SceneMaker:
Intelligent Multimodal Visualisation of Natural Language Scripts

Eva Hanser, Dipl.-Des. (FH), M.Sc.

Supervisors:
Prof. Paul Mc Kevitt, Dr. Tom Lunney, Dr. Joan Condell

Confirmation Report
June, 2009

Presented as a requirement for Ph.D. in
School of Computing & Intelligent Systems
Faculty of Computing & Engineering
University of Ulster, Magee
Derry/Londonderry BT48 7JL
Northern Ireland

E-mail: hanser-e@email.ulster.ac.uk

Abstract

Performing plays or creating films and animations is a complex creative, and thus expensive, process involving various professionals and media. This research project proposes to augment this process by automatically interpreting film and play scripts and automatically generating animated scenes from them. For this purpose a web based software prototype, *SceneMaker*, will be developed. During the generation of the story content, special attention will be given to emotional aspects and their reflection in the execution of all types of modalities: fluency and manner of actions and behaviour, body language, facial expressions, speech, voice pitch, scene composition, timing, lighting, music and camera. The main objective of this work is to demonstrate how a scene and actor behaviour changes, when emotional states are taken into account, e.g. walking down a street in a happy versus a sad state. Consequently, more realistic and believable story visualisations are an expected outcome. Literature on related research areas is reviewed involving natural language and text, screenplay or play script, layout processing with regard to personality and emotion detection, modelling affective behaviour of embodied agents, visualisation of 3D scenes with digital cinematography and genre-based presentation, intelligent multimedia selection and mobile 3D technology.

SceneMaker's architecture with a language and text analysis module, a reasoning and decision making module based on cognitive, emotional information and a multimedia module for multimodal visualisation are presented along with a development plan and test scenarios. Technologies and software relevant for the development of *SceneMaker* are analysed and prospective tools for text and language processing, visualisation, media allocation and mobile user interfaces are suggested. Accuracy of content animation, effectiveness of expression and usability of the interface will be evaluated in empirical tests. In relation to other work, this project will present a genre-specific text-to-animation methodology which combines all relevant expressive modalities. Emotional expressivity inspired by the OCC (Ortony, Clore and Collins) emotion model will influence all modalities to enhance believability of virtual actors as well as scene presentation. Compared to other scene production systems, *SceneMaker* will infer emotions from the story context, rather than relying on explicit emotion keywords, *SceneMaker* will automatically detect genre from the script and apply appropriate cinematic direction. Also *SceneMaker* will enable 3D animation editing via web-based and mobile platforms.

In conclusion, *SceneMaker* will reduce production time, save costs and enhance communication of ideas through the development of an intelligent, multimodal animation generation system providing quick pre-visualisations of scenes from script text input.

***Keywords*:** Natural Language Processing, Script Writing, Text Layout Analysis, Automatic 3D Visualisation, Affective Embodied Agents, Emotion Reasoning, Intelligent Cinematography, Virtual Director, Multimedia Scene Production, Genre Specification, Mobile Multimodal Application, *SceneMaker*

Table of Contents

1 Introduction

This section covers the theoretical foundations of this research as well as highlighting the practical need for, and usefulness of, this project for movie and play production. The research areas and features to be addressed through the development of the software prototype, *SceneMaker*, are presented. The principle aims and objectives of this work are summarized.

1.1 Background – Real Life and Virtual Movie and Play Production

The production of plays or movies is an expensive process involving planning and rehearsal time, actors, technical equipment for lighting, sound and special effects. It is also a creative act which might not always be straightforward, but requires experimentation, visualisation of ideas and their communication between everyone involved, e.g. play writers, directors, actors, camera men, orchestra, managers, costume and set designers. This creative act usually begins with ideas being written down as story scripts, which are then pre-visualised as storyboards or rough animation sketches. In the next step actors practise and perform the scenes and scripts may be altered until the final performance or movie is created. This research intends to simplify the production process by computerising the pre-production stage by automatically producing 3D animation pre-visualisations from written scripts and cutting out storyboarding and many practice runs.

Directors aim for the best possible impact of a scene on the viewer, aesthetically appealing views and believable characters. It is important to not just show an action or event, but how exactly it is shown in cinematic presentation and in the behavioural manner of the actors. Affect and emotions are continuously exchanged in human interaction and manipulate decision making. Accordingly, they are relevant for interpreting and reasoning in storytelling. Realistic and appropriate presentation of emotional aspects attracts the audiences' interest and gives an illusion of believability, thus conveying the right sense or feeling, is a significant element of entertainment, film making and play performance. Consequently, an automatic scene generation system faces challenges in achieving the same quality and sensitivity in its visualisations. Unfortunately the behaviour of digital agents in current systems is still very artificial. Various research investigates human-like reasoning, behaviour models and decision making to model realistic, emotionally influenced behaviour of virtual humans, but results are still limited in expressiveness, because of narrow emotion structures and restricted visual mapping. This project will focus on reflecting individual personality and emotional states of virtual actors impacting on their behaviour, as well as cinematic techniques consistent with the genre, theme and overall moods.

Another focus of this work is the creation of the connection between language elements, written words found in the scene script and matching visual elements, 3d objects and their animation. Language and vision (and music) share meanings, or semantics, which humans hold in their mind as common knowledge. This knowledge has to be captured and structured in databases for computational access and processing. The formal structure of screenplays or play scripts, which defines scene settings, location and time, actions of characters and dialogues well, is intended to support language processing to extract semantic information.

Such technology could be applied in the training of film/drama directors without having to continuously utilise expensive actors and actresses. Alternatively, it may be useful for advertising agencies, which constantly need rapid visualisations of various ideas and concepts. At the Ohio State University a virtual theatre interface (Virtual Theatre, 2004) for teaching drama students about lighting, positioning on stage and different view points was considered very beneficial and improved training methods.

1.2 Key Features of SceneMaker

This research proposes a web based software prototype, *SceneMaker*, which will assist in the film or play production process. *SceneMaker* will provide a facility for everyone involved in the creation of animated scenes to test and pre-visualise scenes, before putting them into action. Users input a natural language text scene script and automatically receive multimodal 3D visualisations taking into account considerations such as aesthetics and emotions. The user can refine the output through an interface which facilitates the direct control of character behaviour, actions and cinematographic settings, e.g. lighting and camera settings, within the 3D environment, as well as access to cognitive settings such as personality and emotional states. *SceneMaker* focuses on the precise representation of emotional expression in all modalities available for scene production and especially on most human-like modelling of body language as it is the most expressive modality in human communication, delivering 60-80 percent of our messages. Actual words only represent 7-10 percent of all modalities delivering a message in conversation (Su et al., 2007). Further modalities include voice tone, volume, facial expression, gaze, gesture, body posture, spatial behaviour and aspects of appearance. These facts show the importance of the visualisation of body language in film/play production, but also point out the challenges in deriving information for animation from scripts containing mostly dialogues. Much research is dedicated to modelling of emotion and facial expressions, gaze and hand gestures (Kopp et al., 2008; Sowa, 2008), but body posture has yet to be addressed extensively (Gunes and Piccardi, 2006).

The *SceneMaker* prototype will be developed using appropriate multimodal technology. Natural language processing, affective reasoning and visualisation tools will be combined to form a unique research contribution. *SceneMaker* will extend an existing software prototype, CONFUCIUS (Ma, 2006), which performs automated conversion of natural language to 3D animation for action verbs. Computational emotion processing can be divided into the five phases: classification, quantification, interaction, mapping and expression (Bartneck, 2002). In the classification phase the emotional effect of the story content – events, actions and objects – is evaluated and appropriate emotion categories are selected according to the chosen emotion model and common knowledge. In the quantification phase the intensity of the triggered emotions is calculated. In the interaction phase the new emotion values are linked to the character’s personality, current mood and circumstances to filter relevant emotions for display. In the mapping phase the resulting emotional values need to be mapped to possible emotional expressions through, e.g. Extensible Markup Language (XML) annotations. Finally, in the expression phase the emotional state is visualised through the defined bodily expressions and behaviour. *SceneMaker* will add screenplay/play script layout analysis and genre recognition to the pre-processing of text. Emotion and genre will be directing the cinematographic presentation of the scene environment affecting lighting and camera work. *SceneMaker* will be accessible over the internet and thus will be an easily available tool for script writers, animators, directors, actors or drama students to creatively and inexpensively express their ideas. They can prove the ideas’ effectiveness in achieving their desired effect whilst writing or advising directors and actors on set. Successful example scenes can be saved and shared with other scene producers. A Graphical User Interface (GUI) suitable for mobile devices is intended to facilitate the use of *SceneMaker* on stage or on set.

1.3 Research Aims and Objectives

This research aims to solve three research questions: How can emotional information be computationally interpreted from screenplays? How can emotional states be structured for visualisation purposes and synchronised in presenting all relevant modalities? Can compelling, life-like animations be achieved? These questions will be addressed through the design and implementation of an automated animation system, *SceneMaker*, featuring affective actor modelling and scene production based on personality, social and narrative roles and emotions. The software prototype, *SceneMaker*, will be a multimodal content generation system, with a user interface for script input and manual manipulation of the automatically outputted scenes, accessible on mobile devices, which can be applied seamlessly for testing and customizing performances according to the producers' intentions. The objective is to give directors or animators a reasonable idea of what the scene they are planning will look like.

The key objectives of this work are:

- To interpret natural language text input, in particular screenplays, extracting semantics and emotional aspects essential for visualization
- To integrate knowledge bases of common sense and emotional relations required for affective reasoning and decision making rules
- To map language elements to visual elements to present actors, objects and actions or events with affective properties
- To generate virtual scenes automatically, with 3D animation, speech and non-speech audio
- To coordinate the timing of different media/modalities to form a coherent animation
- To apply genre-specific presentation aesthetics
- To design and implement a software prototype, *SceneMaker*, realising the above mentioned objectives
- To evaluate and test the effectiveness of *SceneMaker* and the realism achieved compared to other work in the field

This includes potential unique contributions to:

- Enhance believability of virtual actors and scene presentation
- Incorporate the cognitively inspired OCC emotion model within visualisation tools
- Extract emotions from context, rather than explicitly keywords
- Automatically detect genre and animate accordingly
- Editing 3D content on mobile devices

These potential contributions are a means to achieve the main aim of this research namely to automatically create believable characters and realistic presentations and convey meaning in telling stories computationally.

Section 2 of this report gives an overview of current research on scene production for multimodal and interactive storytelling, virtual theatre, affective agents and mobile 3D applications. In section 3, the project proposal and prototype, *SceneMaker,* are described. *SceneMaker* is compared to related multimodal visualisation applications in section 4. Section 5 concludes the report.

2 Literature Review

Successful storytelling, with the focus of this paper being movie or play production, requires sensitivity and tact in the translation and execution of all actions and content. The attention to detail and reflection of subtle expressions, emotions and personality traits is significant for the believability of characters and animations (Bates, 1994). Hence visualisation tools have to provide affective 'understanding' and 'reasoning' similar to human cognitive processes. In order to appear believable virtual scenes do not have to be real-life imitations, but need to be as realistic as filmic/theatrical performances. The automatic and intelligent production of film/theatre scenes, with characters expressing emotional states and with affective design of the animation, involves seven development stages:

1. Semantic processing of the screenplay or play script – to draw context information from the formal text layout of the script and from the dramatic structure of typical stories;
2. Computation of emotions and personality – defining and computationally recognizing emotion and personality models from natural language text;
3. Annotation of emotions and affective behaviour – assigning appropriate behaviour to detected emotions, bringing the behaviour instructions into a computer readable notation;
4. Modelling affective behaviour of embodied agents – applying emotional expressions to 3D models;
5. Multimodal visualisation of 3D scenes and virtual theatre – bringing the scene environment together including automatic cinematic direction and genre-oriented animation styling;
6. Automatic sound selection and affective speech synthesis – adding a further modality reflecting mood and theme through audio;
7. Multimodal web-based and mobile interfaces – introducing intelligent and multimodal applications and 3D rendering techniques on mobile devices.

This section reviews state-of-the-art advances in these areas and discusses related research projects.

2.1 Semantic Processing of Screenplays/Play Scripts

The specific format of scripts allows easy access to content information. The layout analysis alone provides semantic clues. Breaking a script down into its structural components assists in identifying, location, time, actors and actions. Determining a story's general dramatic structure creates presumptions about emotional states and the development of characters. This section will discuss the specifics of screenplay and play script formatting and dramatic narrative structure and review techniques for their automatic processing.

2.1.1 Text Layout Analysis

Professional screenplays and play scripts follow a unified, semi-regular writing style with strict formatting rules in order to serve practical purposes of the production process. They are written plans to translate a story into a film, play or animation visualisation and to direct scene flow, composition and content. The manifestation-oriented prose of scene descriptions focuses on what is audible and visible with sparse interpretations and explanations only when clarification is needed. Scene headings answer the question of ‘where’ and ‘when’, action descriptions make it easy to recognise ‘who done what’ and parenthesis refine the manner ‘how it was done’. The well structured text layout eases the machine parsing of scripts. Elements such as dialog, location, time, present actors, actions and sound cues can be visually recognised through capitalisation, indentation, parentheses and text alignment or tab settings. The writing style may vary slightly between different screenplays, but it is always coherent within one document.

Layout of Screenplays

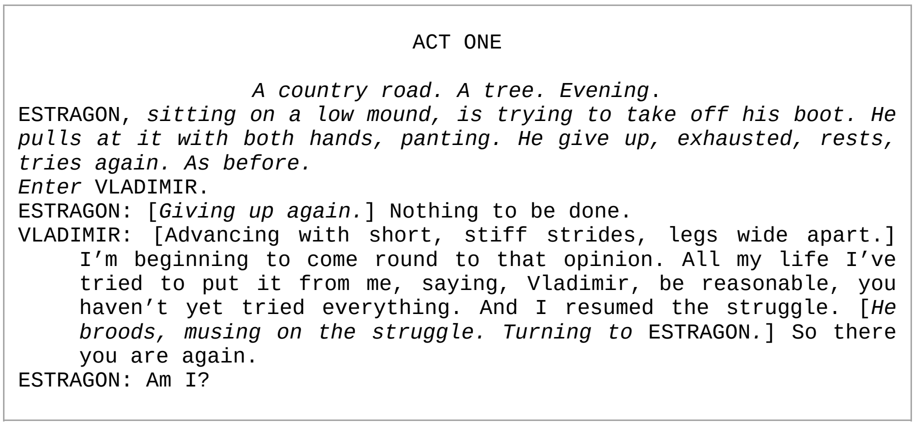
Final shooting scripts include numbered scenes and a codified notation to specify technical or dramatic elements such as scene transitions, changes in narrative perspective, sound effects, emphasis of dramatically relevant objects and characters speaking from outside a scene. One script page usually equals one minute of screen time. The beginning of a new scene is indicated by the ‘slug line’ which consists of three parts, the indoors or outdoors environment, location and time of the day, usually separated with a dash and in capitalised letters, e.g. ‘EXT – Hotel pool – NIGHT’. A more detailed description in prose of the scene and action may follow, though sentences may not always be complete and for instance verbs could be missing in set descriptions. Newly introduced actors, objects, sounds and camera instructions are capitalised. In the dialogue part, the actor’s name is capitalised and centred (2 tabs) and his/her spoken text is one tab indented. Further actions or manners during the dialogue are given in parenthesis. An example extract of a professional screenplay is given in Figure 1.

```
INT. M.I.T. HALLWAY -- NIGHT

Lambeau comes out of his office with Tom and locks the door.
As he turns to walk down the hallway, he stops. A faint TICKING
SOUND can be heard. He turns and walks down the hall.

Lambeau and Tom come around a corner. His P.O.V. reveals a
figure in silhouette blazing through the proof on the
chalkboard. There is a mop and a bucket beside him. As Lambeau
draws closer, reveal that the figure is Will, in his janitor's
uniform. There is a look of intense concentration in his eyes.

                   LAMBEAU
         Excuse me!

Will looks up, immediately starts to shuffle off.

                   WILL
         Oh, I'm sorry.

                   LAMBEAU
         What're you doing?

                   WILL
                   (walking away)
         I'm sorry.
```

Figure 1: Screenplay Extract from ‘Good Will Hunting (1997)’

Layout of Play Scripts

The layout of play scripts differs slightly and is more varied among different plays. An example of a play script excerpt is shown in Figure 2. Play scripts generally begin with a title page which might potentially state the genre and number of acts of the play. A cast list follows naming all actors to appear on stage. At the beginning of an act the act number is written centred and in capital letters potentially followed by a statement line similar to the slug line indicating location, props and time. Scene, actors and actions are described in more detail in the following paragraph. All descriptions are italised, actions and manners can be inserted into dialogue with squared brackets and all names are capitalised. An actor’s name with a colon indicates spoken dialogue. The keyword ‘curtain’, capitalised and centred, marks the end of an act.



ACT ONE

A country road. A tree. Evening.

ESTRAGON, *sitting on a low mound, is trying to take off his boot. He pulls at it with both hands, panting. He give up, exhausted, rests, tries again. As before.*

Enter VLADIMIR.

ESTRAGON: [*Giving up again.*] Nothing to be done.

VLADIMIR: [Advancing with short, stiff strides, legs wide apart.] I’m beginning to come round to that opinion. All my life I’ve tried to put it from me, saying, Vladimir, be reasonable, you haven’t yet tried everything. And I resumed the struggle. [*He broods, musing on the struggle. Turning to* ESTRAGON.] So there you are again.

ESTRAGON: Am I?

Figure 2: Play Script Extract from ‘Waiting for Godot’ by Samual Beckett

A number of research projects focus on screenplay processing and analysis. Preuß and Loviscach (2007) make use of the formal text layout of screenplays to automate the production of comic books. Through text analysis, dialogue elements are easily identified and placed into speech balloons, verbal directions and interrupted dialogues are mirrored in the corresponding speech balloon style and sound descriptions are presented through noise graphics. Choujaa and Dulay (2008) translate the high-level content of plain text screenplays into a machine readable XML representation. The screenplay layout is converted into an XML structure categorising items of content. A grammar for parsing screenplays is developed by Turetsky and Dimitrova (2004). Location, time, description of a scene, individual lines of dialogue and their speaker, the manner and action direction for the actors and transition between scenes, e.g. cut, fade, wipe, dissolve, are extracted. Thus screenplays are a rich source of high level semantic information.

2.1.2 Dramatic Structure

Generally, stories follow the dramatic structure of three acts. They are divided into *beginning*, *middle* and *end* (Choujaa and Dulay, 2008). In the beginning, the main character or protagonist is introduced as he/she is confronted with a new experience or a problem to solve. The middle typical accelerates in suspense and shot pace when actions/events densify, as the protagonist struggles to achieve his goal, leading to a climax or turning point where the character changes or finds a more successful approach. The decelerating end reveals the final conclusion, potentially with another moment of suspense, where the protagonist achieves his goal or tragically fails. Salway and Graham (2003) reveal the narrative structure of films by analysing the characters’ emotional course filtered from scene descriptions.

2.2 Computation of Emotions and Personality

All modalities of human interaction express personality and emotional states namely voice, word choice, gestures, body posture and facial expression. Therefore, many projects aiming to create life-like characters integrate affective modelling. Affective computing brings the concept of emotions, feelings or moods into the virtual world. Firstly, emotions have to be defined and structured for machine processing. Defining a universal set of emotions, their relation and causes have been widely discussed in literature. Emotion structures commonly applied in computing projects are presented below. A wide range of text processing techniques have been tested to computationally recognize and human-like interpret these affect structures in natural language text.

2.2.1 Emotion and Personality Models

Psychological theories for emotion, mood, personality and social status are translated into computable methods. These emotional models can be classified into three main emotional theories, categorical, dimensional and procedural (Masuch et al., 2006), which are mostly employed in artificial intelligence (AI). Categorical theories define sets of primary or fundamental emotions. Commonly referenced, Ekman's six basic emotions (Ekman and Rosenberg, 1997), happiness, surprise, sadness, disgust, fear and anger, are universally recognized across ages, cultures and species. Dimensional theories express different degrees of intensity of emotions and are often represented in a three-dimensional vector space. An example is the 'Pleasure, Arousal and Dominance' (PAD) mood space (Mehrabian, 1995). Procedural or Cognitive Appraisal theories account for the cognitive and subjective interpretation of emotion eliciting conditions by individuals. Based on appraisal models stimuli are evaluated with respect to goals, intentions, norms or standards, taste and attitudes of a character, as in the OCC-Model (Ortony et al., 1988). In order to model emotions as reactions to a character's context, human-like cognitive processing and evaluation of actions, objects and environment is required. The OCC emotion model provides an appraisal theory reflecting the subjective interpretation of such emotion eliciting conditions leading to a broad range of 22 emotions. The desirability, praiseworthiness and appeal of these conditions are evaluated according to the goals/intentions, norms/standards and taste/attitudes of an actor. McDonnell et al. (2008) have proven that the perception of Ekman's six basic and universally recognizable facial expressions also applies to emotional body language and virtual characters independent of their body representation. 3D figures reflect the richness of human bodily expression with varying intensity as perceived in Shaarani and Romano (2008).

Personality filters emotions and determines the emotions humans act upon. Thus personality is an integral part of emotion processing. Computation of personality is dominated by the psychological *OCEAN or Five-Factor-Model* (McCrae and John, 1992) which organizes personality traits into the five basic dimensions openness, conscientiousness, extraversion, agreeableness and neuroticism. Dependencies between emotions and the OCEAN-model for personality are often implemented with *Bayesian belief networks* or with *fuzzy logic* as in Su et al. (2007) (see Section 2.2.2).

Emotion and personality models influence human practical reasoning about plans to pursue an action to achieve a goal and can be integrated with the Belief-Desire-Intention (BDI) model (Bratman, 1987) The BDI model, readily implemented in robotics, is a software model for intelligent agents to reason about problems to solve and select plans from a plan library. Agents are equipped with the mental attitudes: beliefs, desires and intentions. Beliefs represent information about the world, the agent itself and other agents that the agent beliefs to be true.

They need not be true and may change over time, for instance inference rules may lead to new beliefs. Desires or goals are states that the agent would like to accomplish or see happen. Intensions are plans that the agent has committed to achieve. Plans are sequences of actions that an agent can perform and may contain other plans. The BDI approach allows the specification of large plans for complex and even uncertain domains where e.g. the agent's actions have probabilistic outcomes.

An example implementation can be found in the virtual human, Max (Kopp et al., 2008), who engages museum visitors in face-to-face small talk. Max listens while the users type their input, reasons about actions to take, has intentions and goal plans, reacts emotionally and gives verbal and non-verbal feedback. His emotion-model is based on PAD, with intensity values decaying over time and mutual interaction between emotions and moods. Default values set for the probability of Max taking a certain action determine his personal character.

2.2.2 Recognition of Affect from Natural Language Scripts

The word choice gives a good clue about a writer's or a story character's personality, social situation, emotional state and attitude. Sensing affective and attitudinal information from natural language text has intrigued researchers of various domains. Textual emotion extraction is significant for:

- improving human computer interaction through affective computing, e.g. in educational applications such as natural language based e-learning/tutoring systems (Tutoring Agent, *INES*: Heylen et al., 2005);
- opinion mining and market analysis, e.g. evaluating sentiment and opinions expressed in news, blogs, user feedback and reviews (*Sentiment Analysis*: Nasukawa and Yi, 2003);
- embodied conversational agents (ECAs), e.g. online chat agents (*EmpaphyBuddy*: Lui et al., 2003);
- storytelling systems and interactive computer games (Applications in Masuch et al., 2006)
- information retrieval and browsing (*Colour Bar*: Lui et al., 2003) and
- creating individual user profiles, e.g. to incorporate personalised advertisement.

Here we review six different approaches to textual affect sensing including keyword spotting, lexical affinity, machine learning methods, hand-crafted rule systems and fuzzy logic, statistical models, common knowledge based approaches and a cognitive inspired model, as well as combinations of these.

Keyword Spotting

Keyword spotting requires a fixed list of emotion-related keywords. The presence of individual keywords is 'spotted' in a given text. Dictionaries dedicated to affective words and phrases provide a source of fairly unambiguous emotional expressions. *Ortony's Affective Lexicon* (Ortony et al., 1987) consists of a collection of 500 affect-related words categorised into affective groups distinguishing affect, behaviour or cognition induced expressions and physical or bodily states. *WordNet-Affect* (Strapparava and Valitutti, 2004) extends the semantic synonym-sets of the WordNet lexicon (Fellbaum, 1998) with affective domain labels of pure emotion terms, personality trait terms and physical and cognitive state terms. With these lexicons affective keywords can be filtered from a scene script by simply looking up lexical entries. Even though being very precise, this approach has two main weaknesses. It is rather naïve, filtering emotions only superficially by relying on the presence of obvious affect words and disregarding underlying emotional meaning conveyed in non-emotional expressions. Thus the result is fairly incomplete. The second weakness is that keyword spotting does not account for negation, e.g. ‘not happy’ results in the detection of ‘happy’ and consequently a positive emotion. Salway and Graham (2003) apply keyword spotting of 679 emotion tokens to scene description scripts in order to extract visibly manifested emotions. Plotted against the time-code of the movie the distribution of the emotions reveals the film’s narrative structure.

Lexical Affinity

Lexical Affinity predicts emotions beyond obvious affect words extending keyword spotting to include arbitrary non-affective words. On a word-by-word basis Lexical Affinity can assign a probabilistic value to the co-occurrence of related word pairs or phrases (Terra and Clarke, 2004) or to the affinity/association of an arbitrary word with a particular emotion (Francisco et al., 2006). These probabilities are trained from linguistic corpora and therefore may be domain specific.

Statistical Models

Some underlying semantic context can be detected with statistical natural language processing. *Latent Semantic Analysis (LSA)* (Strapparava and Mihalcea, 2008) is an example of a statistical model. LSA evaluates semantic similarity of a given text and affective concepts. Through LSA the affective valence of emotion words, their synonyms or all WordNet (Fellbaum, 1998) synonym sets (synsets) labelled with the particular emotion can be represented as vectors and thus compared to generic words, word sets, sentences or full texts. *Term Frequency – Inverse Document Frequency (TF–IDF)* (Term Frequency, 2009) is a statistical measure to evaluate the importance of a term within a document, by calculating how many times a term appears in a text, where the weight of too frequent terms like ‘the’ is diminished and more weight is given to content containing words. Statistical methods score good results for coarse-grained evaluations in accuracy and precisely identifying relevant emotions (Strapparava and Mihalcea, 2008). But they require sufficiently large input texts to produce acceptable accuracy and are less successful on the sentence-level.

Hand-crafted Rules and Fuzzy Logic

Fuzzy logic can be applied to problems dealing with uncertainty. The degree of truth of a statement can range between 0 and 1, when reasoning about approximate rather than precise values. In linguistic applications fuzzy logic is used to express rules and facts through IF-THEN rules and Boolean logic (AND, OR, NOT). Su et al. (2007) decode the meaning of story scripts/scene descriptions and classify the affective state of the story characters. A psychological model of personality, the OCEAN model, Ekman’s six basic emotions (Ekman and Rosenberg, 1997) and a model of story character roles are combined through a hierarchical fuzzy rule-based system to control the body language of the characters. The fuzzy rule based system is applied to combine the various input and output parameters of emotions, personality and behaviour to produce complex emotional states. These emotional states are mapped to personality patterns providing guidelines for body language selection. They formulate 48 if-then statements that describe emotional and 32 if-then statements that describe personality conditions and relations. The fuzzy logic approach improves the classification and decision making process of personality and emotion for a character’s behaviour and the results are interpretable. It achieves the best results in terms of fine-grained evaluation due to the deep semantic analysis.

Machine-Learning

The intent to reproduce the dynamic and temporary nature of emotional behaviour has introduced various machine learning methods to emotion extraction and processing such as neural networks and Bayesian networks. Emotions and the affected behaviour are often spontaneously formed through situational motivation and reaction, rather than following predictable rules. Machine learning approaches take a character's memory into account including information about its identity, long-term goals and experiences with known objects and other characters.

Masuch et al.'s (2006) 'Emotion Engine' processes virtual agents' behaviour through a *neural network*. The neural network connects current emotional dimensions and personality traits to basic behaviour classes based on weighted links. Perceived stimuli from the agent's external environment or internal states can slowly change the agent's emotional status and emotions can decay over time.

Bayesian networks are often used to represent the mutual dependencies between emotions and personality as well as varying intensity values. Bayesian networks are also referred to as Bayes nets, Bayesian Belief Networks (BBNs), belief networks or causal probabilistic networks (CPNs). Ball and Breese (2001) choose a Bayesian network to represent the probabilistic and causal interactions between emotions, personality and their behavioural expressions of embodied conversational agents. Bayesian networks are particularly beneficial when dealing with uncertainty and predicting the likelihood of different outcomes. Thus they help to avoid artificial, repetitive behaviour and to account for human-like, variable and inconsistent reactions to the same event. As Bayesian networks consist of meaningful links and parameters, they represent the connections between causes and their effects, but initial probabilistic values need to be feed into the system. Ball and Breese constructed behaviour nodes for posture, linguistic, vocal and facial expressions, which are influenced by the internal, unobservable states of emotions and personality. The Bayesian model can be used for emotion recognition as well as for affective behaviour modelling.

Corpus-based machine learning algorithms, such as neural and Bayesian networks, learn the affective valence of keywords, word co-occurrences frequencies and punctuation from large corpuses of training data obtained by human hand annotation of affects in texts. Strapparava and Mihalcea (2008) use emotion tagged blog-posts to feed a *Naïve Bayes classifier* to automatically identify emotions in text. All blog entries tagged with one specific emotion were considered a positive example for the selected emotion, all other blogs provided negative examples. Alm et al. (2005) define a handcrafted feature set (or network of linear functions) covering linguistic, lexical and semantic rules with continuous value ranges. The feature set is trained with the *SNoW learning architecture* on a data set of 22 fairy tales. In both cases, a large training corpus is crucial for the precision of the textual emotion recognition. Thus machine learning methods are domain-specific depending on the training data and fail to incorporate contextual affective meaning.

Common Knowledge

Common sense knowledge about real word situations is essential to the emotional understanding of everyday events. WordNet (Fellbaum, 1998), Open Mind Common Sense (OMCS, 2009) and Cyc (Lenat, 1995), for instance, provide such knowledge in digital format. The WordNet database is optimised for lexical categorisation and word-similarity as it distinguishes discrete senses of words and links them by a small set of semantic relations such as synonyms or hierarchical ‘is-a’ relations. OMCS represents facts in a limited range of English sentence templates expressing different relations between concepts, e.g. ‘The effect of … is …’ or ‘… is used for …’. The OMCS project is a public collaboration, where web users enter common knowledge sentences through a website. Many practical, textual reasoning tasks are supported by ConceptNet (Liu and Singh, 2004), a commonsense knowledge base with an integrated natural language processing tool-kit. Context oriented inferences allow for a deeper and more meaning full understanding of text than keyword-based or statistical methods. A semantic network, encompassing spatial, physical, social, temporal and psychological aspects of everyday life, is automatically generated from the English sentences of the OMCS corpus. ConceptNet interrelated compound concepts, e.g. verb and noun phrases like ‘drink coffee’, through twenty semantic relations, e.g. EffectOf, DesireOf or CapableOf. Cyc utilises its own structured language, CycL, to formalise commonsense knowledge into a logical framework of largely handcrafted assertions. Cyc focuses on logical reasoning rather than practical reasoning. Liu et al. (2003) employ OMCS in order to retrieve the underlying affective semantics in natural language text. A subset of affective knowledge is extracted from the OMCS knowledge corpus through affective keyword spotting. Syntactic sentence analysis with consideration of negation and hand-coded rules for valence interpretation is performed to assign weighted emotions to sentences, words and phrases. Smoothing models aim to achieve affective coherence throughout a story by smoothing the transition of emotions between sentences. Therefore annotated emotions have a decay rate, neutral sentences are assigned an emotional value through hand-coded interpolation rules, a global mood is determined for each paragraph which influences the emotion of sentence and meta-emotions are blended from certain patterns of the basic emotions.

The common knowledge approach facilitates the integration of correlations between concepts and the corresponding words. Limitations of this method are that the relation of emotions to individual human beliefs, desires and intentions is not considered. This means that commonsense knowledge bases provide disputable default assumptions about typical cases, laying a foundation for more nuanced interpretation. Only ConceptNet with its NLP toolkit compensates for the lack of integration of semantic relationships and linguistic components which otherwise may lead to incorrect interpretations.

Cognitive Model

The OCC model has been implemented in several virtual agent systems to model agents’ emotional states and personality. Depending on an agent’s temperament, his/her emotional state changes in reaction to personal appraisal of events, actions and objects in their environment. Cognitive appraisal models are particularly appealing as a basis of computational models, because they can easily be integrated with the BDI framework (see Section 2.2.1) which is often used in agent systems. A computer implementation of the OCC model can be found in Elliott’s Affective Reasoner (Elliott, 1992). Elliott’s model also includes the representation of moods which can influence an individual’s appraisal of the world. The Affective Reasoner supports social interaction as it is capable of evaluating an event from multiple perspectives, from the agent’s own perspective and from the supposed view of other agents. However, the rules to appraise events in this model are domain specific.

Shaikh et al. (2009) implement a system that recognises the fine-grained emotion structure of the OCC model from text, whereas previously mentioned methods only distinguished either positive or negative sentiment or 6-8 basic emotions. Natural language processing tools, e.g. the Machinese syntax, semantic parser (Connexor, 2003), are applied to identify emotion-causing situations (events and actions expressed through verb-object pairs) and the cognitive state of an actor (mainly expressed through adjectives and adverbs). A database of verbs, adjectives and adverbs is created where each word is assigned a prior valence, prospective and praiseworthy value. These values are maintained by manually counting and calculating the positive and negative senses of each word in WordNet (Fellbaum, 1998). Real-world knowledge is used to determine variables influencing emotions. The semantic network of common sense knowledge, ConceptNet (Liu and Singh, 2004), automatically assigns prior valence to nouns and their matching concepts. Named entities, like peoples' names, are manually looked up in the web search engine, Opinmind (2008). Opinmind displays the relative number of positive and negative opinions expressed on a topic by users, which is used to assign a valence score to named entities. Based on the valence values of words, SenseNet (Shaikh et al., 2007) calculates the positive or negative sentiment carried by the sentences. Through combination and calculation of the valence values obtained by the linguistic resources mentioned above, specific values are assigned to cognitive variables according to the appraisal structure of the OCC model. Those cognitive variables are presumptions concerning the character itself or other agents, prospect, status, agent fondness, self-appraisal, object appealing, unexpectedness, deservingness, effort of action, expected deviation and event familiarity. They represent an understanding of an individual's beliefs about how events will affect him/her and predict how those beliefs lead to an emotional response. Rules for emotion types are then defined using the values of associated cognitive variables.

In conclusion, the linguistic interpretation of the OCC model for affect sensing from text does not only support explicit reasons for the detected emotions by extracting the eliciting conditions of emotions, but is also grounded on the semantic structure of sentences as well as common knowledge. Missing features of the OCC model would be a history function or memory of past emotions and experienced events, a personality designer for individual agents and the interaction and mutual influence of the emotional categories (Bartneck, 2002).

2.2.3 Emotion Processing

The above mentioned appraisal theory describes an agent's cognitive interpretation of their physical and social relationship with the environment. Considerable research has addressed this interpretation to select appropriate external actions. Human *coping* behaviour (Lazarus, 1966) involves either acting externally by attempting to alter the factors in the environment that are causing the emotion, or acting internally by changing one's beliefs about those factors, e.g. deciding a threatened goal is not so important, or altering the attention to those factors, e.g. not thinking about the threat. Therefore appraisal theories are incomplete and need to be extended to integrate the process of coping with emotions and internally managing emotions. Finding the most effective coping response depends on seriousness and likelihood of the threat, the relevance to the goal and a person's capability to deal with the threat. Personality influences the process of how an individual appraises and copes with events, as a person may not always choose the most effective coping response which may lead to emotional stress and adaptive or maladaptive behaviour. Marsella and Gratch (2003) implement a general, domain-independent model of coping as a framework for human-like autonomous agents. The coping framework is built on artificial intelligence reasoning techniques, such as plan-based causal representation, decision-theoretic planning and methods that model commitments to beliefs and intentions. The selection of a coping strategy involves four stages: identifying a coping situation, proposing alternative coping strategies, assessing the agent's coping potential and selecting the relevant strategy to be applied. Coping strategies comprise *planning* (forming the intention to take an action), *positive interpretation* (finding a positive meaning in a negative event), *acceptance* that a negative event is unavoidable, *denial/wishful thinking*, *mental disengagement* (a change of attention), *accept or shift of blame* and combinations of these strategies. A *reflexive agent* who is able to decide whether to express an emotion to hide its feelings in a given context is proposed by De Carolis et al. (2001). The behaviour planner applies regulation rules to triggered emotions considering the agent's personality and goals, the dialogue partner's characteristics and the situational and physical context. Synchronised presentation plans with XML-tags are created based on the available modalities, the cognitive ease to produce and process the signals, the expressiveness in communicating a specific meaning and the appropriateness to the social situation. Finally the XML-tags are converted into MPEG-4 facial parameters to render the animation. These two projects take emotion detection a step further introducing models to imitate human reactions to encountered emotions.

The various computational approaches, reviewed in this section, to identify emotions and psychological structures from text demonstrate a range of promising results, all with significant contributions and disadvantages within their area of focus.

2.3 Annotation of Emotions and Affective Behaviour

Tools for scripting the visual appearance of animated character scenarios compile and synchronize all information needed for 3D animation. Scripting characters' bodies and minds requires a declarative language catering for behaviour specifications including body parts and changes to each individual degree of freedom in a character's motion model, motion type, manner, duration and intensity as well as measures to express a character's personality by means of its movements. The social interaction intended through dialogue in play scripts is of more importance than the actual actions carried out or topics discussed. For this reason, Paggio and Music (2001) suggest a modality unifying representation to solve such hidden intentions and ambiguities. A considerable amount of notations have been invented for text-to-visual systems to translate the interpreted behaviour actions and affective expressions from the text input into a machine readable language. Most behaviour descriptions are Extensible Markup Language (XML)-based annotations added into the original script. These data structures provide a visualising language to express events affectively, expressing character features, roles and gestures and spatial information like path or place, and bringing various modalities together. The advantages of XML are its modularity and extensibility and the fact that it provides a natural way to represent information which spans intervals of text. This section will discuss four such XML-based annotations for expressive characters.

A mark-up language specifically intended for multimodal systems, the Extensible Multi-Modal Annotation (EMMA) (2003), provides an XML specification for containing the semantic interpretation of multimodal user input. EMMA is suitable for a variety of input recognisers, including natural language text, speech, graphical user interfaces among others. The mark-up language supplies a set of elements and attributes for annotating user inputs with their interpretations, e.g. attribute/value pairs describe the meaning of the input or describe a gesture. The interpretation is expected to be generated by an interpretation software component which automatically creates the EMMA notation to be used by another software component which integrates multiple modalities for visualisation for instance. EMMA is expected to be a standard data interchange format between the components of multimodal systems, but only focuses on single inputs, rather than multiple turns of dialogue.

Conversational, non-verbal behaviour, like facial expressions, speech pauses, gaze direction and gestures, is automatically modelled from dialogue text with the Behaviour Expression Animation Toolkit (BEAT) (Cassell et al., 2001). Utterances are broken down into an XML tree by the language tagging module, the generation model assigns behaviours to the speech parts and the scheduling module compiles a linear XML script including timing specifications. BEAT supports user defined filters, knowledge bases and tag set extensions. BEAT can be combined with systems that assign personality profiles, motion characteristics, scene constraints, or the animation style of a particular animator. In SPARK (Vilhjálmsson and Thórisson, 2008), for instance, BEAT is used to annotate chat messages to automate avatar behaviour in an online virtual environment.

The Multimodal Presentation Mark-up Language (MPML) is one scripting level higher as the above mentioned scripting languages. MPML supports the control of a character’s embodied behaviour as well as the synchronisation of the animation and synthetic speech of multiple characters. The XML-style mark-up language provides a tagging scheme for the control of predefined motions of a character, generating scripted descriptions of animated virtual characters which can be run in a web browser. MPML is employed for scripting the animation of emotion-based characters in SCREAM (SCRipting Emotion-based Agent Minds) (Prendinger et al., 2002), a web-based scripting tool for multiple characters which computes affective states based on the OCC-Model of appraisal and intensity of emotions, as well as social context. A character's mind contains a user-extensible set of rules and facts which evaluate events to generate emotionally and socially appropriate responses. An author defines the agent’s personality and social role through a graphical user interface. Breitfuss et al. (2007) take the SCREAM system a step further in the ‘Behaviour Generation System’ and automatically add non-verbal behaviour descriptions to dialogue scripts. Derived from linguistic and contextual analysis, gestures are generated and the annotated dialogue script is transformed into the extended MPML for 3D (MPML3D) to model facial and body animations and speech synthesis of 3D agents. The MPML3D player displays the animation of the embodied agents.

ALMA (Gebhard, 2005), 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 by mapping the five personality traits of the OCC-Model onto the Pleasure, Arousal and Dominance (PAD) mood space. Appraisal rules determine how a character appraises its environment, events and its own or other characters’ acts or emotional displays to filter out relevant emotions for display.

The presented annotation languages focus on different modalities, but due to XML’s modular nature, these scripting languages can easily be extended or combined to incorporate further modalities.

2.4 Modelling Affective Behaviour of 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 viewers' interpretation of the behaviour. The behaviour of affective agents also needs to adapt to changes of emotional states.

An architectural foundation for modelling intelligent and emotional agents has been proposed by Okada and implemented in the AESOPWORLD project (Okada et al., 1999) which integrates comprehension and generation of vision, motion and language. A computer model of the mind fulfils tasks such as mentally surveying objects, events and attributes, organizing high-level thought and communicating with others. The emotional states, mental and physical behaviour of the intelligent agent, the protagonist or fox of an Aesop fable, "The Fox and the Grapes", are expressed through graphical animation, voice and music generation. In doing so, AESOPWORLD focuses on understanding the meaning of each media. The agent lives in a natural environment, has desires or goals, makes plans to achieve them, takes actions to execute the plans, recognises this surroundings and gets emotional. These processes are called sequentially in a chain activation. The emotional character evaluates given data and makes subjective judgements based on eight primitive emotions leading to special actions or expressions in affective communication. Knowledge and emotion can be learned through the memory-learning domain. To describe body movements the human posture is roughly divided into 9 joint points and 7 connecting vectors.

The high-level control of affective characters in Su et al. (2007) is mapped from the output of the 'Personality & Emotion' engine to graphics and animation using Maya as the visualisation environment. Four main body areas are identified for human motion: head, trunk, upper and lower limbs. Possible postural values are supplied to a *Dependency Graph* to manipulate the shape and geometry of the model, e.g. through the stretch and squash technique. The joint values of the character's skeleton are updated according to the '12 principles of typical animation techniques for believable characters' (Thomas and Johnson, 1981; Disney Animation, 2008) based on physical characteristics of sex, space, timing, velocity, position, height, weight and portion of the body. Sequences can be layered, blended and mixed through non-linear animation. Greta (Pelachaud, 2005 and De Rosis et al., 2003) is modelled as an expressive multimodal Embodied Conversational Agent (ECA). The Affective Presentation Mark-up Language (APML) defines her facial expressions, hand and arm gestures for different communicational functions 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 (Gunes and Piccardi, 2006). The captured animation data can be mapped to 3D models, which is useful for instructing characters precisely on how to perform desired actions. Even though the above approaches enable automatic emotional 3D animation, a challenge remains in assigning appropriate animation data to seamless nuances of emotion intensities and reasoning about an individual's expression of an emotion.

2.5 Multimodal Visualisation of 3D Scenes

This section discusses all visual 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. Examples of complete text-to-visual systems, SONAS, WordsEye, CONFUCIUS and ScriptViz, and the scene directing system, CAMEO, are presented.

2.5.1 SONAS – Environment Visualisation

An early system to interpret language and automatically construct a three-dimensional virtual world display is the Spoken Image (SI) System (Ó Nualláin and Smith, 1994), or its successor SONAS (Kelleher et al., 2001). The Spoken Image system accepts verbal or written natural language scene descriptions of a town scenario, given in the form of dialogue between the system and a human user. Spoken Image builds the described scene sentence by sentence as the user speaks, placing new objects with default values until further specified by the user. As human language leaves many details unspecified, a great deal of reasoning and common sense grounding must be performed by the computer in order to maintain consistent visual scenes. SONAS integrates a combination of several input modalities, like spoken language and gesture, to allow a user to manipulate the objects in the 3D environment. To achieve this task, sentences are parsed and broken down into visual objects, action and spatial relations, then the virtual scene is searched for the specified elements and once they have been identified, action classes are instantiated to calculate the path from initial to final position of the figure when executing the action. The main effort of the SONAS project has been to define common semantics between language and vision and to develop a meaning representation that is common to both language and vision.

2.5.2 WordsEye – Scene Composition

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 such as background noises or music. Visualisation of scenes from text input is realised in WordsEye (Coyne and Sproat, 2001), which creates static 3D images from specific descriptive texts. 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. The objects are connected with their linguistic counterparts in an extended WordNet version containing links to their corresponding nouns.

2.5.3 CONFUCIUS – Scene Animation

CONFUCIUS (Ma, 2006) is an intelligent, multimedia storytelling system which produces multimodal 3D animations with audio from single natural language sentences. The 3D animation generation involves virtual humans performing actions, dialogues being synthesised and basic cinematic principles determining the camera placement. Linguistic and visual semantics are connected through the proposed Lexical Visual Semantics Representation (LVSR) which focuses in particular on visual semantics of verbs and suitability for action execution/animation to generate virtual humans' movements for verb classes. The LVSR structure lists each action mentioned in a sentence, the agent who performs the action, the theme, time and location of the action. Investigating the meaning of verbs in language visualisation leads to the distinction of two *visual roles*, human and object, because they require different processes in animation generation. Since visual modalities require more specific information than expressed in language (syntax and semantics), the notion of *visual valency* is introduced, which states the number of visual roles involved in the animation of an action. For example, the sentence ‘Jane cut the cloth’ syntactically presents a valency of two, the subject ‘Jane’ and the object ‘cloth’, but for visualisation a valency of three is necessary including another object, ‘scissors’, the instrument used for cutting the cloth.

Natural language understanding and generation is realised with existing tools for syntax parsing and text-to-speech synthesis. Name recognition, negation, idioms and indirect speech are solved in the pre-processing module. The semantic analyser performs semantic inferences, word senses disambiguation, anaphora resolution of third person pronouns and temporal reasoning expressing different tenses in sequential order. Lexicons are availed of for word sense disambiguation. Ambiguous verbs are analysed with hypernym relations (more general words) and word frequencies of WordNet (Fellbaum, 1998) and thematic roles of the Lexical Conceptual Structure database (LCS, 2000).

The visual knowledge base encompasses H-Anim (2001) 3D models of characters defining geometry and joint hierarchy, physical, spatial and functional information, models of props and places and event models describing animations of actions. The H-Anim modelling language is a subset of the Virtual Reality Modeling Language (VRML, 1995). An animation blending approach combines pre-created and dynamically generated animation sequences in smooth transitions. The media allocator module decides which content is represented though with modality assigning content to either the animation engine, text-to-speech engine or the narrator. The animation and speech output is synchronized through a VRML file. VRML relieves the computation of path-finding, enables the encapsulation of sound effects, like ambient background noise or music, within object models and provides viewpoints to adjust position and orientation of the camera view to any desired perspective in storytelling and to achieve movie-editing effects like ‘cuts’ or ‘pan shots’. Output Examples of CONFUCIUS are given in Figure 3 to 5.

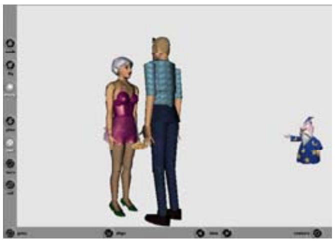


Figure 3:
CONFUCIUS Generated Scene - Narrator Merlin Speaks Alongside a Story

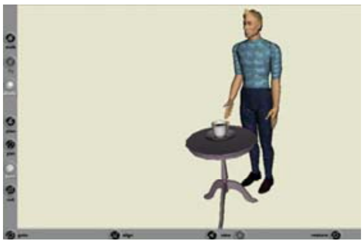


Figure 4:
CONFUCIUS Output Animation of “John put a cup on the table.”

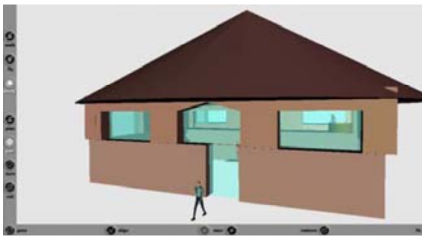


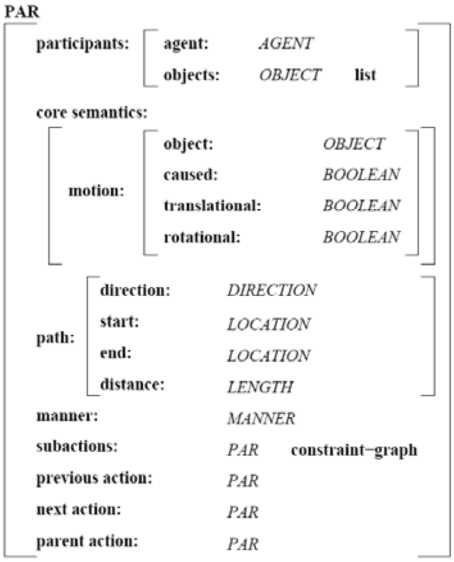
Figure 5:
CONFUCIUS Output Animation of “John left the gym.”

2.5.4 ScriptViz – Screenplay Visualisation

ScriptViz (Liu and Leung, 2006) is a software implementation which allows users to visualise screenplays automatically via realistic, animated 3D graphics. Though marginally based on screenplays, but rather short stories of a few well-formed, unambiguous English sentences, ScriptViz performs natural language processing, constructs plans of action and renders scenes. The user interface consists of three sections, a textbox for the input of a natural language story sentence, the full script display and the virtual stage for viewing the animation output in real-time (Figure 7). It appears that a scene specification naming the set environment, actors, and objects to appear in the scene and their type and position is required before the story writing process begins. The text-understanding module processes one sentence at a time. A dependency parser extracts verbs and their corresponding subject and object from the sentence structure. The meaning of the verbs is identified and actions or emotion changes for the relevant agents are triggered. Emotions in the ScriptViz system only reflect explicitly stated emotion words.

The action planning process involves four phases. The high-level planning module retrieves a plan outline from its library for the action described by the verb. Each plan describes one action which may consist of a number of subtasks or action primitives. In the second phase, the subject and object of the sentence are matched with the agents and objects present in the current scene. In the third phase, the feasibility of the planned task is tested checking whether the requested action can be performed by the specified agents and if necessary supporting actions are added in, such as getting an agent into a suitable position. In the final planning phase, the task plan and its required steps are refined with concrete parameters of the current state of the virtual world and a Parameterised Action Representation (PAR) (Badler et al., 2000) is generated (see Figure 6). Thus the PAR provides detailed instructions for animation. While an agent is carrying out his planned task, the action primitives can only be performed sequentially and the state of the virtual world is frozen to elude having to monitor the environment and other agents and potentially having to adapt and update the plan while being processed. The scene generator passes the task plans on to the relevant agents for animation, coordinates different agents in the scene and updates resulting state changes of the virtual world caused by the execution of the task. The scene generator executes actions through a hierarchical structure of 'bubbles', which are objects composed of primitive animation elements like graphics and motion-capture data files in a database. Bubbles contain methods for blending primitives in smooth transitions.

The scene generator provides two camera views, a close-up view for each agent and an independent camera view. The camera can be moved manually to any position and orientation. ScriptViz demonstrates a simplified screenplay animation software providing immediate visual feedback during the creative writing process.



PAR

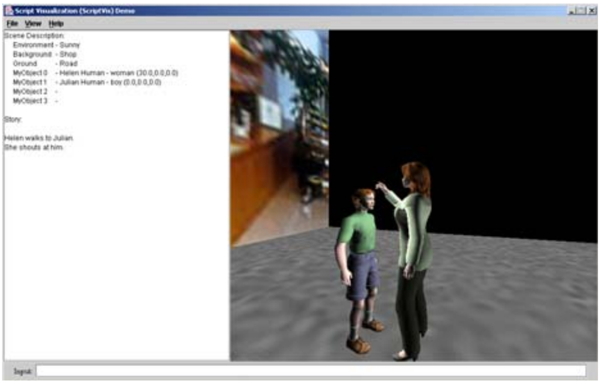
participants: agent: *AGENT*
objects: *OBJECT* list

core semantics:
motion: object: *OBJECT*
caused: *BOOLEAN*
translational: *BOOLEAN*
rotational: *BOOLEAN*

path: direction: *DIRECTION*
start: *LOCATION*
end: *LOCATION*
distance: *LENGTH*

manner: *MANNER*
subactions: *PAR* constraint−graph
previous action: *PAR*
next action: *PAR*
parent action: *PAR*

Figure 6: Syntactic Representation of PAR



Scene Description
Environment - Sunny
Background - Shop
Ground - Road
MyObject 0 - Helen Human - woman (30.0,0.0,0.0)
MyObject 1 - Julian Human - boy (0.0,0.0,0.0)
MyObject 2 -
MyObject 3 -

Story:

Helen walks to Julian.
She shouts at him.

Figure 7: ScriptViz User Interface

2.5.5 Automated and Intelligent Cinematography

Furthermore, another modality, cinematography, can assist in conveying themes and moods in animations. Through the application of cinematographic rules defining appropriate placement and movement of camera, lighting, colour schemes and the pacing of shots, communicative goals can be expressed. The psychological effects of film techniques are easily understood by viewers and affect their emotional disposition. Smith and Bates (1989) point out how adept film making is at communicating within a context of significance and which strong impact film editing has. The same scene, presenting the same physical information, edited in different ways, will evoke completely different moods and deeper ideas. Therefore they include film making techniques in the Oz project. The Virtual Theatre Interface project (Virtual Theatre, 2004) offers a web-based user interface to experiment with actors' positions on stage, lighting effects and audience view points.

Kennedy and Mercer (2002) developed an application which automatically applies film techniques to existing animations. Reasoning about the plot, theme, character actions, motivations, emotions and narrative goals described by the human animator, it creates a communicative plan, automatically maps appropriate cinematographic effects from a knowledge base and renders the final animation. A cinematic knowledge-base for camera behaviour in different cinematic genres is developed by Friedman and Feldman (2004). Rules and principles of cinematic expression were collected from text-books and domain experts, converted into declarative rules and captured mathematically. DirectorNotation (Christodoulou et al., 2008) is an alternative, under-development, open-source system capturing cinematography knowledge. DirectorNotation generates ontologies to formalise and reason over the film-making domain. Film directors can formally record their intentions using the logical symbolic notation system which is fully computer-processable and can be integrated in various applications for automatic animation synthesis. Ramírez (2005) describes a concept for the integration of such cinematic knowledge bases into the automatic creation of scenes with a reasoning engine.

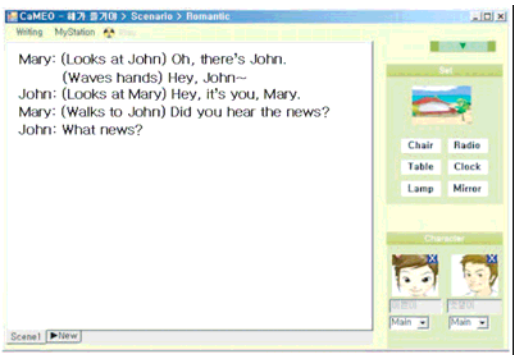
De Melo and Paiva (2006) introduce a high-level synchronized language, the Expression Mark-up Language (EML), which integrates environmental expressions like cinematography, illumination and music as a new modality into the emotion synthesis of virtual humans.

2.5.6 The Effect of Genre on Animation Style

The time, milieu or environment where a story takes place (setting), the issues or the concept the story revolves around (theme) and the emotional tone (mood) of fictional films, plays or literature classify different genres with distinguishable presentation styles. Genre categorisation is not clearly defined and two or more genres may be combined in one production. Nevertheless commonly applied genres are action, adventure, comedy, crime, documentary, drama, fantasy, horror, musical, romance, science-fiction, thriller, war and western (Film Genre, 2009). Genre is reflected in the detail of a production, efficiency, exaggeration and fluency of action movements, pace (shot length), lighting, colour and camera movement. These parameters are responsible for viewer perception, inferences and expectations and thus for an appropriate affective viewer impression. The audience is 'trained' to interpret low-level symbols like lighting, background music and shot length. For example, dim lighting is associated with fear and silence with a feeling of expectation. Therefore it is relevant to determine the genre before modelling a scene and to apply the correct animation style to the scene.

Vish (2004) examines the relationship between genre recognition and animation characteristics such as body movement parameters (efficiency, fluency and detail), body portion and film velocity. Vish compares thematically similar movie scenes of four genres (comic, drama, action and non-fiction) and models the same scene with varying parameters. Rasheed et al. (2005) investigate the cinematic principles of average shot length, colour variance, motion content (amount of activity) and lighting keys in different genres. Dramas are slower paced with longer dialogues, whereas action movies have rapidly changing, shorter average shot length. Comedies tend to be presented in a large spectrum of bright colours, whereas horror films adopt mostly darker hues. High-key lighting with bright scene illumination and little contrast is mostly found in comedies and actions films. Low-key lighting which is predominantly dark with high contract between lighted areas and shade appears more dramatic and is mostly found in horror films. Dramatic and romantic movies have less visual activity compared to action movies. The automatic 3D animation production system, CAMEO, implements distinguishable direction styles for user selected genres and automatically applies them to the scene presentation.

2.5.7 CAMEO



Mary: (Looks at John) Oh, there's John.
(Waves hands) Hey, John~
John: (Looks at Mary) Hey, it's you, Mary.
Mary: (Walks to John) Did you hear the news?
John: What news?

Figure 8: CAMEO User Interface

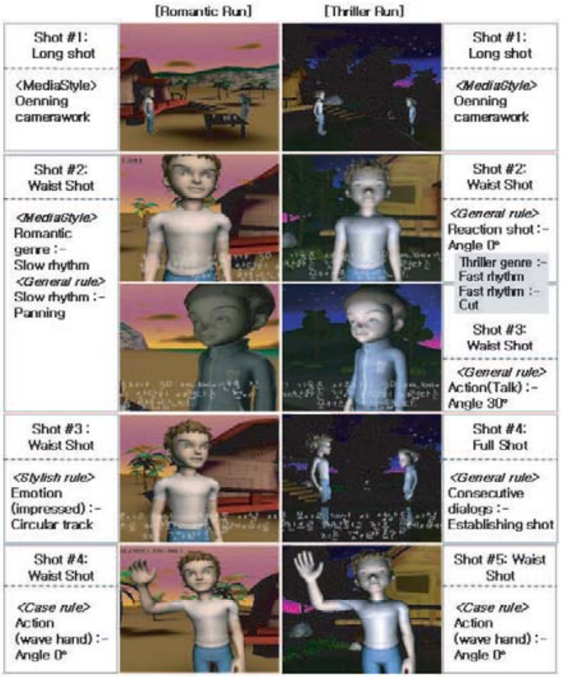

[Romantic Run]
[Thriller Run]
Shot #1: Long shot
<MediaStyle> Oenning camerawork
Shot #2: Waist Shot
<MediaStyle> Romantic genre :- Slow rhythm
<General rule> Slow rhythm :- Panning
Shot #3 : Waist Shot
<Stylish rule> Emotion (impressed) :- Circular track
Shot #4: Waist Shot
<Case rule> Action (wave hand) :- Angle 0°
Shot #1: Long shot
<MediaStyle> Oenning camerawork
Shot #2: Waist Shot
<General rule> Reaction shot :- Angle 0°
Thriller genre :- Fast rhythm
Fast rhythm :- Cut
Shot #3: Waist Shot
<General rule> Action(Talk) :- Angle 30°
Shot #4: Full Shot
<General rule> Consecutive dialogs :- Establishing shot
Shot #5: Waist Shot
<Case rule> Action (wave hand) :- Angle 0°

Figure 9: Romantic and Thriller Style Output of CAMEO

CAMEO (Shim and Kang, 2008) is an automatic animation production system which creates 3D scenes based on cinematic direction knowledge for characters, camera, light, film editing and sound. CAMEO requires three inputs, besides writing the screenplay, the story in the form of dialogues, the user selects the setting, the characters and the media style defining the theme or genre of the animation (Figure 8). Based on these initial settings three XML representations are constructed for the story, the characters, set and shooting characteristics and the production style. The XML documents are merged into one 'SceneScipt' by applying a set of film direction rules. Depending on the atmosphere and genre style, shots are determined for the dialogue elements and directions for camera, light and sound are inserted according to text, emotion and action values. Cinematic effects are defined in the direction knowledge base which includes clichés in direction and film practices, general filming rules and user invented rules. The most appropriate directions are decided with a statistics module. Thus, scenes can be automatically rendered with different presentation styles (Figure 9).

These projects demonstrate that the automatic application of cinematic rules supports the expression of mood and genre of the narrative and assists users in considering the composition of a scene, camerawork and illumination efficiently.

2.6 Virtual Theatre

Interactive, virtual theatre projects are presented, since live performances such as theatre serve as concise test scenarios for interaction between humans and robots, or virtual agents (Breazeal et al., 2003). The storyline defines the scenario, the script provides constrained dialogues and interaction, the stage or set constrains the environment and 'actors' have to act and react in a compelling and convincing manner, which requires sophisticated perceptual, behavioural and expressive capabilities.

In the Improv system (Perlin and Goldberg, 1996) an author can create virtual actors to interact with human actors in a virtual theatre. Through an English-style scripting language the author defines the virtual character's behaviour set and personality which act as rules governing how the actor behaves, changes, makes decisions and communicates. Intelligent multimodal virtual theatre productions in real space provide human user interaction through sensitive vests or head-mounts to recognize user movements and speech (Cavazza et al., 2007; RIVME, 2004). Thus the actor is integrated into the virtual performance and directly influences the plot and the virtual character's reactions as in Cavazza's Mme Bovary project. The mixed reality project 'Casino Virtuell' (Gebhard and Schröder, 2008) realises an artificial intelligence Poker game with two virtual players competing against human players. The virtual characters whose comments and actions are influenced by their personality and emotional state give a believable impression of expressive behaviour. Emotions are simulated with the above mentioned ALMA model and a novel speech synthesis system controls the quality of emotional expression in the characters' synthesised voice. The game plot is generated through the 'authoring toolkit', SceneMaker (Gebhard et al., 2003). SceneMaker treats content in separate scenes (pieces of dialogue) which can be pre-scripted or author written. An author can define the logical scene flow (transitions between the scenes) of the story, in finite state machine graphs.

2.7 Automatic Sound Selection and Affective Speech Synthesis

Audio elements, such as synthesised speech, music and sound play an important role in conveying emotions in movies or plays and therefore need to be as carefully selected as the visual elements. A wide range of research is devoted to emotional speech synthesis, automatic music classification, intelligent music retrieval including audio recommendation systems, song identification and similarity matching, visual and acoustic matching and experimental sound generation.

While advances in different aspects of speech synthesis contribute to a natural impression of voice, emotional expressivity still needs further development. Several emotional speech synthesis systems are reviewed in (Schröder, 2001). Most studies model only three to nine discrete, extreme emotional states, rather than reflecting the variety of subtle, sometimes language-specific, human emotions. Prosody rules have been outlined to determine, how a given emotion is expressed through voice, but techniques are yet missing to assess the appropriateness of the acoustic output for a given communication context.

Cano et al. (2005) developed a content-based music recommendation system which categorizes songs by rhythm, tonality, melody, main chords and variations of loudness over time. The system allows to make a distinction between sad or happy moods conveyed in a given piece and to estimate a probable genre membership. A system which recommends music based on emotion is proposed by Kuo et al. (2005). Associations between emotions and music features in movies are discovered by extracting chords, rhythm and tempo. Inspiration for the integration of expressive audio according to relevant visual cues can be found in experimental projects for sound effects (Physically Informed Audio Synthesis, 2008) and music (Rebelo et al., 2005).

2.8 Multimodal Web Based and Mobile Interfaces

In order to make script visualisation and scene editing available on mobile devices, technical restraints have to be agitated. Mobile and handheld devices are limited in memory size, processing power, screen size and graphics acceleration. Even though recent technological advances make 3D development and display on mobile devices possible, it is still limited in its performance and requires the adaptation of existing desktop 3D development techniques, software and file structures. This raises an issue of version control between desktop and mobile application file formats. System architectures have to distribute the user interface between servers and mobile devices (Turunen and Hakulinen, 2007). The integration of and adaptation to the mobile context poses an interesting research question. Examples of multimodal systems and 3D rendering methods for mobile devices are given now.

2.8.1 Multimodal Mobile Applications

SmartKom Mobile (Wahlster, 2006) brings the multimodal system, SmartKom, onto mobile devices such as Personal Digital Assistants (PDAs). Supported modalities include language, gesture, facial expression and emotions carried through speech emphasis. The user interacts with a virtual character, Smartakus, through dialogue. SmartKom supports tasks of 50 different domains for scenarios in the car or for pedestrians with an extendable unified knowledge base. It is possible to place the system architecture and rendering on the mobile device itself or to distribute these from a server via wireless broadband networks. Turunen et al. (2005) introduces a mobile system architecture for multimodal dialogues. The server handles the coordination of the dialogue and generates dialogue task descriptions in VoiceXML (Voice Extensible Markup Language) (2004), which are executed on the mobile device selecting modalities from available resources.

2.8.2 3D Rendering on Mobile Devices

As more powerful mobile devices are available, software architectures carrying out the entire 3D rendering process on the mobile deliver promising results. Remote 3D rendering on powerful servers and sending results via wireless network or including 3D rendering hardware on the small mobile devices becomes unnecessary.

For the purpose of integrating 3D computer graphics with Java, the *Mobile 3D Graphics API* (M3G) (M3G, 2004 and JSR 184, 2005) has been designed to meet the specific needs of mobile devices. This specification allows developers to write Java programs that produce 3D graphics on mobile phones and PDAs. M3G is separate and incompatible to the desktop version Java 3D, but it specifies a file format for 3D models and animations which allows the creation of content on PCs and can be loaded on mobile devices. The main advantage of M3G is that Java is widely supported on mobile devices, but the M3G API still runs quite slow on most mobile phones. OpenGL for Embedded Systems, OpenGL ES (2005), is the emerging standard for 3D rendering on mobile devices. OpenGL ES consists of a well-defined subset of the desktop 2D and 3D API of OpenGL and thus enables full programmable 3D graphics which are easily migrated across platforms. Nadalutti et al. (2006) built a rendering engine using the OpenGL ES API that implements shading, lighting and navigation algorithms. This rendering engine is integrated into their mobile 3D player, MobiX3D, which fully supports the H-Anim (2001) standard. The mobile 3D player is able to display sign language sentences performed by an H-Anim model on mobile devices. To improve the performance models are rendered with a low polygon count as the small display of handheld devices doesn't require any greater detail.

Virtual Reality Modelling Language, VRML (1995), is a scene description language to represent interactive and animated 3D models in web browsers through freely available plug-ins. In VRML 3D geometric objects, their shape, the appearance of the shape and their behaviour or transformations, that can be applied, are described in nodes arranged in scene graphs. A mobile VRML browser and development kit, Pocket Cortona (2009), has been developed for wireless devices supporting most of the VRML specification. VRML models developed on desktop applications can be viewed without modification on the mobile device and developers can write 3D applications on wireless platforms using the Cortona SDK. A useful feature of VRML is its efficiently small file size. Another format used to realise mobile 3D animation is MPEG-4 coding for audio-visual scenes which is particularly suitable for mobile applications because of its low bit rates. DeCarolis et al. (2001) convert XML-tags to MPEG-4 parameters.

An interface for editing and manipulating virtual characters on mobile devices is tested on the 'mobile animator' interface (Gutiérrez et al., 2004) which adapts the intuitive graphical interaction of professional, commercial desktop 3D packages, like Maya (2009) or 3ds Max (2009), using virtual manipulators. 3D objects and camera are directly controllable on the handheld display by selecting and dragging body parts with a stylus. 2D controls, like sliders, buttons and menus are avoided. To maintain the level of detail of the 3D characters for display on bigger screens and more powerful computers, handheld users are provided with a simplified representation. Bounding boxes roughly section the 3D model providing enough information about the actual dimensions of the character's limbs. The user can zoom into the simplified sections to access the full set of joints. Full H-Anim models contain close to 80 joints. In the mobile animator interface only one virtual character can be modified at a time. A list of characters gives users the option to select another 3D character to be edited.

These examples show that software methodologies developed for online access can be adapted for mobile devices and enable mulitmodal and 3D content on handheld devices with some performance drawbacks. The wide range of approaches to modelling emotions, moods and personality aspects in virtual humans and scene environments along with first attempts to bring multimodal agents onto mobile devices provide a sound basis for the *SceneMaker* prototype.

3 Project Proposal

This research will investigate Natural Language Processing, knowledge and reasoning methods with regard to extracting theme and mood from film/play scripts. The main aim is to visualise the detected emotions in virtual animations. Affective scene visualisation is discussed in section 3.1. Tools and techniques will be tested with a software prototype, *SceneMaker,* which will augment 3D scenes with emotional influences on body language and environmental expression. *SceneMaker's* architecture is presented in section 3.2. Section 3.3 covers prospective tools for affect, text and language processing, automatic visualisation, sound and speech generation and the user interface to be combined in *SceneMaker* to form a complete text-to-animation system. A multimodal scene will be composed of affective actors and multimedia like music, sound, illumination, timing and camera work to support a given mood and theme. A front-end user interface, featuring a mobile 3D editor, will be created for directors or animators with a focus on high usability and intuitive operation on computers and mobile devices to allow fine-tuning of the automatically created animations. Created animations can be saved and shared with co-workers. In section 3.4 methodologies for testing the functionality and effectiveness of the prototype are discussed.

3.1 Affective Scene Visualisation with SceneMaker

Existing AI methodologies assist users in different aspects of creating virtual scenes. *SceneMaker* aims to enhance these approaches with more realistic and believable visualisations through affective computing. The more complex OCC emotion model will build the affective basis of *SceneMaker* to provide a finer grained emotion distinction and realistic cognitive processing. Emotions will be computationally extracted from natural language film or play scripts. In doing so, *SceneMaker* will go beyond filtering explicitly stated affective words and will take into account the context, events before and after the occurrence of an emotion, common knowledge and the meaning of actions. *SceneMaker* will define the genre of a story through text and language analysis to allow theme specific presentation of scenes.

3.2 Architecture of SceneMaker

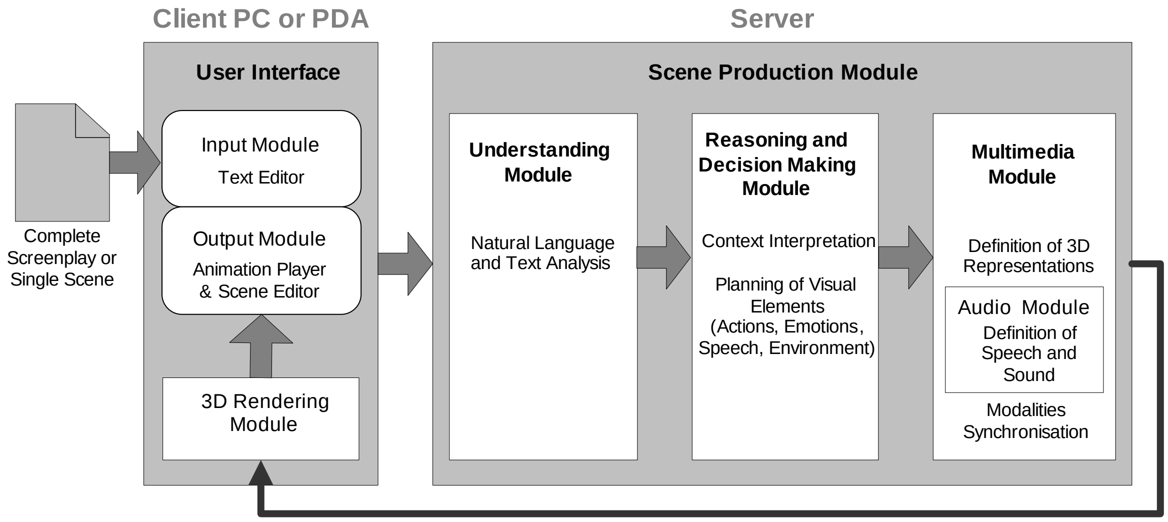

Client PC or PDA
Server
User Interface
Scene Production Module
Input Module
Text Editor
Complete Screenplay or Single Scene
Output Module
Animation Player & Scene Editor
3D Rendering Module
Understanding Module
Natural Language and Text Analysis
Reasoning and Decision Making Module
Context Interpretation
Planning of Visual Elements (Actions, Emotions, Speech, Environment)
Multimedia Module
Definition of 3D Representations
Audio Module
Definition of Speech and Sound
Modalities Synchronisation

Figure 10: *SceneMaker* Architecture

*SceneMaker*s's architecture is given in Figure 10. The main component is the scene production module including three modules for understanding, reasoning and multimodal visualisation situated on a server. The understanding module performs natural language processing and text structure analysis of the input text. The reasoning module interprets the content based on common, affective and cinematic knowledge bases, updates emotional states, creates plans for actions/behaviours and defines their manner and the representation of the set environment. The multimedia module maps these plans to 3D animation data, selects appropriate models from the graphics database, defines their body motion transitions, instructs speech synthesis and selects sound and music files from the audio database, instructs camerawork and lighting. The multimedia 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 for animation playback renders the 3D scene according to the manuscript and allows manual scene editing of all 3D objects to fine-tune the automatically created animations.

3.3 Software Analysis and Prospective Tools

A detailed review of existing affective computing, language processing applications and multimodal presentation will reveal the most suitable components for *SceneMaker*. Different theories on emotion processing and expression will be evaluated for viability. Technologies for *SceneMaker's* online and mobile user interface will be tested.

3.3.1 Methodologies

For the development of *SceneMaker* a Constructionist Design Methodology (CDM) (Thórisson, 2007) or more specifically a Constructionist AI Methodology (CAIM) (Thórisson et al., 2004) will be employed. This methodology suggests a modular approach. Existing software tools fulfil sub-tasks and will be modified, combined and extended to construct the modules of *SceneMaker*. The Psyclone software platform (Thórisson et al., 2004) may be utilised as it simplifies the design of complex systems and their connections between various input and output modules like e.g. vision, body tracking and graphics. By incorporating the OpenAIR specification (Thórisson, 2007), for information exchange and network routing, Psyclone provides a unified messaging format to communicate between different modules. *SceneMaker* will include interacting modules for perception, i.e. text interpretation, for decision, i.e. emotion and action classification, and for multimedia presentation, i.e. 3D character animation, speech synthesis, music, camera and lighting.

3.3.2 Text and Language Processing

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.

For the automatic interpretation of the input scripts, *SceneMaker* will build upon the natural language processing module of CONFUCIUS (Ma, 2006). The syntactic knowledge base of CONFUCIUS parses the input text and identifies grammatical word types, e.g. noun, verb, adjective or other, with the Connexor Part-of-Speech Tagger (Connexor, 2003) and determines their relation in a sentence, e.g. subject, verb and object, with Functional

Dependency Grammars (Tesniere, 1959). Despite variability, ambiguity and imprecision of natural language *syntactic analysis* tools achieve mostly reliable results. *Semantic interpretation* and actual understanding of the meaning of a text is more difficult, because it depends largely on common sense knowledge. Common sense knowledge and mental images need to be structured, related through logical rules and entered into databases before computational text interpretation is possible. A commonly used tool for determining semantic relations between words is WordNet (Miller, 1995, Fellbaum, 1998), an extended dictionary specifying word relations such as similarity, part-of relations, hierarchy or manner. The collected knowledge assists in specifying the story context to consequently evaluate current events and assign appropriate emotions. The semantic knowledge lexicons of CONFUCIUS including WordNet and the LCS database (LCS, 2000) coupled with a visual semantic ontology and temporal language relations solve semantic inferences and word sense disambiguates. The knowledge bases in CONFUCIUS will be extended by ConceptNet (Liu and Singh, 2004) and Open Mind Common Sense (OMCS, 2009) for practical commonsense reasoning about context and Opinmind (2008) to analyse proper pronouns such as peoples' names. An emotional knowledge base will be added, e.g. building on WordNet-Affect (Strapparava and Valitutti, 2004).

SceneMaker will add two new tools to the pre-processing of the input text, a layout analysis tool and a genre recognition tool. The *Layout-Analysis tool* decomposes the text layout structure of screenplays or play scripts into its scene elements in relation to the parsing grammar by (Turetsky and Dimitrova, 2004) or the XML structure by (Choujaa and Dulay, 2008). The Layout-Analysis module will separate scene descriptions from dialogue and thus facilitate access to semantic information. The *Genre-Recognition tool* will apply and combine two language processing methods to determine genre from natural language screenplays or play scripts. First, keyword co-occurrences and term frequencies are examined, because words, terms and concepts that co-occur are likely to be meaningfully related. Hence, highly recurrent keywords, word pairs or phrases are indicators for different themes and genres. When identifying *Keyword Co-Occurrences* (2005) some factors manipulate the result and need to be considered: the word type, e.g. noun, verb or adjective, synonyms which rarely co-occur, but represent the same context, the scope how general or specific a term is in a given context, the order in which terms appear and multiple meanings of words. *Term Frequency – Inverse Document Frequency* (TF-IDF) (Term Frequency, 2009) is a statistical algorithm calculating the number of times a content-relevant term occurs in a document, by diminishing the weight of words that appear very frequently, but have no contextual meaning, like 'the', 'a' or 'and'. The TF-IDF weighting scheme can be used in the *vector space model*, an algebraic model for representing text documents as vectors, together with *cosine similarity,* a measure of similarity between two vectors by finding the cosine of the angle between them, to determine the similarity between screenplays of the same genre. *Latent semantic analysis* (LSA) (see section 2.2.2, Statistical Models) analyses relationships between documents and the terms they contain by constructing sets of concepts related to the document and term. Second, the length of dialogues, sentences and scenes/shots can be determined respectively by a simple word count, which allows assumptions on the genre type. A *genre knowledge base* will account for presentation differences between older and more recent (movie) productions. For example, the shot pace has drastically increased in modern film-making as more quickly observed visual effects are employed and less dialogue is shown.

3.3.3 Visualisation

SceneMaker's multimedia module will provide tools to specify the visualisation of actors, objects and set, providing 3D models, a scripting notation for events and actions, media allocation and synchronisation. Visual knowledge as defined in CONFUCIUS, such as object models and event models, will be related to emotional cues applying the OCC emotion structure to 3D models. Reasoning about context and decision making on behaviour and other visual elements will be inspired by models of the mind as developed in AESOPWORLD (Okada et al., 1999) and implemented with AI techniques like fuzzy rules or machine learning algorithms (see section 2.2.2). Cinematic principles will be classified into expressive categories.

Resources of 3D models are available on the internet, for instance Microsoft Agents (MS-Agents, 2008) which have functionalities to model personalities and synchronise speech or H-Anim (Humanoid Animation) models (H-Anim, 2001), as used in CONFUCIUS, which include geometric or physical, functional and spatial properties. Maya and 3D Studio Max may be considered for the generation of 3D models. Maya also incorporates an editor for scripting characters as used in Su et al. (2007).

SceneMaker will extend the action representation and event structure, employed in CONFUCIUS, which is based on the event parameters of the LSC database and the Parameterised Action Representation (PAR) (Badler et al., 2000) and defines the event (verb), agent, objects, space, time and manner, to include affective parameters. In CONFUCIUS, the media allocation module selects and coordinates which information is conveyed by which medium or modality based on a limited set of principles. These media allocation rules will need to be expanded for all new modalities planned for *SceneMaker*. Human animation, the transformation of H-Anim models, e.g. moving body and limbs or grasping objects, is prefabricated manually, exported to VRML (1995) and made available in the *animation library*. The *animation engine* checks whether simultaneously occurring motions can be executed or need to be resolved by the user. Microsoft-Agents are controlled with Javascript. The *animation controller* schedules the execution of sequences of basic actions by inserting key-framing information. *SceneMaker* will consider automating this visualisation process and enlarging the animation library by integrating animation descriptions from multimodal corpora and annotation tools, e.g. ANVIL (Kipp, 2001), an XML-based annotation tool for multimodal dialogue. Multimodal annotation and its application are further discussed in Rehm and André (2008). EML (De Melo and Paiva, 2006) is an XML-based scripting language for modelling expressive modalities of virtual humans and cinematic settings and is combined with SMIL (2008) for synchronisation of the various modalities. Hence, it may augment the VRML scripts by affective notations, but might need to be extended to serve all multimodal expressions in *SceneMaker*.

SceneMaker will introduce a *Dramatic-Structure* module to estimate the general dramatic flow of the story which is directly linked to cinematic presentation style changes, as suggested in (Choujaa and Dulay, 2008). Variations of the dramatic intensity may be inferred from the course of the protagonist's emotions, accordingly the detected emotions may be plotted against a timeline to determine the beginning, middle and end of the story. This module then assigns default cinematic styling values for each part of the animation, such as shot pace or camera angles.

3.3.4 Media (Sound, Speech)

For the speech generation from dialogue text, the speech synthesis module in CONFUCIUS will be tested for its suitability in *SceneMaker* with regard to mobile applications and the effectiveness of emotional prosody. CONFUCIUS utilises FreeTTS (2005) which is written in Java and derived from the Festival speech synthesis system (Taylor et al., 1998). An automatic audio selection tool, as proposed in Kuo et al. (2005), will be added for intelligent recommendation of appropriate sound and music according the theme and mood of a scene.

3.3.5 User Interface

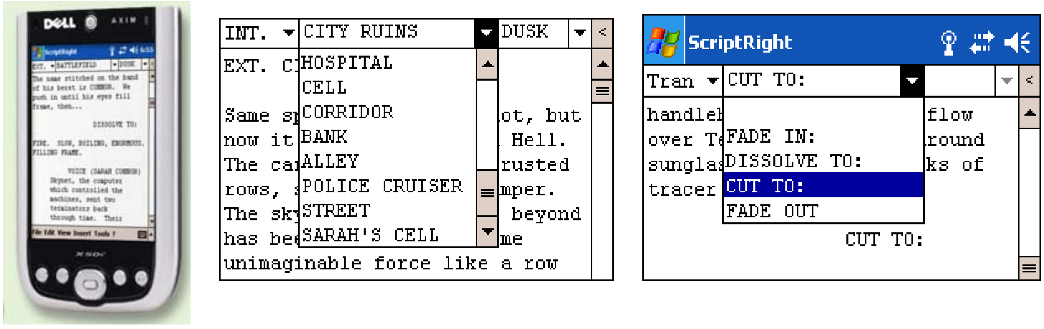
SceneMaker's user interface will consist of three functions, the script input, the animation player and the animation editing. The interface faces usability challenges when operating on desktop computers as well as on small mobile devices. Rendering 3D models on mobile devices is best achieved with low polygon count models as there is no need for more detail on the small display, but a finer polygon structure will be required on the desktop computer. The VRML animation player may be suitable for both output devices.

Web-based, interactive animation presentation including various multimedia may be supported by the Synchronized Multimedia Integration Language, SMIL (2008). SMIL constitutes a W3C standard, XML-style mark-up language to define and synchronise various media data and functionality including audio, video and image. SMIL supports timing, layout, animation, visual transitions and embedded media to be published and interact on the web. The subset, SMIL-Animation (2001), incorporates an animation framework, a timeline and a set of XML animation elements suitable for integration with XML documents. Script input is discussed in detail in the following section.

3.3.6 Online and Mobile Script Writing Software

Script writing software tools assist the writing process of screenplays, stage plays, television shows or commercials. The script is automatically formatted, as the author writes, to meet conventions and submission standards set by the film and theatre industry. The commercial software, FinalDraft (2009), is the professional industry-standard for screenwriting. It offers built-in templates for different script types pre-populated with all standard elements and formats for scene headings, actions, characters, dialogue, shots and transitions. FinalDraft automatically identifies and stores character and location names or complete scene headings and fills select boxes for quick typing. Story scenes can be rearranged and notes attached to script sections. Script changes can be highlighted during the revision process or synchronised in real-time during online collaboration with other users. Powered by the text-to-speech engines built into Windows and Mac OC voices can be assigned to different characters, so that the writer can have the script read back aloud at any time. FinalDraft supports the planning of movie or stage productions beyond the writing process, supplying statistical reports relevant to budget keeping and actor casting, e.g. by giving information on how often and for how long a location or a character appears in the script. FinalDraft files can be saved in an XML based format and thus be easily shared with other software. The relatively comprehensive scriptwriting software, Celtx (2009), is an open-source project and free for download. It also caters for film and theatre scripting, storyboarding and more and can be combined with a web-based non-free collaboration version, Celtx Studios. Furthermore, a great variety of freely available web applications, to be used inside a web browser, come with different degrees of functionality, for example the more complex collaborative story-development or scriptwriting software, FiveSprockets (2009) or the rather simplistic Plotbot (2009).

The commercial ScriptRight (2006) Mobile Edition brings scriptwriting applications onto handheld devices including formatting, indentation and capitalisation, quick tab input for slug lines, character names and transitions, production tracking and note taking features. ScriptRight automatically integrates characters, locations and transitions into select menus as the writer adds them to the script (Figure 11), which is particularly useful on devices where the user mainly interacts using a stylus to pick menu items or to write on the screen. Files can be imported and exported from and to the desktop environment, FinalDraft.



(a) (b) (c)

Figure 11: ScriptRight User Interface

3.4 Testing and Evaluation

Four aspects of this research will be tested and evaluated separately, the correctness and relevance of the screenplay or play script analysis, the effectiveness of the outputted scenes, the suitability for genres and the functionality of the user interface.

The process of computational text interpretation and selection of visualisations for the scenes to be created in *SceneMaker* will be compared against the creative thinking processes as recorded by animators when hand-animating scenes. Test scenarios will be developed based on screenplays of different genres and animation styles, e.g. drama movies, which include precise descriptions of set layout and props versus comedy, which employs plenty of exaggeration techniques for expression. Similarity or closeness of the context interpretation of the human animator and *SceneMaker's Scene Production* module will be qualitatively and quantitatively evaluated. The cinematic effectiveness, emotional believability and appeal of the scenes created in *SceneMaker* will be evaluated against scenes from existing feature films. Three to four example scenes, each of a different genre, will be selected from Hollywood movies and checked against the same scenes produced by *SceneMaker*. In order to analyse whether the animations created by *SceneMaker* suit the intended genre, a test will be conducted where viewers categorise scenes produced by *SceneMaker* of unknown scripts into genre categories and rate the appropriateness in relation to the content. The functionality of *SceneMaker*'s components and their accessibility through the user interface will be tested in cooperation with professional film directors, contrasting the process of directing a scene traditionally with actors or with *SceneMaker*. Time efficiency, levels of creativity and usability of *SceneMaker* will be assessed.

3.5 Project Status, Development and Schedule

This research project involves three key development stages, the literature review, the design, implementation and testing of the *SceneMaker* prototype and the Ph.D. thesis write up. A research schedule is given in Appendix A, Table A.1. In the first year I have reviewed related literature in the areas of natural language processing, affective computing and affective agents, multimodal, interactive storytelling, automatic scene modelling and mobile 3D applications. The general architecture for the *SceneMaker* prototype is outlined and prospective tools for its implementation are surveyed. The next step will be to refine the design of *SceneMaker's* architecture and its individual modules and to determine the layout and interactive elements of the user interface. The prototype will be implemented during the second research year. *SceneMaker* will be tested and improved during the first half of the third research year and the Ph.D. thesis will be written in the second half of the third year. Appendix C provides a preliminary outline of the final thesis, which will be submitted in October 2011.

The overall aim is to have an automated system that can visualise the scenes and scripts and hence assist in the work of film making. These evaluations will give feedback on the viability of realistic and believable text-to-animation production and will indicate features of the prototype that may need to be refined or improved.

4 Relation to Other Work

Research implementing various aspects of modelling affective virtual actors, narrative systems and film-making practices is compared to the proposed *SceneMaker* prototype. Table B.1, Appendix B gives a detailed overview of these related systems. The focus of this research lies on processing natural language screenplays or play scripts to create affective animations, thus mainly systems which use vocal or written language as input to control virtual humans or animate scenes are considered in Table B.1. Multimodal options of input and output media, the extent to which affective or emotional aspects are involved, the level of story processing, the type of story input and interaction with and access to the system are collated.

All systems have their own advantages in the area of multimodal storytelling. The virtual agent systems, Improv (Perlin and Goldberg, 1996), AESOPWORLD (Okada et al., 1999), Behaviour Generation System (Breitfuss et al., 2007), Max (Kopp et al., 2008) and Greta (Pelachaud, 2005 and De Rosis et al., 2003) investigate in great detail, how to model an agent's mind with sophisticated reasoning about emotional states, personality, social and conversational roles. Furthermore, these systems connect those mental processes to facial, gestural and vocal expressions of 3D models achieving human-like conversational behaviour.

The text-to-animation systems concentrate on identifying descriptions of content and activity to animate the body movement and actions of virtual actors. An important characteristic of the fuzzy P&E engine (Su et al., 2007) is that it provides an in-depth rule system tying emotions and personality with affective bodily expression and is the only system integrating story character roles, such as protagonist/hero or villain. ScriptViz (Liu and Leung, 2006), CAMEO (Shim and Kang, 2008) and CONFUCIUS (Ma, 2006) depict the scene environment incorporating cinematic techniques such as effective camera views and, in CAMEO only, theme related lighting. EML (De Melo and Paiva, 2006) supports the affective, multimodal modelling of virtual actors and scene environments proving a comprehensive scripting language. CAMEO is the only system relating specific cinematic direction, for character animation, lighting and camera work, to the genre or theme of a given story. CONFUCIUS and ScriptViz 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 sentence and does not consider the wider context. These text-to-animation systems lack the extensive emotion and personality investigation of the above mentioned virtual agents.

The following observations can be made from Table B.1:

- Concerning emotion detection from input text, none of the discussed projects integrates reasoning about the story context to recognize emotions which requires consulting common knowledge and facilitating a context memory.
- No previous system controls agent behaviour through integrating personality, social status, narrative roles and emotions. SCREAM realises most of these aspects, except narrative roles, though they are not automatically recognised from story text.
- With regard to animation modelling, varying spatial behaviour depending on the mood of characters, for example, whether they approach each other closely or stay at a distance, as well as expressive effects of staging or positioning on stage have not been realised by any system.
- Most of the related systems process one sentence at a time and do not benefit from the analysis of the text layout of screenplays and their dramatic structure. Though the development idea of ScriptViz is motivated by automatic screenplay animation, the current implementation only processes individual well formed sentences.
- Automatic detection of genre or theme from story text has not yet been implemented. The genre type applied to animation directing in CAMEO is explicitly selected by the user.
- No previous system provides a user interface to edit the automatically created output.

SceneMaker will bring all relevant techniques together to form a software prototype system for animation production from natural language scene scripts. *SceneMaker* will be fully automated in intelligent content, action, affect and genre recognition from text input as well as the according 3D visualisation with appropriate application of cinematic techniques. *SceneMaker* will combine existing tools to achieve this goal. *SceneMaker* will not only support screenplays, as suggested in similar systems like ScriptViz and CAMEO, but play scripts or other scripts for different entertainment formats as well. *SceneMaker* will deploy text structure analysis to detect semantic context information. *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 text. *SceneMaker* will introduce a new approach to automatic genre detection from screenplays or play script implementing keyword co-occurrence, term frequency and dialogue and scene length calculations. *SceneMaker* will present a unique user interface for directors or animators to control and adjust the automatic scene production. No other automatic animation production system endows an online or mobile accessible user interface to directly and manually manipulate the scene.

SceneMaker's GUI will consider 3D rendering and editing on small displays and limited power technology to facilitate easy access from mobile devices and rapid scene testing on set.

5 Conclusion

This research proposes to contribute to believability and artistic quality of automatically produced animated, multimedia scenes. The research aims are to advance knowledge in the areas of affective computing, digital storytelling and expressive multimodal systems through the development of the software prototype, *SceneMaker*, which automatically visualises affective expressions of screenplays or play scripts. *SceneMaker's* mobile and web-based user interface will assist directors, drama students, writers and animators in the testing of their ideas. Thus *SceneMaker* will considerably shorten the production time, reduce production costs and improve collaboration and communication. Existing systems solving partial aspects of natural language processing, emotion modelling and multimodal storytelling have been reviewed. 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.

SceneMaker's architecture will comprise a scene production module running on a central server and a web-based, mobile accessible user interface with a 3D player and editor for manual adjustments of the animations. The scene production module will be divided into three modules for (1) natural language processing, (2) semantic and affective reasoning and decision making and (3) multimodal visualisation. *SceneMaker* will integrate the natural language processing module and scene visualisation module of CONFUCIUS (Ma, 2006) with a comprehensive, cognitively-grounded emotion model, a multimodal affective knowledge base, affective reasoning and affective visualisation in behaviour and cinematic expression. Potential software and tools relevant for the implementation of *SceneMaker,* such as script writing interfaces, emotion and genre recognition from scene scripts, affective script annotation and intelligent cinematography, have been presented. Future work involves the implementation of the proposed system prototype, *SceneMaker*, by testing the integration into CONFUCIUS and efficiency of the prospective tools. Accuracy of content animation, effectiveness of expression and usability of the interface will be evaluated in comparison to manual animation, feature film scenes and real-life directing. In relation to similar projects, *SceneMaker,* aiming to enhance believability of virtual actors and scene presentation, will incorporate an expressive model for multiple modalities, including prosody, facial expressions, gestures, body posture, illumination, sound, music, staging and camera work. Emotions will be inferred from context. Genre types will be automatically derived from the scene scripts and so influence the design style of the output animation. The 3D output will be editable on mobile devices.

In conclusion, this research intends to automatically and intelligently produce multimodal animations with heightened expressivity and visual quality from screenplay or play script input.

References

3ds Max (2009). *Autodesk 3ds Max*. http://usa.autodesk.com/adsk/servlet/index?id=5659302&siteID=123112. Accessed June 2009.

Alm, C. O., Roth, D. and Sproat, R. (2005). “Emotions from text: machine learning for text-based emotion prediction”, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Human Language Technology Conference, Association for Computational Linguistics, Morristown, USA, 579-586.

Badler, N. I., Bindiganavale, R., Allbeck, J., Schuler, W., Zhao, L., and Palmer, M. (2000). “Parameterized action representation for virtual human agents”, *Embodied Conversational Agents,* MIT Press, Cambridge, USA, 256-284.

Bartneck, C. (2002). “Integrating the OCC Model of Emotions in Embodied Characters”, *Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges*. Melbourne.

Ball, G. and Breese, J. (2001). “Emotion and Personality in a Conversational Agent”, *Embodied Conversational Agents,* MIT Press, Cambridge, USA, 189–219.

Bates, J. (1994). “The role of emotion in believable agents”, *Communications,* ACM, New York, USA, 37 (7), 122-125.

Bratman, M. (1987). *Intention, plans, and practical reason*, Harvard University Press, Cambridge, Mass.

Breazeal, C., Brooks, A., Gray, J., Hancher, M., McBean, J., Stiehl, W.D., and Strickon, J. (2003). “Interactive Robot Theatre”, *Communications,* ACM, New York, USA, 46 (7), 76-85.

Breitfuss, W., Prendinger, H., and Ishizuka, M. (2007). “Automated generation of non-verbal behavior for virtual embodied characters”, *Proceedings of the 9th international Conference on Multimodal interfaces,* ICMI '07, ACM, New York, USA, 319-322.

Cano, P., Koppenberger, M., and Wack, N. (2005). “Content-based music audio recommendation”, *Proceedings of the 13th Annual ACM international Conference on Multimedia*, MULTIMEDIA '05, ACM, New York, USA, 211-212.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). “BEAT: the Behavior Expression Animation Toolkit”, *Proceedings of the 28th Annual Conference on Computer Graphics and interactive Techniques*, SIGGRAPH '01, ACM, New York, USA, 477-486.

Cavazza, M., Lugrin, J., Pizzi, D., and Charles, F. (2007). “Madame Bovary on the holodeck: immersive interactive storytelling”, *Proceedings of the 15th international Conference on Multimedia*, MULTIMEDIA '07, ACM, New York, USA, 651-660.

Celtx (2009). *Integrated Media Pre-production*, http://www.celtx.com. Accessed May 2009.

Choujaa, D. and Dulay, N. (2008). “Using screenplays as a source of context data”, *Proceeding of the 2nd ACM international Workshop on Story Representation, Mechanism and Context*, SRMC '08, ACM, New York, USA, 13-20.

Christodoulou, Y., Mavrogeorgi, N. and Kalogirou, P. (2008).
“Use of Ontologies for Knowledge Representation of a Film Scene”, *Third International Conference on Internet and Web Applications and Services, ICIW '08,* 662-667.

Connexor (2003). *Connexor Machinese Software*, http://www.connexor.eu/technology/machinese. Accessed December 2008.

Coyne, B. and Sproat, R. (2001). “WordsEye: an automatic text-to-scene conversion system”, *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press, Los Angeles, 487-496.

De Carolis, N., Pelachaud, C., Poggi, I. and De Rosis, F.(2001). “Behavior planning for a reflexive agent”, *International Joint Conference on Artificial Intelligence,* IJCAI'01, Seattle, USA.

De Melo, C. and Paiva, A. (2006). “Multimodal Expression in Virtual Humans”, *Computer Animation and Virtual Worlds 2006,* John Wiley & Sons Ltd., 17 (3-4), 239-348.

De Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V. and De Carolis, B. (2003). “From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent”, *International Journal of Human-Computer Studies,* Applications of Affective Computing in Human-Computer Interaction, 59 (1-2), 81-118.

Disney Animation, (2008). *12 basic principles of animation*. http://en.wikipedia.org/wiki/12_basic_principles_of_animation. Accessed November 2008.

Ekman, P. and Rosenberg E. L. (1997). “What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system”, *Oxford University Press.*

Elliott, C. D. (1992). “The Affective Reasoner: a Process Model of Emotions in a Multi-Agent System”, *Doctoral Thesis,* Northwestern University.

EMMA, (2003). *EMMA: Extensible MultiModal Annotation markup language*. http://www.w3.org/TR/emma. Accessed May 2009.

Fellbaum, C. (1998). “WordNet: An Electronic Lexical Database”, MIT Press, Cambridge, USA.

Film Genre (2009). http://en.wikipedia.org/wiki/Film_genre. Accessed March 2009.

FinalDraft (2009). *Final Draft Software*. http://finaldraft.com/products-and-services/final-draft. Accessed May 2009.

FiveSprockets (2009). *Free online screenwriting and media-production software and resources*. http://www.fivesprockets.com. Accessed May 2009.

Francisco, V., Hervás, R. and Gervás, P. (2006). “Two Different Approaches to Automated Mark Up of Emotions in Text”, *Proceedings of the AI 2006*, Cambridge, 101-114.

FreeTTS (2005). *FreeTTS 1.2 - A speech synthesizer written entirely in the Java^{TM} programming language.* http://freetts.sourceforge.net/docs/index.php. Accessed May 2009.

Friedman, D.A., and Feldman, Y.A. (2004). “Knowledgebased cinematography and its applications”, *Proceedings of ECAI*, 256-262.

Gebhard, P. (2005). ”ALMA - Layered Model of Affect”, *Proceedings of the Fourth International Conference on Autonomous Agents and Multiagent Systems.* AAMAS 05, ACM, New York, USA, 29-36.

Gebhard, P., Kipp, M., Klesen, M., and Rist, T. (2003). “Authoring scenes for adaptive, interactive performances”, *Proceedings of the Second international Joint Conference on Autonomous Agents and Multiagent Systems,* AAMAS '03, ACM, New York, USA, 725-732.

Gebhard, P. and Schröder, M. (2008). “IDEAS4Games – A.I. Poker im Casino Virtuell”, DFKI Newsletter 1/2008.

Good Will Hunting (1997). *Draft Script*. http://sfy.ru/sfy.html?script=good_will_hunting. Accessed June 2009.

Gunes, H. and Piccardi, M. (2006). “A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior”, *18th International Conference on Pattern Recognition*, ICPR, IEEE Computer Society, Washington, USA, 1, 1148-1153.

Gutiérrez, M., Vexo, F., and Thalmann, D. (2004). “The Mobile Animator: Interactive Character Animation in Collaborative Virtual Environments”, *Proceedings of the IEEE Virtual Reality 2004*, VR ’04, IEEE Computer Society, Washington, USA, 125-132.

H-Anim (2001). *Humanoid animation working group*. http://www.h-anim.org. Accessed December 2008.

Heylen, D.K.J., Nijholt, A. and op den Akker, H.J.A. (2005). “Affect in Tutoring Dialogues”, *Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors),* Taylor and Francis Inc., 19 (3-4), 287-310.

JSR 184 (2005). *JSR 184: Mobile 3D Graphics API for J2ME*. http://www.jcp.org/en/jsr/detail?id=184. Accessed May 2009.

Kelleher, J., Doris T., Hussain, Q. and Ó Nualláin, S. (2000). “SONAS: Multimodal, Multi-User Interaction with a Modelled Environment”, Spatial Cognition, John Benjamins Publishing Company, Amsterdam, Netherlands, 171-184.

Kennedy, K. and Mercer, R. E. (2002). “Planning animation cinematography and shot structure to communicate theme and mood”, *Proceedings of the 2nd international Symposium on Smart Graphics,* SMARTGRAPH '02, ACM, New York, USA, 24, 1-8.

Keyword Co-Occurrences (2005). *Keywords Co-Occurrence and Semantic Connectivity*. http://www.miislita.com/semantics/c-index-1.html. Accessed May 2009.

Kipp, M. (2001). “ANVIL - a generic annotation tool for multimodal dialogue”, *EUROSPEECH-2001*, Aalborg, Denmark, 1367-1370.

Kopp, S., Allwood, J., Grammer, K., Ahlsen, E. and Stocksmeier, T. (2008).“Modeling Embodied Feedback with Virtual Humans”, *Modeling Communication with Robots and Virtual Humans,* Springer Berlin/Heidelberg, 18-37.

Kuo, F., Chiang, M., Shan, M., and Lee, S. (2005). “Emotion-based music recommendation by association discovery from film music”, *Proceedings of the 13th Annual ACM international Conference on Multimedia*, MULTIMEDIA '05, ACM, New York, USA, 507-510.

Lazarus, R.S. (1966). *Psychological stress and the coping process,* McGraw-Hill, New York, USA.

LCS (2000). *Lexical Conceptual Structure Database*. http://www.umiacs.umd.edu/~bonnie/LCS_Database_ Documentation.html. Accessed December 2008.

Lenat, D.B. (1995). “CYC: A large-scale investment in knowledge infrastructure”, Communications, ACM, New York, USA, 38(11), 33-38.

Liu, H., Lieberman, H., and Selker, T. (2003). “A model of textual affect sensing using real-world knowledge”, *Proceedings of the 8th international Conference on intelligent User interfaces,* IUI '03, ACM, New York, USA, 125-132.

Liu, H. and Singh, P. (2004). “ConceptNet: A practical commonsense reasoning toolkit”, *BT Technology Journal*, Springer Netherlands, 22(4), 211-226.

Liu, Z. and Leung, K. (2006). “Script visualization (ScriptViz): a smart system that makes writing fun”, *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* Springer Berlin/Heidelberg, 10 (1), 34-40.

M3G (2004). *Getting Started With the Mobile 3D Graphics API for J2ME*. http://developers.sun.com/mobility/ apis/articles/3dgraphics. Accessed May 2009.

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

Marsella, S. and Gratch, J. (2003). “Modeling coping behavior in virtual humans: don't worry, be happy”, *Proceedings of the Second international Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, ACM, New York, USA, 313-320.

Masuch, M., Hartman, K., and Schuster, G. (2006). “Emotional Agents for Interactive Environments”, *Proceedings of the Fourth international Conference on Creating, Connecting and Collaborating Through Computing*, C5, IEEE Computer Society, Washington, USA, 96-102.

Maya (2009). *Autodesk Maya*. http://usa.autodesk.com/adsk/servlet/index?id=7635018&siteID=123112. Accessed May 2009.

McCrae, R. and John, O. (1992). “An Introduction to the Five-Factor Model and its Applications”, *Journal of Personality,* Routledge, 60 (2), 175–215.

McDonnell, R., Jörg, S., McHugh, J., Newell, F., and O'Sullivan, C. (2008). “Evaluating the emotional content of human motions on real and virtual characters”, *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, APGV '08, ACM, New York, USA, 67-74.

Mehrabian, A. (1995). “Framework for a Comprehensive Description and Measurement of Emotional States”, *Genetic, Social, and General Psychology Monographs*, Heldref Publishing, 121 (3), 339-361.

Miller, G. A. (1995). “WordNet: a lexical database for English”, *Communications*, ACM, New York, USA, 38 (11), 39-41.

MS-Agents (2008). *Microsoft Agents*. http://msdn.microsoft.com/en-us/library/ms695784(VS.85).aspx. Accessed December 2008.

Nadalutti, D., Chittaro, L., and Buttussi, F. (2006). “Rendering of X3D content on mobile devices with OpenGL ES”, *Proceedings of the Eleventh international Conference on 3D Web Technology*, Web3D '06, ACM, New York, USA, 19-26.

Nasukawa, T. and Yi, J. (2003). “Sentiment analysis: capturing favorability using natural language processing”, *Proceedings of the 2nd international Conference on Knowledge Capture*, K-CAP '03, ACM, New York, USA, 70-77.

Okada, N., Inui, K. and Tokuhisa, M. (1999). “Towards affective integration of vision, behavior, and speech processing”, *Proceedings of the Integration of Speech and Image Understanding,* SPELMG, IEEE Computer Society, Washington, USA, 49-77.

Ó Nualláin, S. and Smith, A. G. (1994). “An investigation into the common semantics of language and vision”, *Artificial Intelligence Review*, Springer Netherlands,8 (2), 113-122.

OpenGL ES (2005). *OpenGL ES - The Standard for Embedded Accelerated 3D Graphics*. http://www.khronos.org/opengles. Accessed May 2009.

OMCS (2009). *Open Mind Common Sense*. http://commons.media.mit.edu/en. Accessed April 2009.

Opinmind (2008). http://www.opinmind.com. Accessed April 2009.

Ortony, A., Clore, G. L. and Foss, M. A. (1987). “The referential structure of the affective lexicon”, *Cognitive Science,* Cognitive Science Society, Inc., 11 (3), 341-364.

Ortony A., Clore G. L., and Collins A. (1988). “The Cognitive Structure of Emotions”, Cambridge University Press, Cambridge, USA.

Paggio, P. and Music, B. (2001). “Linguistic Interaction in Staging – a Language Engineering View”, *Virtual Interaction: Interaction in Virtual Inhabited 3D Worlds*, Springer, London, 235-249.

Pelachaud, C. (2005). “Multimodal expressive embodied conversational agents”, *Proceedings of the 13th Annual ACM international Conference on Multimedia*, MULTIMEDIA '05, ACM, New York, USA, 683-689.

Perlin, K. and Goldberg, A. (1996). “Improv: a system for scripting interactive actors in virtual worlds”, *Proceedings of the 23rd Annual Conference on Computer Graphics and interactive Techniques,* SIGGRAPH '96, ACM, New York, USA, 205-216.

Physically Informed Audio Synthesis (2008). http://www.sarc.qub.ac.uk/main.php?page=projects&projID=48. Accessed December 2008.

Plotbot (2009). *Write screenplays in your browser*. http://www.plotbot.com. Accessed March 2009.

Pocket Cortona (2009). http://www.parallelgraphics.com/products/cortonace. Accessed May 2009.

Prendinger, H., Descamps, S. and Ishizuka, M. (2002). "Scripting affective communication with life-like characters in web-based interaction systems", *Applied Artificial Intelligence Journal*, Taylor & Francis, 16 (7-8), 519-553.

Preuß, J. and Loviscach, J. (2007). "From movie to comic, informed by the screenplay", *ACM SIGGRAPH 2007 Posters*, SIGGRAPH '07, ACM, New York, USA, 99.

Ramírez, A. (2005). "Towards a cinematically enhanced narrative", *Proceedings of the 2005 ACM SIGCHI international Conference on Advances in Computer Entertainment Technology,* ACE '05, ACM, New York, USA, 265, 365-366.

Rasheed, Z., Sheikh, Y., Shah, M. (2005). "On the use of computable features for film classification", *IEEE Transactions on Circuits and Systems for Video Technology,* IEEE Circuits and Systems Society, 15(1), 52-64.

Rebelo, P., Alcorn, M. and Wilson, P. (2005). "A Stethoscope for Imaginary Sound: Interactive Sound in a Health Care Environment", *International Computer Music Conference Proceedings,* ICMC 05, Barcelona.

Rehm, M. and André, E. (2008). "From Annotated Multimodal Corpora to Simulated Human-Like Behaviors", *Modeling Communication*, Springer-Verlag, Berlin Heidelberg, 1-17.

RIVME (2004). *Real-time Immersive Virtual MOCAP Environment*. http://accad.osu.edu/~sgencogl/mocap/mocap.htm. Accessed November 2008.

Salway, A. and Graham, M. (2003). "Extracting information about emotions in films", *Proceedings of the Eleventh ACM international Conference on Multimedia*, MULTIMEDIA '03, ACM, New York, USA, 299-302.

Schröder, Marc (2001). "Emotional speech synthesis: a review", *Proceedings of the 7th European Conference on Speech Communication and Technology*, EUROSPEECH-2001, Scandinavia, 561-564.

ScriptRight (2006). *ScriptRight Mobile Edition Features*. http://www.scriptright.com. Accessed May 2009.

Shaarani, A. S. and Romano, D. M. (2008). "The intensity of perceived emotions in 3D virtual humans", *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, International Conference on Autonomous Agents, International Foundation for Autonomous Agents and Multiagent Systems, Richland, USA, 3, 1261-1264.

Shaikh, M.A.M., Prendinger, H. and Ishizuka, M. (2007). "SenseNet: A linguistic tool to visualize numerical-valence based sentiment of textual data", *Proceedings of the International Conference on Natural Language Processing*, ICON, Hyderabad, India, 147–152.

Shaikh, M.A.M., Prendinger, H. and Ishizuka, M. (2009). "A Linguistic Interpretation of the OCC Emotion Model for Affect Sensing from Text", *Affective Information Processing*, Springer London, 45-73.

Shim, H. and Kang, B. G. (2008). "CAMEO - camera, audio and motion with emotion orchestration for immersive cinematography", *Proceedings of the 2008 international Conference on Advances in Computer Entertainment Technology*, ACM, New York, USA, 352, 115-118.

SMIL (2008). *Synchronized Multimedia Integration Language (SMIL 3.0)*. http://www.w3.org/TR/2008/REC-SMIL3-20081201. Accessed May 2009.

SMIL-Animation (2001). *SMIL Animation*. www.w3.org/TR/smil-animation. Accessed May 2009.

Smith, S. and J. Bates (1989). “Toward a theory of narrative for interactive fiction”, Technical Report CMU-CS-89-121, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

Sowa, T. (2008). “The Recognition and Comprehension of Hand Gestures - A Review and Research Agenda”, *Modeling Communication with Robots and Virtual Humans*, Springer Berlin/Heidelberg, 38-56.

Strapparava, C. and Valitutti, A. (2004). “WordNet-Affect: an affective extension of WordNet”, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC 2004, 4, 1083-1086.

Strapparava, C. and Mihalcea, R. (2008). “Learning to identify emotions in text”, *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, ACM, New York, USA, 1556-1560.

Su, W-P., Pham, B., Wardhani, A. (2007). “Personality and Emotion-Based High-Level Control of Affective Story Characters”, *IEEE Transactions on Visualization and Computer Graphics*, 13 (2), 281-293.

Taylor, P., Black, A. and Caley, R. (1998). “The architecture of the Festival Speech Synthesis system”, *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 147-151.

Term Frequency (2009). *Tf-idf*. http://en.wikipedia.org/wiki/Term_frequency. Accessed May 2009.

Terra, E. and Clarke, C. L. (2004). “Fast computation of lexical affinity models”, *Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics, Morristown, USA, 1022.

Tesniere, L. (1959). *Elements de syntaxe structurale*. Klincksieck, Paris.

Thomas, F. and Johnson O. (1981, reprint 1997). *The Illusion of Life: Disney Animation*. Abbeville Press/Hyperion, 47-69.

Thórisson, K. R., Pennock, C., List, T., and DiPirro, J. (2004). “Artificial intelligence in computer graphics: a constructionist approach”, *SIGGRAPH Computer Graphics*, ACM, New York, 38 (1), 26-30.

Thórisson, K. R. (2007). “Integrated A.I. systems”, *Minds & Machines*, Springer, Netherlands, 17 (1), 11-25.

Turetsky, R. and Dimitrova, N. (2004). “Screenplay alignment for closed-system speaker identification and analysis of feature films”, *IEEE International Conference on Multimedia and Expo*, ICME '04, 3, 1659-1662.

Turunen, M and Hakulinen, J. (2007). “Spoken and Multimodal Communication Systems in Mobile Settings“, *Verbal and Nonverbal Communication Behaviours*, LNAI 4775, Springer-Verlag, Berlin/Heidelberg, 227-241.

Turunen, M., Salonen, E-P., Hakulinen, J., Kanner, J. and Kainulainen, A. (2005). “Mobile architecture for distributed multimodal dialogues”, *Proceedings of Applied Spoken Language Interaction in Distributed Environments*, ASIDE-2005, paper 22.

Vish, V. (2004). “Moved by Movements: How Character Movements Cue Us to Form Specific Genre and Affective Impressions“, *Entertaining Computing – ICEC 2004*, Springer Berlin/Heidelberg, 3166/2004, 168-171.

Vilhjálmsson, H. and Thórisson, K.R, (2008). “A Brief History of Function Representation from Gandalf to SAIBA”, *Proceedings of the 1st Function Markup Language Workshop*, AAMAS, Portugal, 61-64.

Virtual Theatre (2004). *The Virtual Theatre Project*. http://accad.osu.edu/research/virtual_environment_htmls/ virtual_theatre.htm. Accessed November 2008.

VoiceXML (2004). *Voice Extensible Markup Language*. http://www.w3.org/TR/voicexml20. Accessed June 09.

VRML (1995). *Virtual Reality Modeling Language*. http://www.w3.org/MarkUp/VRML. Accessed May 2009.

Wahlster, W. (2006). “Smartkom: Foundations of Multimodal Dialogue Systems”, Springer Verlag.

Appendix A – Research Schedule

🟦	Activities
🟩	Submissions
🟥	Deliverables

Research Activities	1st year				2nd year				3rd year			
	2008	2009				2010				2011		
	Oct-Dec	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Jan-Mar	Apr-Jun	Jul-Sep
Literature Review	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
100 Day Review and Presentation		🟥										
Submission to ISEA2009	🟩											
In-Depth Review of Systems and Approaches Relevant for Text to Animation Modelling		🟦	🟦	🟦								
Submission to IMVIP 2009			🟩									
Submission to KI 2009			🟩									
Development of detailed System Architecture for SceneMaker			🟦	🟦								
Submission to AICS 2009			🟩									
Confirmation Report and Presentation			🟥									
Submission to 'Artificial Intelligence Review' Journal by Springer				🟩								
Design Automated Scene Production Module of SceneMaker				🟦	🟦							
Design User Interface Module of SceneMaker					🟦							
2nd Year Poster						🟥						
Implementation of Scene Production Module					🟦	🟦	🟦					
Implementation of User Interface in Accordance with HCI							🟦	🟦				
Submission to IEEE Pervasive Computing Journal								🟩				
Test and Evaluate									🟦	🟦		
Improve Developed SceneMaker Prototype										🟦		
Submission to ACM Transactions on Multimedia Computing, Communications and Applications										🟩		
3rd Year Presentation										🟥		
Thesis write up										🟦	🟦	🟥

Table A.1: Research Schedule

NL	Natural Language
SP	Screenplay
D	Dialogue
NS	Not Specified

Appendix B – Relation to Other Work

System	Year	Category	Input (Perception)				Output (Generation)									Emotions					Story Processing			Story Type		Author Options		Access		Implementation
			Text	Speech	Other (Code, GUI)	Other (Vision, Gesture)	Text	Speech	Gaze	Facial Expression	Body Language	Music/Sound	Camera	Lighting	Genre	Basic Emotions	OCC, PAD, BDI	Personality	Social Roles	Narrative Rolls	Layout/Dramatic Analysis	Sentence-based	Context Memory	Linear Storyline	Interactive Story	User Interface	Code/Rules Extendable	Mobile Devices	Internet	
Improv	1996	Virtual Agents		✓	✓	✓		✓		✓	✓	✓				✓		✓							✓	✓	✓		✓	
AESOPWORLD	1999		NL	✓		✓		✓			✓	✓				✓		✓					✓	✓						Prolog, Blackboards
Max	2008		D			✓		✓	✓	✓	✓						✓	(✓)				✓			✓					
Greta	2005		D		✓			✓	✓	✓	✓						✓	✓	✓			✓			✓					APML, Festival, MPEG-4
SCREAM	2002	Scripting Tools	D		✓			✓			✓					✓		✓	✓				✓		✓		✓		✓	Java, Prolog, Ms-Agents
EML	2006			✓				✓		✓	✓	✓	✓	✓		✓	✓					✓		✓					✓	Festival,SABLE, GestuRA, ANVIL, SMIL
WordsEye	2001	Text to Animation	(NL)								✓		✓	✓								✓		✓		✓			✓	WordNet
Spoken Image	94/01		NL	✓		✓																✓		✓						
Fuzzy P&E Engine	2007		SP								✓					✓		✓		✓			✓	✓		✓				Fuzzy Rules, Maya
Behaviour Generation System	2007		NL					✓	✓		✓								✓			✓		✓						BEAT, MPML3D
ScriptViz	2006		SP							✓	✓		✓			✓						✓		✓		✓				OpenGL, Java, ApplePie
CAMEO	2008		SP		✓			✓		NS	✓	✓	✓	✓	✓	NS							✓	✓		✓				XML
CONFUCIUS	2006		NL					✓			✓	✓	✓									✓		✓						Java, Connexor Parser, WordNet, VRML, H-Anim
SceneMaker	2009		SP					✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓		✓	✓	

Table B.1: Comparison of Affective and Multimodal Story Visualisation Systems

Appendix C – Preliminary Thesis Outline

Prelims

- Title page, Table of Contents, List of Figures, List of Tables, List of Abbreviations

Abstract

1. Introduction

- Theoretical foundations
- Overview of thesis

2. Literature Review

- Previous work in area of natural language processing and intelligent multimodal visualisation

3. Automatic, Intelligent, Multimodal and Affective Text-to-Animation Generation

- Theoretical contributions

4. Implementation of the Text-to-Animation System, SceneMaker

- Description of software prototype

5. Evaluation

- Test results and evaluation of text-to-animation generation with SceneMaker

6. Conclusion

- Summary, relation to other work, future work

Appendices

References