Speech and Expression Driven Animation of a Video-Realistic Appearance Based Hierarchical Facial Model

David Marshall, Darren Cosker, Paul L. Rosin
School of Computer Science
Cardiff University, U.K.
Dave.Marshall@cs.cardiff.ac.uk

Yulia Hicks
School of Engineering
Cardiff University, U.K.

1. Introduction

We describe a new facial animation system based on a hierarchy of morphable sub-facial appearance models. The innovation in our approach is that through the hierarchical model, parametric control is available for the animation of multiple sub-facial areas. We animate these areas automatically *both* from speech - to produce lip-synching, and natural pauses and hesitations - *and* using specific temporal variation of appearance parameters to control sub-facial behaviours.

Image-based and morphable models are capable of producing highly-realistic facial animations, but currently only provide parametric control of limited facial areas [3]. We are able to produce realism comparable to that of real video, while also providing parameters for animation usually only associated with 3D parametric models.

Speech driven animation is facilitated using a Hidden Markov Coarticulation Model (HMCM). The model learns visual speech relationships from a training corpus of 2D video, and creates new visual parameters given new speech – which may not have appeared in the original corpus.

For control over non-speech related animation and also for greater control over certain facial expressions in general, parameters responsible for controlling the sub facial movement of the mouth, eyes and eyebrows are extracted from the hierarchy and specific poses identified. These are interpolated to create sub-facial animations for behaviours such as smiling, blinking and winking, and expressions akin to anger, fear, shock and surprise.

2. The Hierarchical Modelling of Facial Shape, Appearance and Dynamics

A training corpus of video and speech has been collected by recording a participant speaking front-on into a standard digital video camera. The video images are automatically land-marked around key-facial areas. Sub-sets of the facial landmarks are used to extract sub-facial images and build appearance models for these areas [1] - thus creating the hierarchy. By decomposing the face in this way (see Figure 1), the major appearance modes of each sub-facial model encode the major variation for that area. For example, the highest mode of variation in our left-eye model encodes a blink. Variation of the appearance parameter value for this mode produces an animation. Figure 1 demonstrates example mode/animation parameters in our model. Facial animation is achieved by creating new sub-facial animations using appearance trajectories, and then merging them in a top-down manner. In order to avoid artifacts in rendering sub-facial components we project back up through the hierarchical PCA model. The combined face is then warped onto background images to increase realism.

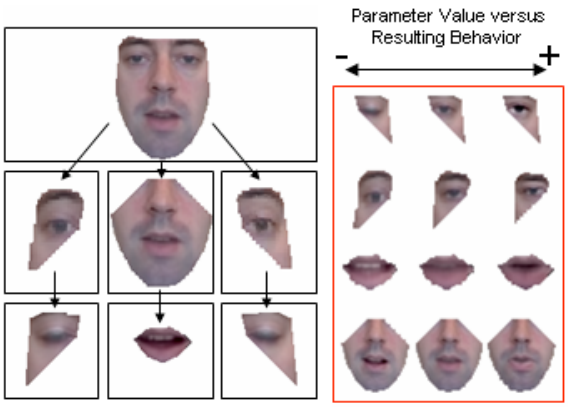

Figure 1. (From left to right)A Hierarchy of sub-facial appearance models. The result of varying animation parameters for a selection of sub-facial areas.

2.1. Dynamics: Appearance Parameter and Speech-Driven Animation

Animations of expressions may be produced by selecting sub-facial appearance parameter values at discrete sampled moments, and interpolating their values *in-between*. The resulting trajectory in parameter space is used to synthesize a corresponding sub-facial animation (see Figure 2). A

complete set of facial parameters provides a pow
for an animator to produce a wide-range of ima
facial animations. Our parameterization also allo
learn *natural animation parameters* from the trai
pus, produced during different facial poses and exp
Analysis of these *real* trajectories gives an insight
our *synthetic* trajectories should behave.

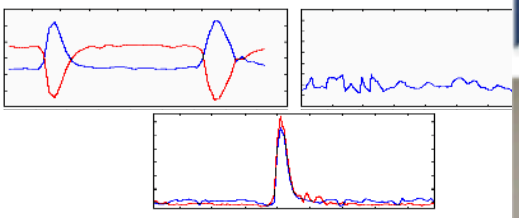

Figure 2. (Top Left) Eye parameter trajectories during t
(Top Right) A mouth trajectory during a smile (Botto
eyebrow trajectories. Blue = Left eye, Red = Right eye.

We animate the mouth from speech using a
trained using our video and speech corpus [2]. This au
tomatically creates mouth parameter trajectories given new speech. The model makes no assumptions concerning the content of the speech, and therefore encodes, and reproduces, natural pauses and hesitations. Essentially, we extract out mel-Cepstral and LPC speech features which we correlate with our video appearance parameters via the HMCM. We do not perform any phonetic transcription of the input speech. This creates a state transition sequence between video and speech features at each frame in the sequence. To resynthesise new video from input unseen audio. We extract out similar audio features and project into the HMCM using the Viterbi Algorithm to assemble to most likely state sequence from the observed training data.

3. Results and Conclusion

Figure 3 shows example frames from two animations. Source videos may be found at *http://users.cs.cf.ac.uk/D.P.Cosker/Videos*. A wide range of facial poses and expressions are possible using simple control of a small number of appearance parameters. Animation from speech also gives equally satisfying results. This has been achieved using a novel hierarchical representation of an audio-visual appearance based talking head model. The hierarchy allows for simple control (via a small number of parameters) of complex expressions of the head. In order to build a final convincing video-realistic rendering care must be taken in combining the sub-facial hierarchy to avoid artifacts — we have developed a novel back projection through our hierarchical model to accomplish this.

4. Future Work

We are currently investigating more complex hierarchies, with aims to create direct mappings to existing 3D parametric models for performance driven animation using our naturally learned animation parameters.

Figure 3. Example frames from two animations (Top) Unseen speech input (Middle) Artificial blink inserted. (Bottom) Temporal control of a single "Smile" parameter

We are also interested in analysing the behaviour of new natural parameters, both for complete facial expressions and individual sub-facial regions. Modelling the interaction between individual regions may be required in some cases in order to account for the dependence between different parts of the face. By tracking the natural parameter dynamics of different facial areas we can attempt to further understand the role of different facial areas in conveying e.g. fake and genuine emotions. We have already demonstrated the usefulness of such an analysis in our work on smiles.

Recent acquisition of a real time 3D scanner will enable us to apply our techniques using 3D facial appearance models. The appearance models in 2D do not work well with any out of plane head rotations, as this clearly distorts to underlying statistical model. This would allow us to re-write performances with our synthetic animations, by estimating head pose orientation in videos and projecting in our model. Using such a device we can also begin to learn the 3D dynamics of expressions.

References

[1] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. PAMI*, 23(6):681 – 684, 2001. 1

[2] D. Cosker, D. Marshall, P. L. Rosin, and Y. Hicks. Speech driven facial animation using a hidden markov coarticulation model. In *Proc. IEEE ICPR*, pages 128 – 131, 2004. 2

[3] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proc. of Computer Graphics and Interactive Techniques*, pages 388–398. ACM Press, 2002. 1