Speech independent emotion transfer for virtual faces

Timothy Costigan
Rachel McDonnell
Trinity College Dublin

Figure 1: We capture a number of clips, isolate their emotion components and use them to learn an emotion transfer system.

1 Introduction

Animating a virtual face is a difficult and expensive undertaking. Whether using a motion capture system or hand animating, it will likely require considerable skill and perhaps the services of a good actor. It is therefore important that mistakes are minimised to avoid having to recapture large sequences. Methods do exist however, that can alter existing clips to take on new properties such as emotion.

Examples include the component weighting method of Shaw et al. [2013] where the motion is split into independent components (IC) which are then compared with a neutral example to derive an emotion weighting. However, this method also does not truly model the variation in expressions across a clip as it simply applies a weighting to each component.

Cao et al.’s [2003] method makes a more explicit distinction between speech and emotion components, allowing the emotional components to be replaced while leaving the speech intact. As this method uses actual emotion components to transfer the emotion, it can take into consideration varying expression intensity. The emotion components however, do need to be from a clip with similar speech content. A system like Ma et al.’s [2009] which uses Gaussian Processes (GP) to model ICs could help break the dependence on speech for emotion transfer. In its current form however, it requires extensive edits by the artist per clip.

2 Our Approach

We aim to produce an automatic system capable of transferring emotion to previously unseen facial motions. To achieve our goal, we combined aspects of the methods of Cao et al. and Ma et al. to create a system that separates motions into speech and emotion components and then models them using GPs. For data, we captured 3 volunteers: one male, two females using FaceShift. We selected 7 sentences to be performed under each basic emotion (neutral, sad, happy, angry, fear, surprise and disgust), taken from the SA VEE, Enterface and Boaz datasets [Haq and Jackson 2010; Martin et al. 2006; Ben-David et al. 2013]. The captured data was then broken up into training (70%) and testing datasets (30%).

For preprocessing, the data was first standardised and then we used principal components analysis PCA (variance 99%) to reduce dimensionality and independent components analysis ICA to separate the motion into components. We aligned all the clips to a single neutral example of each sentence using dynamic time warping (DTW), applied smoothing to remove any artefacts and then compared them using the Kruskal-Wallis H test. In conjunction with visual inspection, the components were then categorised as either speech or emotion components based on their H value. H values above the 70th percentile were found to give good results without interfering with speech. This threshold can however, be lowered to increase expressiveness at the risk of decreased speech motion accuracy.

Finally, the training dataset was separated using 3-fold cross validation and used to train a GP model for each emotion component. We trained a different system per actor. Results showed that our system is capable of applying and transferring between emotions while retaining the same speech content as can be seen in figure 2. The raw motion can be noisy, so additional smoothing was applied to the altered components before reconstructing the facial motion.

Acknowledgements: This work was sponsored by Science Foundation Ireland as part of the “Cartoon Motion” project.

References


