The Prometheus project — the challenge of disembodied and dislocated performances

J M Thorne and D J Chatting

The Prometheus project is seeking to create a virtual studio production chain. This paper discusses the technologies that are being investigated and in particular the research in progress within the Content and Coding Laboratory in BTexact Technologies. Also presented are the historical changes in the production of real and animated performances and how Prometheus and current trends can have an impact upon this. Two areas are identified where further study is required to ensure believability and coherence when virtual performances are created from reusable digital components.

1. Introduction

Prometheus is a three-year collaborative LINK project under the Broadcast Technology Programme funded by the UK DTI and EPSRC, led by the BBC. The project includes markerless face and body tracking, actor and clothing model animation, scene construction and three-dimensional display technologies. It is seeking to build an entire production framework to encapsulate these technologies.

This will open up new possibilities in the creation, distribution and display of multimedia content, and promises a revolution in which the director and viewer of performances will have unprecedented powers. Actors, scenes and props become components to be arranged at will.

This paper considers these technologies, discussing the challenges that actors and directors now face and the new powers of the viewer. It also reflects on the historic changes in 'entertainment experiences'.

2. The Prometheus project

The Prometheus partners are the BBC, BTexact Technologies, AvatarMe, Snell and Willcox, UCL, Queen Mary University of London (QMUL), De Montford University and the University of Surrey. The BBC as lead partner is aware of the limitations of existing 2-D film media and is here trying to create an equivalent 3-D production chain. The project addresses everything from creation and animation of models through to 3-D displays. BTexact Technologies is responsible for the markerless real-time motion capture of the actors' faces and the creation and animation of photo-realistic 3-D head models. A more detailed overview of the project is given by Price and Thomas [1].

Photos or short sequences of frames from multiple cameras are used to automatically create realistic 3-D avatars of the actors complete with bone structure and facial features that can be animated. Markerless real-time motion tracking is then employed to capture the movements, voice and facial expressions of the actors in the studio. This tracking data is used to animate the avatars. Real-time cloth simulations drape the avatars in the latest fashions and virtual studio techniques place them in the scene. MPEG-4 [2] compression and streaming transmits the complete performance to the end user who can explore it in full 3-D.

MPEG-4 is the newest standard to be released by the Moving Picture Experts Group, providing further leaps in audio and video compression technology and a platform for true multimedia experiences. An MPEG-4 stream contains information about a 3-D stage on which are placed images, video, audio sources and 2- or 3-D mesh objects. All of these can be animated. In addition, MPEG-4 defines specific ways of encoding the human form and its movements, paying particular attention to the intricacies of the face. It also includes the Internet concepts of hyper-linking and scripting, thus enabling sophisticated interaction with the viewer. MPEG-4 shares many things in common with VRML [3] and is of a similar file format to QuickTime [4].

While there are still some exciting challenges left to address, many of the initial technological hurdles have been overcome.

A novel editing studio is being developed whereby the director can change the lighting, the set, the clothes worn by the actors, or even replace one actor model with another. For non real-time performances the director may ultimately edit the motions and gestures of the actors — removing a limp here and adding some more anger there. Moreover, because of the use of MPEG-4, the viewer gains the ability to make choices as they watch and choose between threads as the performance unravels.

Full 3-D body models can be assembled from still images and tracked without markers against a blue screen. A cross-platform avatar tool-kit has been developed to interpret the animation streams and manipulate the avatars.

Simulated cloth exhibiting the properties of its real-life counterpart can be draped over these avatars, though for complex cloth the intense computation required limits the speed to less than real time. MPEG-4 compression codecs are being evaluated and integral imaging techniques and displays have been developed to render 3-D scenes so that the viewer can inspect the scene from an infinite number of viewpoints.

Photo-realistic head models can be created from two photographs. Head movements and changes in facial expressions can be tracked in real-time in a normal office environment without placing markers on the face, and MPEG-4 compatible animation streams can be derived to represent this.

2.1 *Creating photo-realistic 3-D head models*

Sufficient information to create a realistic head model can be obtained from just two photographs of the face — one from the front and the other from the side (Fig 1). Image processing techniques can be applied to these photos to automatically locate important feature positions and the shape of the head. A generic head model can then be deformed to look like the person by comparing the locations of these feature points in both the photos and in the generic head. The photos of the person can then be projected on to this conformed head and stitched together to form a seamless texture. The resulting heads are easy to animate because we already know the locations of the important features. The Prometheus head is MPEG-4 compatible.

2.2 *Markerless face tracking*

The challenge of face tracking is to capture the expressive qualities of the face, which manifest themselves visibly through the movement of features and wrinkling of skin, then to store and replicate on a computer-generated puppet. Traditionally these changes have been captured using markers placed on the face; however, using computer vision techniques markerless systems are becoming viable.

The Prometheus face tracker uses a single camera to capture the actor's performance in real time, from which it derives the head's orientation and the facial expression. This can be encoded as an MPEG-4 stream and be used to animate the head previously created. The tracking algorithm is based on earlier work in BTexact by Machin [5] and Mortlock et al [6] (see Fig 2). With a real-time constraint, the tracking data will be inherently probabilistic and unreliable; in addition, with a single camera depth, information is not readily available. For these reasons the interpretation of the raw tracking data is crucial. Where frames are missing or features adjudged unreliable, estimates must be made of reality. However, we must be aware that we are manipulating a very sensitive communications mode, human facial expression. The human face can convey the extremes of emotion, with a small variation of muscle combinations. If we fail to interpret the tracking data correctly, the result can have a

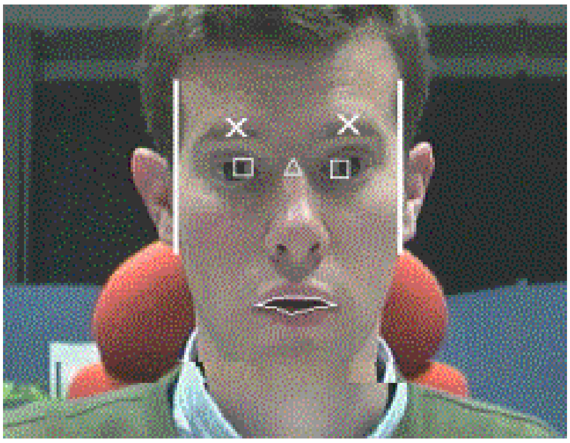

Fig 2 A screen-shot from the face tracker.

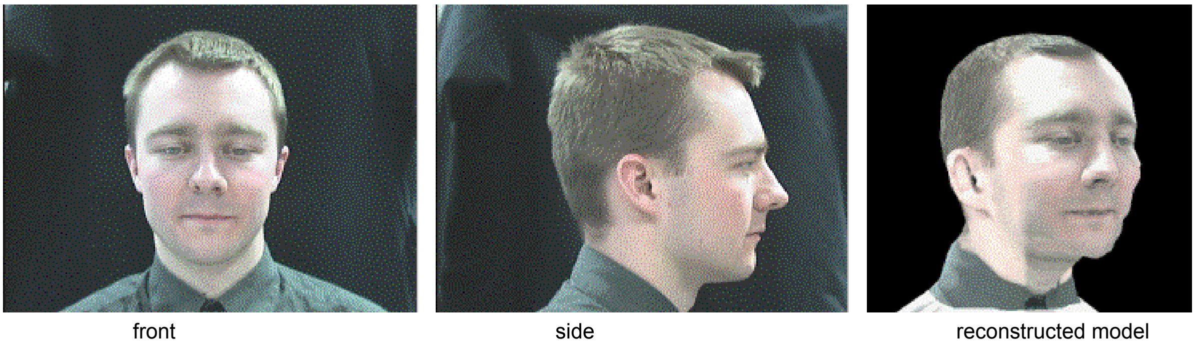

Fig 1 A generic head model constructed from two photographs.

huge impact on the meaning of a performance. Interpretation is the focus of our current research.

3. Considering the future of entertainment experiences

How do technologies like MPEG-4 and projects like Prometheus develop and extend current possibilities for programme makers and consumers and magnify challenges that have existed in the industry for many years? This section first considers the history of entertainment experience and then examines these future challenges more thoroughly.

3.1 A brief history of theatre, film and TV

In the beginning there was theatre. Here actors and audience interacted in a live 3-D arena. The story unfolded before the viewers, adapting and changing in a continual feedback loop with the audience. If the director had allowed for it, the story could diverge down any one of a number of routes (and even if they had not, a few well-aimed tomatoes could dramatically alter the course of a performance). Simple tricks interwoven with the story could provide the impression of the impossible happening. By changing clothes and their style of acting (voice, mannerisms) actors could pretend to be different people. Under the cover of a curtain a quick change of set and lights could conjure up a different place or time. Hidden trapdoors, ropes, props, magic tricks and simple effects (thunder, smoke) could further allow the suspension of reality. However, everything had to be planned carefully in advance with much rehearsing since in the live performance everything has to be right first time.

Along came film — in essence nothing more than a discrete sequence of images. These are replayed fast enough to fool the human eye into accepting the impression of continuous movement. The recorded nature of film however, gives the director great gains in flexibility. They can split up and rearrange the images — to remove and replace bits that went wrong the first time, or to alter the order and pace of story. The soundtrack of the film can also be independently rerecorded, so that effects can be added or language changed.

The ability to make the camera lie created the world of visual effects. Carefully scaled models can create impossible places or events — giant star cruisers or long lost cities. Films themselves can be blended together — for example, the teleportation effects in Star Trek were accomplished through the blending of the actor, the empty stage and some illuminated aluminium foil strips.

Then came TV and video which not only brought the viewer new freedom to watch what they wanted when they wanted it but also imposed tighter production schedules and smaller budgets. This prompted new ideas in the reuse of media.

In both TV and film, however, the viewers are restricted to seeing just what the director intended, they cannot change their point of view and cannot alter the story. Theatre still has not been replaced — the atmosphere and 'liveness' can still hold an audience enthralled.

In parallel with all this was the development of animation — from the first animators, such as Winsor McCay and his hand-drawn animation *Gertie the Dinosaur* in 1914 [7] (see Fig 3), through Walt Disney's lavish productions and Hanna-Barbera's low-budget TV shows, to the computer-generated wonders of *Toy Story* and *Final Fantasy*.

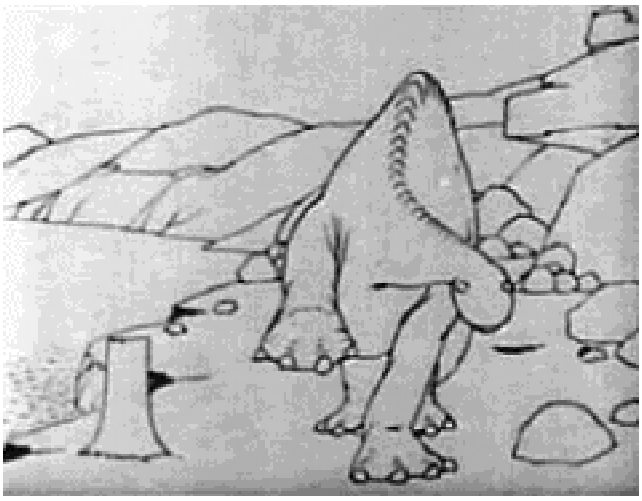

Fig 3 Gertie the Dinosaur.

Throughout this history there has always been a close connection with live action. Winsor McCay conversed on stage with his creation Gertie on screen behind him. Whenever a close-up shot was used of the puppets in *Captain Scarlet* the hands seen were those of real people. Films like *Who Framed Roger Rabbit* made cartoon characters the stars of live action films while *The Labyrinth* brought together actors and the animatronic creations of Jim Henson. The latest blend is the entirely virtual character Jar Jar Binks in the new *Star Wars* films. However annoying we find him, it cannot be doubted that the gap between reality and animation is becoming increasingly thin.

3.2 The future

It would also seem that people are shifting their attentions away from TV and film and into gaming; in its opening weekend the game Resident Evil 2 took more money than any film showing except *Titanic*. People are demanding more from their media, they want interactivity — games, out-takes and access to extra information. With advances in avatar creation technologies, for instance the AvatarBT installation at the Millennium Dome, we are able to create low-cost photo-realistic avatars. A popular application of these avatars is as characters in computer games, allowing people to play as themselves or their heroes. Personal avatars are explored further in Ballin et al [8]. Consumers are able to put together components to customise their own game experience.

Technologies like MPEG-4 and projects like Prometheus promise the same sort of customisation for other entertainment experiences, closer to today's cinema.

As actors, scenes and props become components to be arranged at will, the traditional role of the actors, directors and viewers changes and a new broadcast medium is created. This will have a profound effect on the type of experiences that can be delivered to the consumer. There are many ways in which this can be used creatively by 'experience' producers.

Imagine a soap opera in which the consumer can choose to follow any single character from a group scene. In a football experience you can choose to take the perspective of a player or even the ball. With avatar technology the consumer is able to swap different actors to play each part, including models of themselves.

Directors' powers are also enhanced. They are now able to manipulate the performance of their actors as never before. Every muscle movement is within their control, every glance, shrug, to be edited without any intervention from the original performer. Every element of the set is interchangeable, the dress of the actors, the lighting, the camera position, etc. Scenes, props and performances can be integrated from archives into a new consumer experience, at minimal cost. This is already possible in a limited way — for instance, the synthesised meeting of Tom Hanks and John F Kennedy in the film *Forrest Gump* (1994) — but at considerable expense. Furthermore, the director is able to use and edit an old performance with a new actor. The director may choose to produce an interactive consumer experience or a conventional film, from this process.

Objectised media facilitate a broad range of new opportunities for the director and consumer, but there are a number of challenges which remain unresolved.

3.3 Disembodied acting

Considering a scenario where the director or user is able to select any actors (including themselves) as the lead in an action experience, a number of important questions are raised. For this application it is not enough to simply change the actor model. To give a convincing performance as an identifiable actor, the physical model must be a high-quality facsimile, the voice must be a good mimic, the appropriate vocabulary and turns of phrase should be used and the characteristic expressions and mannerisms need to be replayed too. This is the skill of the impersonator. In fact, it is possible for some performers to adopt the persona without changing the way they look, the effect is achieved with voice and action alone. The television mimic Rory Bremner can 'look' like Tony Blair, without looking like Tony Blair.

In order to facilitate this level of customisation, mannerism and presence need to be describable too. From the recording of an original performance with its associated characterisation, a new sequence needs to be synthesised with an alternative characterisation. There are two general methods of computational impersonation we could apply from the speech synthesis field:

- to capture and store a suitable set of atomic elements of the target person's performance and to recombine these into a new performance — this requires a complete performance database for any given actor and sophisticated interpolation between elements,
- using a generic model with a very rich set of parameters to alter the performance — this requires that a generic model be created and a sufficient set of parameters chosen, while avoiding the need for a complete library of performance (assuming mimic parameters can be generated from a set of sample performances).

In both methods a representation which allows the classification of instances of performance is required — in the first, to find corresponding elements to replace, and, in the second, to select the appropriate model for that behaviour. In speech this could be a phonemic representation, while in facial expression this could be a FACS (facial action coding system) [9] or an extension. However, this representation needs to be abstract enough to allow a wide interpretation — for instance 'character A intends to greet character B'. This intention is then expressed depending on the individual characterisations, using different language, facial expression and body posture, dependent on the individual character. It is debatable whether these intentions are derivable from the original performance by capturing raw motion and sound samples.

Assuming that it is possible to derive a high-level description of the pure performance or disembodied performance, we can consider how this can be combined with a characterisation model. A comical illustration of this is the 'Enchefferizer' [10]. The Enchefferizer converts ASCII text into a sentence uttered in the style of the Muppet's Swedish Chef, using a set of transformation rules. These rules can be considered to encapsulate a characterisation of the Swedish Chef (see Table 1). We can imagine an extended set of rules to encode the vocabulary of Tony Blair. The source dialogue needs to be sufficiently abstract to drive any arbitrary character. Again this raises the question as to whether these descriptions can be generated automatically from a sample performance. They would need to describe both the verbal and non-verbal performance.

Table 1 'Enchefferized' text sample.

Source text	Enchefferized text
Colourless green ideas sleep furiously	*Culuoorless greee idees sleep fooreeuoosly*

There is therefore a need for a language to describe a performance, in such a way that it can be impersonated by any actor. A number of current schemes partially address this need.

HumanML [11] is an XML and RDF schema specification, which embeds human characteristics including emotions, intentions, motivations and allusions within conveyed information. Another example is SABLE [12], which is an XML/SGML-based mark-up scheme for text-to-speech synthesis allowing attributes of the speaker and their delivery to be described.

Laban Movement Analysis (LMA) [13] allows human movement, specifically dance, to be transcribed in Labanotation. LMA analyses five components of motion — body, space, shape, effort and relationship. Body considers which parts of the body are moved during the sequence. Space describes how the body moves through space — the locale, directions and paths of motion. Effort looks at how rather than what the person performs, the dynamic quality of the movement. Shape refers to the way in which someone moves in space, which Laban believed reveals a person's inner thoughts, feelings, drives and emotions. Relationship is the interaction between the performer and their environment.

The EMOTE (expressive motion engine) work, at the University of Pennsylvania [14], is an example of a computation model of LMA which allows the animator to alter the effort and shape of a model.

A description of facial expression, independent of the morphology of the target face, has been addressed by Noh and Neumann [15] among others. MPEG-4 also seeks to encode abstract facial animation parameters (FAPs) which can be applied to any arbitrary MPEG-4 face model. The FAPs encode the required displacement of control points on the face which are then expressed in units derived from the dimensions of the target face. The problem has also been considered for body motion by Gleicher [16], where motion capture data is 'retargeted' from one character to another, using their physical features.

If it were possible to capture and represent all the modes of performance, automatically or in an assisted way from a sample performance, it would be plausible to consider computer-based mimicry.

3.4 Dislocated performances

In animation there has always existed a dislocation of the director and the performance, unlike film making. How do you direct a scene when it takes an animator a week to finish a suggestive wink?

Jane Horrocks, an actress who supplied the voice of a chicken in the animated film, *Chicken Run*, made an interesting observation regarding the divorce of her voice from her character. She commented that she would have spoken differently if she had been aware that the character had such prominent teeth.

With current tracking technologies, such as the Prometheus face tracker, a 'face-over' scenario is considered where the voice, body and face capture for a single performance may not be collocated and the director or viewer may wish to change any aspect of it at any time in the future, i.e. the performance is dislocated from the character's embodiment.

Live action with real and virtual characters creates a problem of dislocation too. How do actors cope with acting in empty studios imagining the virtual characters with whom they are interacting? 'It's really difficult', said Ewan McGregor, of Star Wars Episode II: *Attack of the Clones*, commenting that: 'You just have to remember to look up there,' referring to acting with invisible alien co-stars. Humans are very sensitive to inconsistencies in eye gaze.

Augmented reality systems may address some of these problems in the short-term, enabling more believable and sympathetic performances.

There will exist the potential for a viewer or director to change individual features of an actor, giving the possibility to mix and match characteristics. Here there emerges the problem of characteristic interaction. In this case, the embodiment of the actor and their mannerisms become dislocated, so for example the voice can be changed irrespective of the sex of the character. There are inter-dependencies between characteristics of humans which, if they are violated, create unbelievable characters.

There is a need to develop a language in which dependence between features can be described and a set of mechanisms built to resolve conflicts. So in the *Chicken*

Run example, the character's voice could be changed if the mouth was changed.

4. Conclusions

Technologies like those being developed within Prometheus will bring great flexibility to the director and viewer of future entertainment experiences. Features of the identity of a performance (animated or real) are becoming increasingly parameterised and represented in digital electronic form. These features can be decomposed, duplicated, changed, recalled and reassembled.

The challenges we have identified are those of disembodied acting and dislocated performances. In the first, we must develop novel methods of representing and extracting pure or disembodied performances and applying new characterisation to them, while in the second, we must ensure coherence between the interdependent features of a performance even when that action is assembled from multiple discrete components.

References

1 Price M and Thomas G A: '3D virtual production and delivery using MPEG-4', International Broadcasting Convention (IBC 2000), Amsterdam, IEE Conference Publication (2000) — http://www.bbc.co.uk/rd/pubs/papers/pdffiles/ibc00mp.pdf

2 MPEG-4 — http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm

3 VRML: 'International Standard ISO/IEC 14772-1:1997', — http://www.vrml.org/Specifications/VRML97/index.html

4 QuickTime — http://www.apple.com/quicktime/

5 Machin D J: 'Real-time facial motion analysis for virtual teleconferencing', Proc of the Second Int Conf on Automatic Face and Gesture Recognition, IEEE Comput Soc Press, pp 340—344 (October 1996).

6 Mortlock A N, Machin D, McConnell S and Sheppard P J: 'Virtual Conferencing', in Sheppard P J and Walker G R (Eds): 'Telepresence', Kluwer Academic Publishers, Boston, pp 208—226 (1999).

7 Gertie the Dinosaur — http://www.dinosaur.org/Gertie.htm

8 Ballin D et al: 'Personal virtual humans — inhabiting the TalkZone and beyond', BT Technol J, 20, No 1, pp 115—129 (January 2002).

9 Ekman P and Friesen W V: 'Facial action coding system: A technique for the measurement of facial movement', Consulting Psychologists Press, Palo Alto, California (1978).

10 Hagerman J posted the original chef.x program to Usenet in 1993 — an on-line version of the current Enchefferizer can be found at — http://www.cs.utexas.edu/users/jbc/home/chef.html

11 HumanMarkup — http://www.humanmarkup.org/

12 SABLE — http://www1.bell-labs.com/project/tts/sabpap/sabpap.html

13 Laban/Bartenieff Institute of Movement Studies (LIMS) — http://www.limsonline.org/

14 Chi D, Costa M, Zhao L and Badler N: 'The EMOTE model for effort and shape', Proc of SIGGRAPH 2000, ACM Computer Graphics Annual Conference (University of Pennsylvania), pp 173—182 (2000).

15 Noh J Y and Neumann U: 'Expression Cloning', Proc of SIGGRAPH 2001, ACM Computer Graphics Annual Conference, pp 277—288 (2001).

16 Gleicher M: 'Retargetting motion to new characters', Computer Graphics Proc, Annual Conference Series, pp 33—42 (1998).

Jeremy Thorne joined BT in 1999 as a graduate. Prior to this he obtained an MEng in Electrical and Information Sciences at the University of Cambridge.

He works in the Future Content Group in the Content and Coding Laboratory at BTexact Technologies, Adastral Park.

His current research activities involve image-based head modelling and the creation of personalised media through the use of Smart Content.

David Chatting joined BT Laboratories as an apprentice technician in 1993. Upon completing his apprenticeship, he worked in the distributed systems research group. In 1997 he was awarded an employee scholarship to study Computer Science and Software Engineering at the University of Birmingham, obtaining a first class BSc in 2000.

He is currently working in the Future Content Group, within the Content and Coding Laboratory of the Research Department, in BTexact Technologies. The focus of his current work is the interpretation of facial expression.