Facial Performance Transfer via Deformable Models and Parametric Correspondence

Akshay Asthana, Miles Delahunty, Abhinav Dhall and Roland Goecke

Supplementary Material

1 Active Appearance Models (AAM)

In recent years, model-based approaches have gained momentum as researchers have realised the potential benefits of a model that can replicate deformations of a non-rigid object. In the original work of [1], the model was learnt by applying Principal Component Analysis (PCA) to the set of labelled data in order to model the intrinsic variation in shape and texture of deformable objects. As a result, a parameterised model is formed that is capable of representing large variation in shape and texture by a small set of parameters.

For constructing the AAM, each training image is represented by a $2n$-dimensional shape vector $\mathbf{S} = [x_1, y_1, \ldots, x_n, y_n]^T$, where n is the number of landmark points describing the object's shape. These shapes are aligned into a common coordinate frame by Procrustes analysis. The modes of shape variation are obtained by applying PCA to the set of aligned shapes. Similarly, the texture from each training image is warped to a canonical shape (i.e. the mean shape in this case) and is represented by the $3m$-dimensional texture vector $\mathbf{T} = [r_1, g_1, b_1, \ldots, r_m, g_m, b_m]^T$, where m is the number of pixels describing the canonical shape. The modes of texture variation are modelled by applying PCA to a set of these texture vectors. The parametrisation of shape and texture variations (Figure 1) can be written as

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{\Phi}_s \mathbf{s} \qquad \mathbf{T} = \bar{\mathbf{T}} + \mathbf{\Phi}_t \mathbf{t} \tag{1}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean shape and texture vectors, respectively, $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_t$ are the shape and texture basis matrices, respectively, and $\mathbf{s}$ and $\mathbf{t}$ are the non-rigid *local* shape and texture parameters, respectively.

(a) (b) $\bar{\mathbf{S}}$ (c) $\bar{\mathbf{T}}$ (d) $\mathbf{S}_1$ (e) $\mathbf{T}_1$

Fig. 1. (a) Landmark Points, (b) (c) Mean Shape and Texture, (d) (e) Sample new shape and texture. Texture shown is warped from the mean shape to the new shape.

If we wish to fit the AAM to a face in a new, unseen image $\mathcal{I}$, we need to find the set of model parameters $\mathcal{P} = \{\mathbf{s}_g, \mathbf{s}_l, \mathbf{t}_g, \mathbf{t}_l\}$ that best fits the model to $\mathcal{I}$. Here, $\mathbf{s}_g$ and $\mathbf{s}_l$ represent the global and local shape parameters, $\mathbf{t}_g$ and $\mathbf{t}_l$ represent the global and local texture parameters. This model fitting process is performed by iteratively updating the model parameters $\mathcal{P}$. A number of AAM fitting algorithms have been proposed and almost all of them adhere to one or more basic principle, namely accuracy, efficiency, robustness, generalisability and applicability [2].

The global shape and texture parameters represent the *scaling, rotation and position* of the face, whereas the local shape and texture parameters represent the non-rigid variations or local changes in pixel intensity in the normalised frame. This work focuses on transferring the local changes occurring in the shape and texture from one model to another. We deliberately do not transfer global head movements, e.g. nodding, in the work presented here to precisely study the effects of the different methods of transferring the local shape and texture parameters. For simplicity, we refer to the local shape and local texture parameters as the shape and texture parameters, respectively, in the remainder of the paper. Note that the AAMs used for the experiments presented in this paper were trained on images annotated with 75 landmark points each, with an average face cropped area of 130×140 pixels.

2 AV Recording of subject *B1*

In order to demonstrate our *Cross-Language Facial Performance Transfer* framework, we separately recorded a video of a female subject of *Indian* origin (Figure 2) repeating the sequences present in AVOZES along with some more complex (and new) sequences in both English and *Hindi* (National language of India). This subject is referred as ***B1*** in this paper.

Fig. 2. Subject ***B1*** - Female Subject of Indian Origin

Refer to Section 3.3 for the videos of ***B1***. Details of the relevant video sequences, used for experiments in this paper, are as follows :

2.1 B1 *Module 6*

This module contains three *4 second* videos (15fps) of continuous speech sequences spoken by subject ***B1*** :

- "*Joe took fathers green shoe bench out.*"
- "*Thin hair of azure colour is pointless.*"
- "*Yesterday morning on my tour, I heard wolves here.*"

These sequences are exactly the same as present in AVOZES *Module 6* [3].

2.2 B1 *Hindi*

This module contains a *19 second* video (15fps) of continuous and complex speech sequence spoken by subject ***B1*** in **Hindi** (National language of India). Refer to Figure 3 for the script of this sequence.

हिन्दू	धर्म	के	उनुसार	,	दिवाली	का	उत्सव	भगवान्	राम	के
HINDU	DHARM	KE	ANUSAR	,	DIWALI	KA	UTSAAV	BHAGWAN	RAM	KE
वनवास	से									
VANVAAS	SE									
वापस	घर	लौटने	की	ख़ुशी	मैं	,	दिए	जला	कर	मनाया
WAPAS	GHAR	LAUTNE	KI	KHUSHI	MEIN	,	DIYE	JALA	KAR	MANAYA
गया	था !									
GAYA	THA .									
आज	भी	ये	दिए	जलने	की	परंपरा	कायम	है	,	जो
AAJ	BHI	YEH	DIYE	JALANE	KI	PARAMPARA	KAYAM	HAI	,	JO
बीते	हुए									
BEETE	HUE									
साल	की	ख़ुशी	को	सम्भोदित	करती	है	और	आने	वाले	सालों
SAAL	KI	KHUSHI	KO	SAMBHODIT	KARTI	HAI	AUR	AANE	WALE	SAALON
के	लिए	शुभकामनाएं	देती	है !						
KE	LIYE	SHUBHKAMNAYEIN	DETI	HAI .						

Fig. 3. Complete Script of the B1 *Hindi* Module

3 Facial Performance Transfer Result Videos

Refer to the supplementary videos[1,2] submitted along with this paper for the complete set of video results.

3.1 Facial Performance Transfer from *A1* to *A2*

In Videos/A1toA2/Validation/ : AVOZES *Module 6* sequences synthesised for validating the proposed approach.

In Videos/A1toA2/Testing/ : AVOZES *Module 5* sequences synthesised for testing the proposed approach.

3.2 Facial Performance Transfer from *A1* to *A3*

In Videos/A1toA3/Validation/ : AVOZES *Module 6* sequences synthesised for validating the proposed approach.

In Videos/A1toA3/Testing/ : AVOZES *Module 5* sequences synthesised for testing the proposed approach.

1. Videos:http://users.rsise.anu.edu.au/∼aasthana/TVCG11/Supplement.tar
2. Document:http://users.cecs.anu.edu.au/∼aasthana/TVCG11/ReadMe.pdf

3.3 Cross-Language Facial Performance Transfer

In Videos/B1toA1/B1_Dataset/ : Set of relevant videos for subject ***B1***. Refer to Section 2.

In Videos/B1toA1/Final_Result/ : Facial Performance Transfer from ***B1*** to ***A1***. B1 *Hindi* sequence synthesised for testing the Cross-Language Facial Performance Transfer framework.

3.4 Original AVOZES *Module 6* Videos

In Videos/Original_Videos/ : Original AVOZES *Module 6* sequences used for training and validation.

References

[1] G. Edwards, C. Taylor, and T. Cootes., "Interpreting Face Images Using Active Appearance Models," in *Proc. of IEEE Int. Conf. Automatic Face and Gesture Recognition FG'98*, 1998, pp. 300–305. 1

[2] S. Romdhani, "Face Image Analysis using a Multiple Feature Fitting Strategy," Ph.D. dissertation, University of Basel, Switzerland, 2005. 1

[3] R. Goecke and B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES," in *Proc. Int. Conf. Spoken Language Processing ICSLP2004*, 2004, pp. 2525–2528. 2