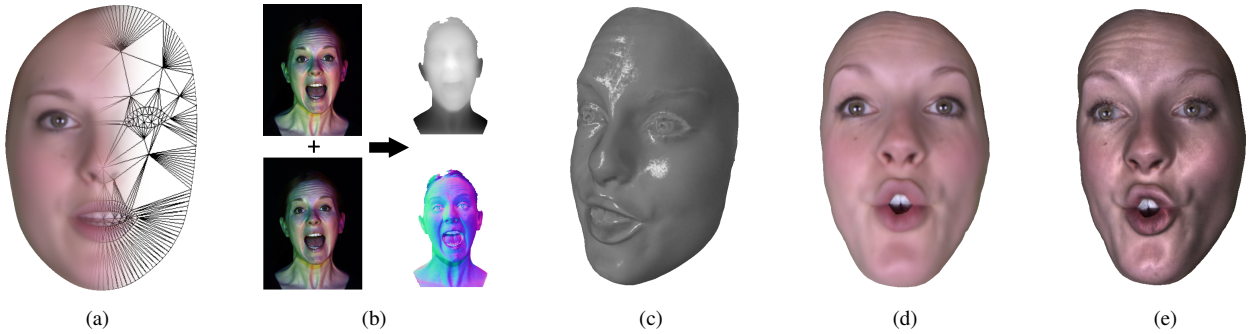
An Expressive Text-Driven 3D Talking Head

Robert Anderson[1], Björn Stenger[2], Vincent Wan[2], Roberto Cipolla[1]
[1]Department of Engineering, University of Cambridge, UK *
[2]Toshiba Research Europe, Cambridge, UK †



(a) (b) (c) (d) (e)

Figure 1: *An Active Appearance Model (AAM) based, expressive 2D talking head is trained on 7000 sentences of video data (a). Given only 30 3D scans obtained by combined multiview and photometric stereo we obtain a set of training samples with both depth and normal maps (b). Depth and normals are mapped to the existing AAM, allowing the same synthesis pipeline used for the 2D talking head to drive renderings of either geometry (c), texture (d) or a combination of the two (e).*

1 Introduction

Creating a realistic talking head, which given an arbitrary text as input generates a realistic looking face speaking the text, has been a long standing research challenge. Talking heads which cannot express emotion have been made to look very realistic by using concatenative approaches [Wang et al. 2011], however allowing the head to express emotion creates a much more challenging problem and model based approaches have shown promise in this area. While 2D talking heads currently look more realistic than their 3D counterparts, they are limited both in the range of poses they can express and in the lighting conditions that they can be rendered under. Previous attempts to produce videorealistic 3D expressive talking heads [Cao et al. 2005] have produced encouraging results but not yet achieved the level of realism of their 2D counterparts.

One challenge in building 3D talking heads is the collection of sufficiently large training datasets. While 2D systems can be trained from thousands of example sentences from video, capturing the same amount of 3D data at high quality is currently still time consuming and challenging. We propose a method for creating an expressive 3D talking head by leveraging a large amount of 2D training data and a small amount of 3D data.

2 Technical Approach

Given a training set of 7000 recorded sample sentences, we first construct a 2D talking head using an Active Appearance Model (AAM) as a representation of the face. Synthesis is performed with a Hidden Markov Model text-to-speech (HMM-TTS) system that uses cluster adaptive training (CAT) as described in [Anderson et al. 2013]. This allows for synthesis over a continuous space of expressions created by setting the weights of five basic emotions (happiness, sadness, fear, anger and tenderness). Given input text an audio file is generated along with a set of AAM parameters that allow for rendering of the face. As additional training data we recorded 5 sentences in each emotion using a 3D scanner, from which we used 30 frames covering a variety of expressions and phonemes. The 3D data is captured using a system consisting of two video cameras and three colored light sources which allows for a combination of multiview stereo and color photometric stereo to produce high quality geometry of the face region at video framerates.

In order to avoid retraining the CAT speech synthesis model we add data from the 3D scans to the existing AAM. This allows the 3D talking head to be driven in exactly the same way as the 2D one, from the same AAM parameters. To model coarse geometry we add a depth component to each vertex in the AAM, essentially turning it into a morphable model. To model fine geometry we add a normal map component, which is dealt with in the same way as the texture component of a standard AAM, except that an additional normalization step is used. To find the depth and normal map modes which correspond to the original 2D AAM, it is first registered to the 30 3D training scans. Depth and normal maps are then extracted from these training examples and the modes are found by minimizing the difference between the observed samples and reconstructions synthesized by the AAM in a least squares sense.

The 3D scanning method used does not accurately reconstruct the teeth or the inside of the mouth and so in future work we would like to better model these. The 2D position of the teeth is given by the original AAM and we aim to drive a rigid 3D teeth model from this.

*e-mail: {ra312, rc10001}@cam.ac.uk
†e-mail: {bjorn.stenger, vincent.wan}@crl.toshiba.co.uk

References

ANDERSON, R., STENGER, B., WAN, V., AND CIPOLLA, R. 2013. Expressive visual text-to-speech using active appearance models. In *CVPR*.

CAO, Y., TIEN, W., FALOUTSOS, P., AND PIGHIN, F. 2005. Expressive speech-driven facial animation. *ACM TOG 24*, 4, 1283–1302.

WANG, L., HAN, W., SOONG, F., AND HUO, Q. 2011. Text driven 3D photo-realistic talking head. In *Interspeech*, 3307–3308.