body areas. Embodied Conversational Agents (ECA) are capable of face-to-face conversations with human users or other agents, generating and understanding NL and body movement, e.g., Max [17] and Greta [18]. Multimodal annotation coding of video or motion captured data specific to emotion collects data in facial expression or body gesture databases [19]. The captured animation data can be mapped to 3D models instructing characters precisely on how to perform desired actions.

2.3 Visualisation of 3D Scenes

The composition of the 3D scene environment or set, automated cinematography, acoustic elements and the effect of genre styles are addressed in text-to-visual systems. WordsEye [20] depicts non-animated 3D scenes with characters, objects, actions and environments considering the attributes, poses, kinematics and spatial relations of 3D models. CONFUCIUS [1] produces multimodal 3D animations of single sentences. 3D models perform actions, dialogues are synthesised and basic cinematic principles determine the camera placement. ScriptViz [23] renders 3D scenes from NL screenplays, extracting verbs and adverbs to interpret events in sentences. Film techniques are automatically applied to existing animations in [21]. Reasoning about plot, theme, character actions, motivations and emotions, cinematic rules define the appropriate placement and movement of camera, lighting, colour schemes and the pacing of shots according to theme and mood. The Expression Mark-up Language (EML) [22] integrates environmental expressions like cinematography, illumination and music into the emotion synthesis of virtual humans. Films, plays or literature are classified into different genres with distinguishable presentation styles, e.g., drama or comedy. Genre is

reflected in the detail of a production, exaggeration and fluency of movements, pace (shot length), lighting, colour and camerawork [24]. The automatic 3D animation production system, CAMEO [25], incorporates direction knowledge, like genre and cinematography, as computer algorithms and data. A system which automatically recommends music based on emotion [26] associates emotions and music features, chords, rhythm and tempo of songs, in movie scenes.

2.4 Intelligent Storytelling

Intelligent, virtual storytelling requires the automatic development of a coherent plot, fulfilling thematic, dramatic, consistency and presentation goals of the author, as in MINSTREL [27]. The plot development can be realised with a character-based approach, e.g., AEOPSWORLD [28], where characters are autonomous intelligent agents choosing their own actions or a script-based approach, e.g., Improv [29], where characters act out a given scene script or an intermediate variation of the two.

The wide range of approaches for intelligent storytelling and modelling emotions, moods and personality aspects in virtual humans and scene environments 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 and moods within the story or scene context. SceneMaker 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. 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 the input module, assisting film script writing and editing and the output module, rendering 3D scene according to the manuscript and allowing manual scene editing to fine-tune the automatically created animations.

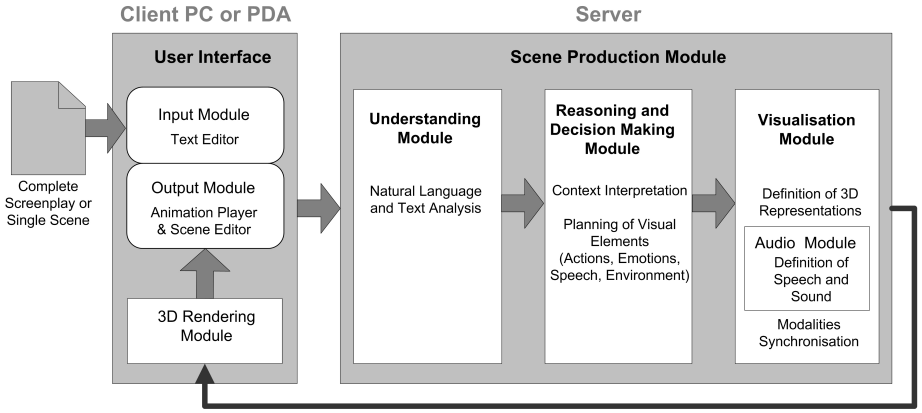

Fig. 1. SceneMaker architecture

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 multimodal system with grammatical structures, semantic relations, visual descriptions and common knowledge to be able to match suitable graphics. A pre-processing tool will decompose the layout structure of the input screenplay to facilitate access to semantic information. SceneMaker's language interpretation will build upon the NLP module of CONFUCIUS [1]. The Connexor Part of Speech Tagger [30] parses the input text and identifies grammatical word types, e.g., noun, verb or adjective, and determines their relation in a sentence, e.g., subject, verb and object with Functional Dependency Grammars [31]. CONFUCIUS's semantic knowledge base (WordNet [32] and LCS database [33]) will be extended by an emotional knowledge base, e.g., WordNet-Affect [34], 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. 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. Resources for 3D models are H-Anim models [35] which include geometric or physical, functional and spatial properties. The speech synthesis module used in CONFUCIUS, FreeTTS [36], will be tested for its suitability with regard to effective emotional prosody. An automatic audio selection tool, as in [26], will be added for intelligent, affective selection of sound and music according the theme and mood of a scene.

4 Relation to Other Work

Research implementing various aspects of modelling affective virtual actors, narrative systems and film-making applications relates SceneMaker. CONFUCIUS [1] and ScriptViz [23] 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 single sentences and does not consider the wider context of the story. 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. Emotion cognition and display will be related to commonsense knowledge. CAMEO [25] 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 automatically recognise genre from script text with keyword co-occurrence, term frequency and calculation of dialogue and scene length. SceneMaker will form a software system for believable affective computational animation production from NL scene scripts.

5 Conclusion and Future Work

SceneMaker contributes to believability and artistic quality of automatically produced animated, multimedia scenes. The software system, SceneMaker, will automatically visualise affective expressions of screenplays. Existing systems solve partial aspects of NLP, emotion modelling and multimodal storytelling. Thereby, this research focuses on semantic interpretation of screenplays, 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 SceneMaker's mobile, web-based user interface and will assist directors, drama students, writers and animators in the testing of scenes. 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 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 (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, 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. 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)

18. 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)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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. 352, 115-118 (2008)
26. 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)
27. Turner, S. R.: The Creative Process: A Computer Model of Storytelling and Creativity. Lawrence Erlbaum Associates. Hillsdale, USA (1994)
28. Okada, N., Inui, K. and Tokuhisa, M.: Towards affective integration of vision, behavior, and speech processing. In: Proceedings of the Integration of Speech and Image Understanding, SPELMG. IEEE Computer Society. Washington, 49-77 (1999)
29. Perlin, K. and Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In: Proceedings of the 23rd Annual Conference on Computer Graphics and interactive Techniques, SIGGRAPH '96, ACM, New York, 205-216 (1996)
30. Connexor, `http://www.connexor.eu/technology/machinese`
31. Tesniere, L.: Elements de syntaxe structurale. Klincksieck, Paris (1959)
32. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press. Cambridge (1998)
33. Lexical Conceptual Structure Database, `http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html`
34. 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)
35. Humanoid Animation Working Group, `http://www.h-anim.org`
36. FreeTTS 1.2 - A speech synthesizer written entirely in the JavaTM programming language: `http://freetts.sourceforge.net/docs/index.php`