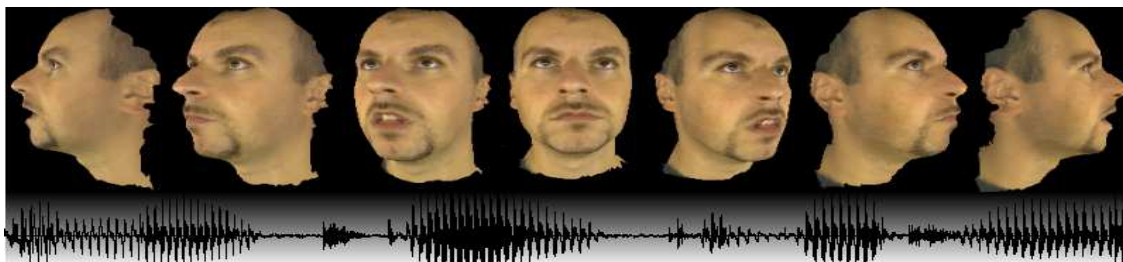


Speech-driven Face Synthesis from 3D Video

Ioannis A. Ypsilos, Adrian Hilton, Aseel Turkmani and Philip J. B. Jackson
Centre for Vision Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
i.ypsilos,a.hilton,a.turkmani,p.jackson@surrey.ac.uk



Abstract

This paper presents a framework for speech-driven synthesis of real faces from a corpus of 3D video of a person speaking. Video-rate capture of dynamic 3D face shape and colour appearance provides the basis for a visual speech synthesis model. A displacement map representation combines face shape and colour into a 3D video. This representation is used to efficiently register and integrate shape and colour information captured from multiple views. To allow visual speech synthesis viseme primitives are identified from the corpus using automatic speech recognition. A novel non-rigid alignment algorithm is introduced to estimate dense correspondence between 3D face shape and appearance for different visemes. The registered displacement map representation together with a novel optical flow optimisation using both shape and colour, enables accurate and efficient non-rigid alignment. Face synthesis from speech is performed by concatenation of the corresponding viseme sequence using the non-rigid correspondence to reproduce both 3D face shape and colour appearance. Concatenative synthesis reproduces both viseme timing and co-articulation. Face capture and synthesis has been performed for a database of 51 people. Results demonstrate synthesis of 3D visual speech animation with a quality comparable to the captured video of a person.

1. Introduction

Realistic face synthesis is a challenging problem due to the sensitivity of viewers to subtle discrepancies between the dynamic behaviour of real and synthetic faces. Due to the importance of this problem for remote communication

and entertainment production, face synthesis has received considerable research interest in both computer graphics and vision. However, automatic production of convincing animated face models of real people remains an open problem. In film production realistic animated faces have been achieved by highly skilled manual animation. Recent results in computer vision and graphics have also demonstrated realistic reproduction of realistic animated faces of individual people by concatenative synthesis from 2D video segments [5, 9, 6]. However, the concatenation of 2D video limits this approach to a 2D typically frontal view with un-naturally restricted head rotation.

In this research we aim to achieve the same visual quality as previous concatenative face synthesis approaches without the restrictions inherent in the use of 2D video. A novel video-rate shape capture technology [19] is used to simultaneously acquire 3D face shape and colour appearance. This provides the basis for reproduction and synthesis of 3D face shape and colour with a visual quality comparable to the captured video. In this paper we introduce a framework for synthesis of novel face sequences from speech. The principal contributions of this paper are: (1) a novel method for 3D face synthesis from a speech viseme sequence; and (2) a novel optical flow formulation using both shape and colour to estimate dense face correspondence. The introduction of a face synthesis framework based on 3D shape capture enables accurate reproduction of facial dynamics, natural head movement and arbitrary viewing direction.

1.1. Previous Work

Passive face reconstruction techniques have been developed to reconstruct animated models of the face from images [3]

and video [10, 15]. Blanz and Vetter [3] used learnt statistical models of 3D face shape and appearance to reconstruct photo-realistic 3D face models from a single image. Other face reconstruction methods [10, 15] use model-based bundle adjustment techniques to reconstruct realistic static face models from image sequences. Currently these approaches are limited to reconstruction of static face shape for a single pose.

Video based approaches to facial animation have been introduced [5, 6, 9, 4] which avoid the requirement for facial reconstruction by resampling captured video sequences of a person to produce novel image sequences. Such approaches achieve photo-realistic synthesis of facial appearance but are currently restricted to the viewpoint and lighting of the captured video data. This limits their use for both entertainment and communication applications.

Extensive research in face tracking from video [7] has demonstrated automatic tracking using optic flow together with sparse facial features such as lips and eyes. However, video-based tracking does not achieve accurate reconstruction of the detailed surface deformation required for convincing animation. Phigin et al.[13] used marker based capture to accurately reconstruct the movement of a sparse set of 3D points on the face surface. The marker movement was then used to morph a face model reproducing the coarse facial dynamics. To overcome the limitation of sparse reconstruction Kalberer and Van Gool [11] combine markers with active structured light capture to acquire dense facial shape deformation for animation. This shape deformation is then mapped onto a generic face model using the known marker point correspondence to produce high-quality facial animation. Due to the use of visible markers and structured light previous approaches do not allow simultaneous capture of facial appearance.

The video-rate capture system [19] used in this paper uses infra red structured light to allow simultaneous capture of shape and appearance. Automatic facial alignment is used to avoid the requirement for known marker correspondence. This results in a representation of real facial dynamics which reproduces both the 3D shape and colour appearance with a visual quality comparable to the original video.

2. Video-rate Shape & Appearance

The visual face synthesis framework introduced in this paper is based on a corpus of captured dynamic 3D colour face sequences of a person. Simultaneous video-rate capture of face shape and colour appearance is achieved using a novel capture system [19]. In this section we provide a brief description of both the capture system and displacement map representation of the face data as a 3D video used in subsequent processing for face synthesis.

2.1. Face Capture System

Video-rate simultaneous capture of face shape and colour appearance is achieved using a multiple camera system with infra-red (IR) stereo capture for shape reconstruction and visible capture with uniform white-light illumination for colour appearance. Figure 1 shows a schematic of the capture system. The system comprises three units each capturing shape and colour at video-rate designed to give ear-to-ear face coverage. Each unit comprises a colour camera, IR pattern projector and two IR cameras.

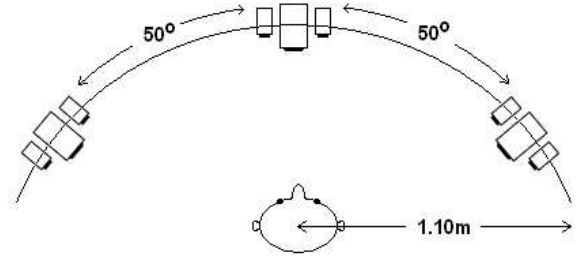


Figure 1: Schematic of the capture equipment.

The system operates at full CCIR601 PAL resolution 25Hz progressive scan using standard PAL video capture cards. For each pair of IR cameras we estimate stereo correspondence and reconstruct the visible surface shape. An IR speckle pattern is projected onto the face surface to allow shape reconstruction in areas of uniform visual appearance. Stereo correspondence between each pair of IR cameras results in three sets of surface measurements captured at video-rate. All cameras are synchronised and time stamped to ensure temporal alignment. Synchronised digital audio is simultaneously captured at 48 KHz with 16 bits per sample.

2.2. 3D Video Face Representation

The capture system acquires three sets of facial surface measurements with corresponding colour images at video-rate. To integrate these data into a single 3D face representation we use an ellipsoidal displacement mapping to combine both shape and colour from multiple views. This represents both shape and colour as a 2D image sequence of colour and displacement, referred to as ‘3D Video’.

An ellipsoid mapping is defined by the location of its centre, \vec{c} together with the magnitude and direction of the three orthogonal axis vectors, $\vec{v}_1, \vec{v}_2, \vec{v}_3$ giving 9 degrees-of-freedom (6 pose and 3 radii). At each frame t we capture a set of N_t face surface measurement points $X(t) = \{\vec{x}_i^t\}_{i=1}^{N_t}$ where $\vec{x} = (x, y, z) \in \mathbb{R}^3$. A sampled displacement map image representation, $D_t(x, y)$ for $x = 1 \dots N_x, y = 1 \dots N_y$, of the surface shape for each frame is then computed by sampling the distance to captured data along equi-angular rays from the ellipsoid centre as shown in Figure 2. $N_x,$

N_y are the numbers of columns and rows of the displacement image which define the sampling resolution. The 3D surface can then be reconstructed from a displacement image $D_t(x, y)$ by sampling the distance from the ellipsoid centre \vec{c}_t to the measured data along each ray $v(x, y)$ such that the intersection \vec{x} with the measured surface is given by $\vec{x} = \vec{c}_t + D_t(x, y)v(x, y)$.

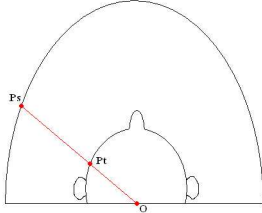


Figure 2: Mapping the geometry of a face to an ellipsoid.

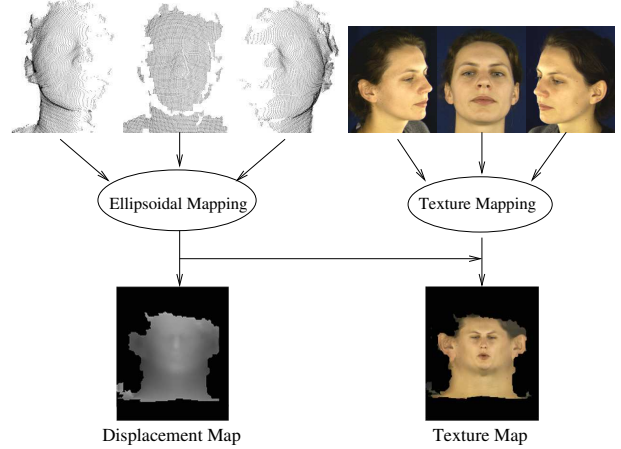
The displacement image representation provides a simple mechanism for integration of overlapping surface measurements by combining their corresponding displacement values:

$$D_t(x, y) = \sum_{p=1}^3 w_t^p(x, y) D_t^p(x, y) \quad (1)$$

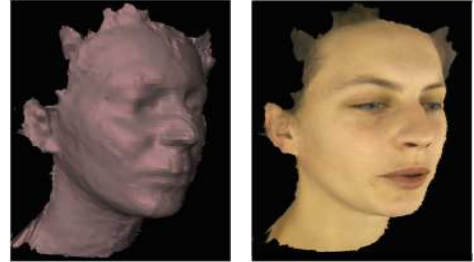
where $w_t^p(x, y) = 0$ if no surface measurement is obtained along $\vec{v}_t(x, y)$ for the p^{th} stereo pair, otherwise $w_t^p(x, y)$ is proportional to the stereo correlation value at (x, y) as a measure of surface confidence with $\sum w_t^p(x, y) = 1$. This provides a computationally efficient method for fusion of the face surface measurements into a single face surface representation $D_t(x, y)$. The integrated face displacement map can be rendered as a single triangulated mesh M_t by triangulating adjacent pixels in the displacement image $D_t(x, y)$ and re-projecting along the corresponding rays $v(x, y)$ to obtain the vertex positions: $\vec{x}_{xy} = \vec{c}_t + D_t(x, y)v(x, y)$.

The second problem that has to be addressed is the integration of colour images into a single texture map. The elliptical mapping also provides an efficient mechanism for integration of the three colour images into a single texture map. The registration of the three colour images with the reconstructed surface shape is known from the colour calibration giving overlapping texture maps from each image $I^p(x, y)$. The problem is then to integrate the texture maps in the overlapping regions. Measured 3D surface shape is used to obtain the correspondence between the observed colour images and resample as an elliptical texture map. Initially we correct for differences in camera colour response by fitting a linear model of the gain and offset between overlapping regions of the texture images relative to the central camera. Given accurate registration and correction for

colour response a simple colour blending[12] of overlapping pixels is then performed. Figure 3 shows results of the shape and colour integration to reconstruct a single 3D colour face model at each time. The sequence of colour and displacement map images provides a 3D video representation of the face.



(a) The displacement and texture map generation.



3D Shape Textured 3D Face

(b) The reconstructed 3D shape and appearance.

Figure 3: Combining colour and shape to produce a 3D video of the face.

3. Temporal Processing

The integrated 3D video of captured face shape and appearance provides an efficient representation for rigid registration, temporal filtering and non-rigid alignment of face sequences. The operations are required to characterise the facial dynamics of a person for synthesis.

3.1. Rigid Registration

Rigid registration of the temporal face sequence is used to remove head movement to obtain an aligned sequence of displacement map and colour images representing non-rigid face deformation such as mouth opening/closing, expressions and eye-movement. This facilitates subsequent analysis and modelling on non-rigid facial dynamics. Iterative closest point (ICP) [2] has been widely used for registration

of rigid point sets and surfaces. In this work we use the ICP algorithm to register the rigid (upper-half) of the face over time. The displacement map representation is used to approximate the nearest-point computation as a single lookup operation enabling registration in constant time. ICP using the displacement map distance is used recursively to estimate the change in head pose between successive frames $T_{t-1,t}$ in order to minimise the registration error between $D_{t-1}(x, y)$ and $T_{t-1,t}(D_t(x, y))$. An estimate of the transformation between the initial pose and the t^{th} frame can then be obtained as the product of intermediate transformations $\tilde{T}_{0,t} = \prod_{s=1}^t T_{s-1,s}$. The initial estimate is refined by direct ICP registration between distance functions $D_0(x, y)$ and $T_{0,t}(D_t(x, y))$ to eliminate propagation of errors. The estimated pose $T_{0,t}$ is used to define the ellipsoid parameters at time t as $(\tilde{c}^t = T_{0,t}\tilde{c}^0, \{\tilde{v}_i^t = T_{0,t}\tilde{v}_i^0\}_{i=1}^3)$, giving a registered displacement image $D_t(x, y)$. In the registration process we consider only the top half of the face since the lower face contains the non-rigid deformations of the mouth and jaw. Registration converges to an RMS error of approximately 0.5mm which is due to measurement error and residual non-rigid deformation. This registration forms the basis for subsequent temporal analysis of non-rigid surface shape change.

3.2. Temporal Filtering

The displacement map representation together with temporal registration provides an image sequence of surface shape and appearance over time. This representation can then be used to analyse the spatio-temporal characteristics of face shape. We use this representation for noise removal, hole-filling and non-rigid alignment of the mouth and lower-face area. Temporal averaging is used to improve estimates for face shape from individual frames. As the surfaces are registered we ensure that corresponding points are averaged over time using a spatio-temporal window of $n \times m \times T$ where n, m and T are spatial and temporal resolution respectively. Spatio-temporal smoothing is performed by convolution with a Gaussian filter. Missing data (such as occluded areas under the chin) are then filled by fitting a smooth 3D surface using a moving least squares approach as described in [17]. This hole-filling strategy has the advantage that the reconstructed patches smoothly blend into the original surface. Results show that temporal smoothing and hole-filling reduces measurement noise in individual frames without significant loss of spatial resolution.

3.3. Non-rigid Alignment

In this section we introduce a non-rigid alignment technique that computes a dense 2D vector field F_{pq} which aligns displacement map D_p to D_q and texture map T_p to T_q captured at times t_p and t_q respectively. Optical flow estimation [1]

has been widely employed to estimate a dense flow field for colour image sequences. In this work we introduce an optical flow algorithm which uses both shape and colour information to estimate a vector field for non-rigid alignment. In non-rigid alignment it is important to obtain accurate correspondence of both visible features such as the edge of the lips and shape features such as regions of high-curvature such as the lips. We therefore introduce an error function which ensures alignment of regions with prominent appearance or shape variation and obtains a smooth alignment in uniform regions. Shape and colour variation are normalised to ensure equal influence on the estimated flow field.

Initially two vector fields are computed, one from shape F_{pq}^S and one from texture F_{pq}^T to produce the motion vector field that deforms frame p to frame q as:

$$F_{pq} = G(\lambda_1 F_{pq}^S + \lambda_2 F_{pq}^T) \quad (2)$$

where λ_r are weights with $\lambda_1 + \lambda_2 = 1$ and $\lambda_r \in [0, 1]$ and G is an averaging 2D Gaussian function which smooths the concatenation of the two vector fields to obtain a dense vector field. To obtain accurate correspondence in either shape or colour we require a high local variation of the gradient in two orthogonal directions. Regions with smooth or linear variation are ambiguous and likely to produce inaccurate correspondences. To perform matching for the optical flow computation we use a *sum of squared differences* (SSD) dissimilarity measure in a local $n \times m$ window around each pixel (typically 10×10). Therefore, to identify and emphasise local regions of high variation which are suitable for matching we analyse the distribution of image gradients within each $n \times m$ window. We use the sum of the horizontal and vertical image gradients as a metric of local variation and the grey-scale eccentricity, to define a matching quality function as:

$$W(x, y) = \frac{\sum_x \sum_y |\nabla f(x, y)|}{\eta} \quad (3)$$

where $\nabla f(x, y)$ is the image gradient at pixel (x, y) and η the grey-scale eccentricity within the window. Grey-scale eccentricity provides a local statistical measure of the variation in image gradients for orthogonal directions as the ratio of the principal to the minor axis of image gradient variation[16]. This is equivalent to the ratio of principal eigenvalues used in the Harris corner detector. We penalise regions with high eccentricity since their topology is likely to extend in nearby regions. Eccentricity can be estimated using the method of statistical moments as:

$$\eta = \frac{\mu_{20} + \mu_{02} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}} \quad (4)$$

where μ_{ij} is the (i, j) 2D grey-scale central moment which

is defined as:

$$\mu_{ij} = \sum_x \sum_y (x \cdot |\nabla f(x, y)| - \bar{x})^i \cdot (y \cdot |\nabla f(x, y)| - \bar{y})^j \quad (5)$$

where

$$\bar{x} = \frac{1}{n} \cdot \sum_x \sum_y x \cdot |\nabla f(x, y)|$$

$$\bar{y} = \frac{1}{m} \cdot \sum_x \sum_y y \cdot |\nabla f(x, y)|$$

and $f(x, y)$ is the grey-scale image intensity at pixel location (x, y) .

A binary threshold is then used to select 5% of local regions with the highest W as candidates for matching. This process results in the identification of image regions based on local shape or colour variation which are suitable for accurate correspondence and flow field estimation. Figure 4 illustrates the regions identified as having sufficient local variation for accurate flow field estimation for the lower regions of the face. Features are clearly identified in the lips, nose and chin areas. The combined shape and colour flow field is estimated from local region correspondence using SSD according to equation 2.

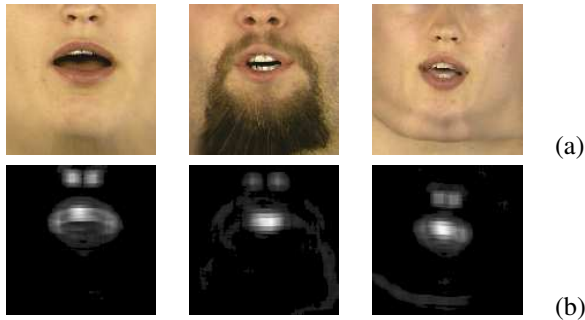


Figure 4: (a) The texture images. (b) Selected features for template matching. The intensity represents the $W(i, j)$ quality measure.

Using this vector field we forward-warp the displacement and texture images and perform a bi-linear interpolation to fill the gaps. Figure 5 illustrates the flow field computed between various visemes and silence $/\#$. The result of forward warping is presented in Figure 5(c) which shows good non-rigid alignment of the outside of the lips. Regions where there are no significant features are aligned smoothly. The dark region in the centre of the mouth results from forward warping the open mouth to a closed mouth where there is no correspondence. The resulting non-rigid alignment is suitable for synthesis of transitions between visemes as will be illustrated in the next section.

4. Face Synthesis from Speech

In this section we present a framework for visual 3D face synthesis driven by speech. First, we define a set of 17

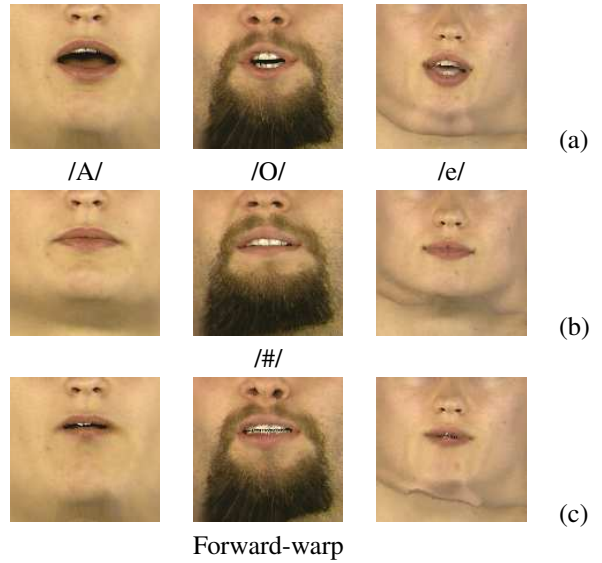


Figure 5: Non-rigid alignment of the mouth area. (a) The image to align, (b) The reference image, (c) The aligned image.

visemes (visual phonemes) which are aligned to the input audio and then we interpolate between them using the non-rigid flow field to achieve smooth transitions at the rate of 25 Hz.

4.1. Viseme Segmentation

In order to drive the visual synthesis from an audio speech sequence, we need to automatically segment and label input utterances in terms of phonemes. Phoneme to viseme mapping is then used to identify the corresponding viseme sequence which is rendered and synchronised with the audio stream. For this work we developed a system using the HMM toolkit (HTK version 3.2.1)[14] to classify each input utterance using word-level transcriptions and the British English Pronunciations (BEEP) dictionary [8]. A Hidden Markov Model (HMM) phoneme recogniser is trained using the MOCHA (Multi-CHANNEL Articulatory) database [18], which consists of 460 phonetically balanced sentences from both a male and a female speaker.

During the video-rate face shape capture session, we acquired 3D face and audio data from 51 speakers, 26 male and 25 female. For each person the following sentence which elicits the 17 visemes that cover the major mouth shapes during English was captured:

"She put the red textbook on top of the cold bed and said with a loud voice, sir do not park that car in the gap and please give me a tip."

Table 1 presents the phoneme to viseme mapping to-

gether with the corresponding exemplars for the above sentence. Forced alignment of the test utterances is used to accurately identify the start and end time stamps for each phoneme in the known transcription. Based on the speech timings the 3D video sequence is segmented into visemes.


















Vis	Phoneme		Example	Mouth
	MPEG-4	BEEP eq.		
0	none (#)	sil	N/A	
1	p,b,m	p,b,m	<u>put</u> , <u>bed</u> , <u>me</u>	
2	f,v	f,v	<u>far</u> , <u>voice</u>	
3	T,D	th,dh	<u>think</u> , <u>the</u>	
4	t,d	t,d	<u>tip</u> , <u>do</u>	
5	k,g	k,g	<u>cold</u> , <u>gap</u>	
6	tS,dZ,S	ch,jh,sh	<u>chair</u> , <u>join</u> , <u>she</u>	
7	s,z	s,z	<u>sir</u> , <u>zeal</u>	
8	n,l	n,l	<u>not</u> , <u>loud</u>	
9	r	r	<u>red</u>	
10	A	aa,ae,ah	<u>car</u> , <u>and</u>	
11	e	eh,ax	<u>bed</u> , <u>the</u>	
12	I	ih,iy	<u>in</u> , <u>me</u>	
13	U	aw,uh,uw	<u>loud</u> , <u>book</u> , <u>do</u>	
14	-	er	<u>sir</u>	
15	-	w	<u>with</u>	
16	O	oh,ow,oy	<u>top</u> , <u>cold</u> , <u>voice</u>	

Table 1: The phoneme to viseme mapping. Words in italics are not included in the example sentence.

4.2. Synthesis from Speech

For synthesis we define a viseme as $V_i = \{D_i, T_i, F_{i,0}\}$, D_i is the displacement map, T_i is the texture map and $F_{i,0}$ is the vector field deforming the viseme V_i to the neutral viseme V_0 . For a novel speech sequence analysis of the audio provides the viseme timing as shown in figure 6, where t_i^s and t_i^e is the start and end time of the i^{th} viseme with $t_i = \frac{t_i^s + t_i^e}{2}$ being the middle point representing the time instance for the static viseme V_i frame. If there are N visemes in the audio

sequence then the problem is to uniformly sample the t_{N-1}^e time period at 25fps to produce $25 \times t_{N-1}^e$ frames of animation.



Figure 6: Viseme timing

At the sampled point $t \in [t_j, t_k]$ we define an intermediate frame V^t by blending the visemes V_j and V_k . The knowledge of the t_k^s time is used to define a piecewise linear weighting function as:

$$W(t) = \begin{cases} \frac{t-t_j}{2 \cdot (t_k^s - t_j)} & (t \leq t_k^s) \\ 0.5 + \frac{t-t_k^s}{2 \cdot (t_k - t_k^s)} & (t > t_k^s) \end{cases} \quad (6)$$

with $0 \leq W(t) \leq 1$. This function can be used to obtain a weighted average between V_j and V_k at point t . To reduce the effect of the blurring in the mouth region, where most of the motion exists, we first non rigidly deform V_j and V_k to align them with V^t before the blending. The two aligning vector fields \widetilde{F}_k and \widetilde{F}_j are defined as:

$$\widetilde{F}_k = W(t) \cdot (F_j - F_k) \quad (7)$$

$$\widetilde{F}_j = (1 - W(t)) \cdot (F_k - F_j) \quad (8)$$

Then we can obtain the new frame V^t by defining D^t , T^t and F^t as:

$$D^t = W(t) \cdot \widetilde{D}_j + (1 - W(t)) \cdot \widetilde{D}_k$$

$$T^t = W(t) \cdot \widetilde{T}_j + (1 - W(t)) \cdot \widetilde{T}_k \quad (9)$$

$$F^t = F_j - \widetilde{F}_j$$

This combines the two vector fields $F(j, 0)$ and $F(k, 0)$ to obtain the interpolated vector field $F(j, k)$ as a weighted average. The resulting vector field is used to synthesise a displacement and colour image at each intermediate time frame. This results in the synthesis of a sequence of face shape and appearance corresponding to the input audio speech sequence.

5. Results

In this section we present results of the 3D face speech synthesis system. Synthesis has been performed for the 51 captured faces in the database. The front page figure illustrates a synthesised 3D face from multiple views together with the

corresponding speech waveforms. Figure 7 shows a close up of the mouth for the synthesised sequence of 7 consecutive frames of example male and female faces driven by an audio speech for the word 'hello'. We can notice how the individual differences are encoded and reconstructed by the system. The non-rigid alignment of mouth shape and appearance used to map between visemes allows synthesis without visible blurring or distortion at intermediate frames. Figure 8 illustrates three different views of 5 frames in the transition from viseme /p/ to viseme /a/. This demonstrates the advantage of 3D visual face synthesis of real people over previous 2D approach in allowing arbitrary viewing direction. Figure 9 illustrates other examples of 3D visual speech synthesis for multiple people pronouncing the consonant-vowel(CV) syllable showing the transition from /t/ to /o/. The synthesised sequences for different people reproduce their individual facial characteristics. Results demonstrate that the proposed framework achieves synthesis of 3D faces of real people with a comparable visual quality to the captured video.

Frame synthesis is implemented off-line with approximately 1 second per frame on a Pentium III 900MHz CPU while the triangulation and rendering is done in real-time on a GeForce4 Ti4200 GPU.

6. Conclusion

In this paper we have presented a novel framework for synthesis of 3D faces of real people from audio speech. The approach is based on the video-rate capture of face shape and appearance for a person. Results demonstrate synthesis of novel speech sequences with a visual quality comparable to the captured video.

To facilitate speech synthesis a novel non-rigid alignment algorithm has been presented to obtain dense correspondence between faces using shape and appearance. The algorithm integrates shape and appearance information to ensure accurate alignment in regions with significant local variation and obtain smooth correspondence in regions of uniform shape and appearance. The resulting non-rigid alignment enables transition of shape and appearance between visemes without visible blurring or distortion.

The framework and results presented in this paper demonstrate the potential for videorealistic concatenative face synthesis based on 3D video sequences rather than 2D video. The use of 3D video has the advantage of allowing arbitrary viewpoint and relighting.

References

- [1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vision*, 12(1):43–77, 1994.
- [2] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(2):239–255, 1992.
- [3] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proc. ACM SIGGRAPH*, pages 187–194, 1999.
- [4] M. Brand. Voice puppetry. In *Proc. ACM SIGGRAPH*, pages 21–28, 1999.
- [5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proc. ACM SIGGRAPH*, pages 1–8, 1997.
- [6] E. Cosatto and P.F. Graf. Photo-realistic talking heads from image samples. *IEEE Transaction on Multimedia*, 2(3):152–163, 2000.
- [7] D. DeCarlo and D. Metaxas. Deformable model-based shape and motion analysis from images using motion residual error. In *International Conf. on Computer Vision*, 1999.
- [8] British English Pronunciations (BEEP) dictionary. <http://mi.eng.cam.ac.uk/comp.speech/Section1/Lexical/Beep.html>.
- [9] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proc. ACM SIGGRAPH*, pages 388–398, 2002.
- [10] P. Fua. Regularized bundle-adjustment to model heads from image sequences without calibration data. *International Journal of Computer Vision*, 38(2):153–171, 2000.
- [11] G.A. Kalberer and L. Van Gool. Realistic face animation for speech. *Journal of Visualization and Computer Animation*, 13(2):97–106, 2002.
- [12] W.-S. Lee and N. Magnenat-Thalmann. Head Modeling from Pictures and Morphing in 3D with Image Metamorphosis Based on Triangulation. In *Modelling and Motion Capture Techniques for Virtual Environments - Magnenat-Thalmann, N. and Thalmann, D. (Eds.)*, pages 254–268. Lecture Notes in Artificial Intelligence 1537, Springer Verlag, 1998.
- [13] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin. Synthesizing Realistic Facial Expressions From Photographs. In *Proc. ACM SIGGRAPH*, pages 75–84, 1998.
- [14] S.J. Young, D. Kershaw, J. Odell, P. Woodland. *The HTK Book (for version 3.2.1)*. Cambridge University, <http://htk.eng.cam.ac.uk/docs.shtml>, 2002.
- [15] Y. Shan, Z. Liu, and Z. Zhang. Model-based bundle adjustment with application to face modeling. In *International Conf. on Computer Vision*, pages 644–651, 2001.
- [16] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing: Analysis and Machine Vision*. O'Reilly, 1999.
- [17] J. Wang and M.M. Oliveira. A hole-filling strategy for reconstruction of smooth surfaces in range images. In *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI03)*, Sao Carlos, Sao Paulo, Brazil, October, 2003.
- [18] Alan A. Wrench. *The Articulatory Database Registry: MOCHA-TIMIT*, EPSRC grant GR/L78680. Centre for Spch. Tech. Res., Univ. of Edinburgh, UK, 1999. [<http://www.cstr.ed.ac.uk/artic/mocha.html>].
- [19] I.A. Ypsilos, A. Hilton, and S. Rowe. Video-rate Capture of Dynamic Face Shape and Appearance. In *IEEE Automatic Face and Gesture Recognition*, pages 117–122, 2004.

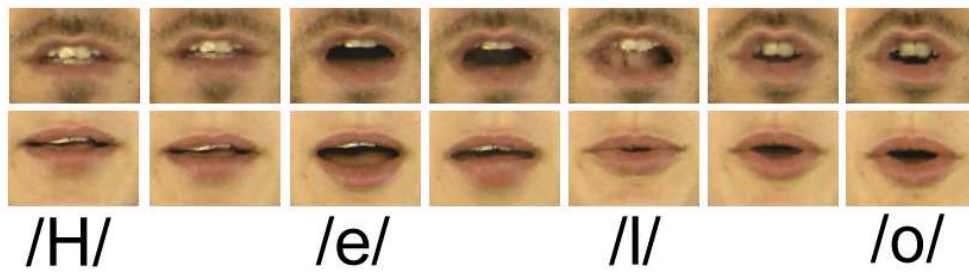


Figure 7: Consecutive frames (25fps) for (top) male and (bottom) female subjects synthesising the word 'Hello'.

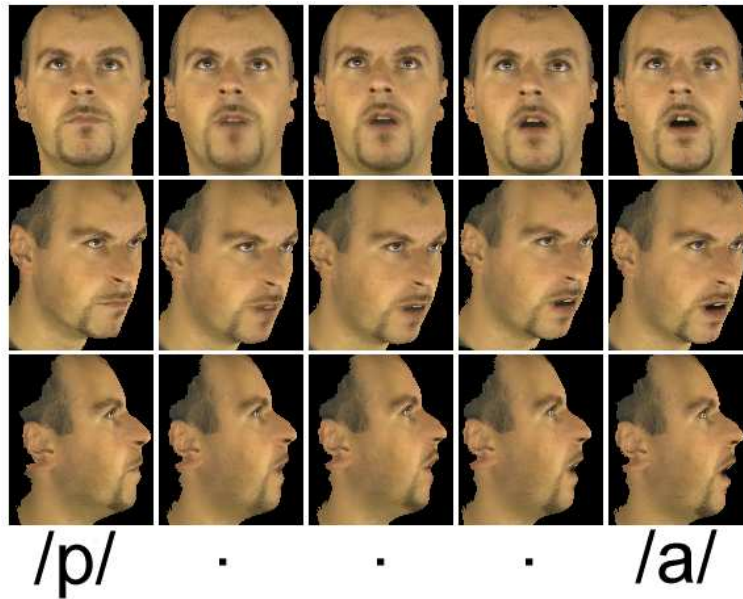


Figure 8: 3D face synthesis of a /p/ to /a/ transition for one person from multiple views.

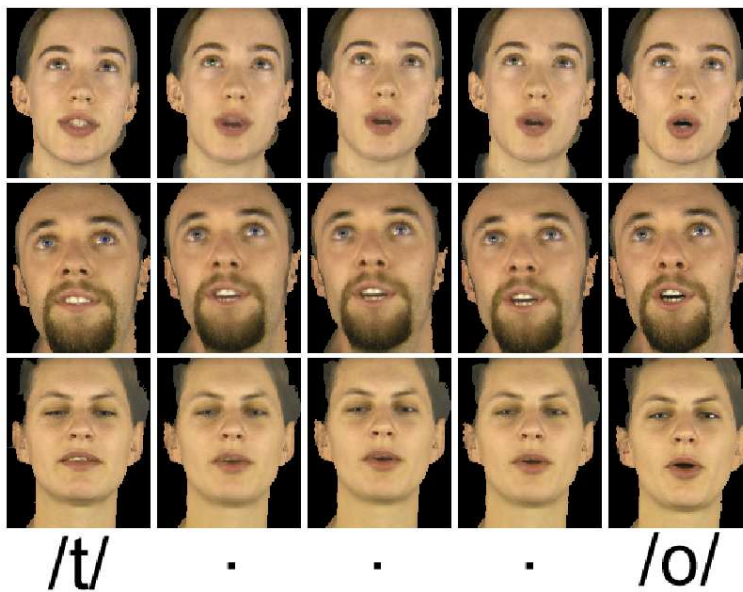


Figure 9: 3D face synthesis of multiple people pronouncing a /t/ to /o/ transition.