Lip Tracking for 3D Face Registration

Robert Anderson
Cambridge University
ra312@cam.ac.uk

Björn Stenger
Toshiba Research Europe
bjorn@cantab.net

Roberto Cipolla
Cambridge University
cipolla@eng.cam.ac.uk

Abstract

This paper presents a method for 3D lip registration when tracking a face model from video and depth map data. The inner and outer lip contours are tracked in 2D using features obtained from Boosted Edge Learning (BEL). Given the lip contours and a reconstructed depth map, the 3D mouth region of a mesh is registered at each frame independently, accounting for the self occlusions of the lips that occur during speech. Experiments show that this leads to significantly higher registration accuracy in the perceptually important mouth region. The proposed method is also shown to have lower error than an Active Appearance Model based registration method.

1 Introduction

Realistic facial animation has long been a goal in the graphics community and is typically achieved through motion capture and significant manual work by skilled artists. Computer vision techniques offer the potential to reduce the amount of work necessary to produce animations by using data-driven approaches to capture detailed 3D models of actors performing. Recent work, such as [5, 13], has demonstrated the quality of models captured using such methods.

To fit into existing animation pipelines reconstructions captured by these techniques must be registered throughout time. The lip region is difficult to track due to rapid movements and large deformations, as well as occlusions and disocclusions along the inner edge of the lips. However it is perceptually very important, especially during speech. In this paper the lips are tracked in input images and used to constrain the registration of the rest of the face in 3D. Fig. 1 shows the stages in the registration process.

2 Prior Work

One popular method for coarse registration over time is to add markers to an actor's face [3, 17, 19]. This provides robust registration, at the expense of significant set up time and occlusion of any detail behind the markers. Another approach is to use optical flow to track motion between frames and to use some form of regularisation to counter the effect of small errors [5, 6, 12, 22, 23]. Furukawa and Ponce [12] make flow computation more reliable by adding extra texture to the face while Bradley *et al.* [5] use several high resolution cameras so that fine skin details are visible. Weise *et al.* [22] aid optical flow calculation around the lips by using a colour space transformation to increase their contrast with the rest of the face.

Tracking over long sequences can lead to drift. This has recently been addressed by using either anchor frames [2] or tracking non-sequentially [16]. Whilst both of these problems reduce drift significantly they are still susceptible to errors when registering the lip region.

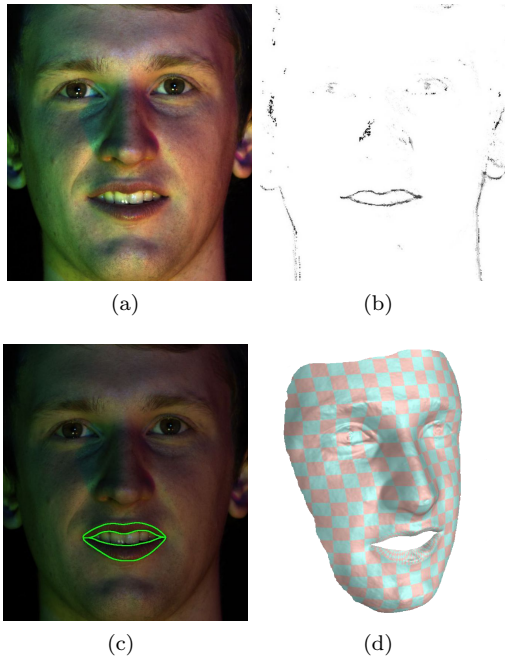


Figure 1: Algorithm overview. **(a)** *Input image,* **(b)** *BEL probability map for inner lip contour,* **(c)** *estimated lip contour positions, and* **(d)** *reconstructed mesh with consistent parametrisation.*

Morphable models provide another means of registration. Blanz and Vetter [4] demonstrated the power of these models for representing static faces and DeCarlo and Metaxas [9] showed how optical flow could be used to drive them for dynamic facial registration. While morphable models generally give visually pleasing results they cannot capture motion outside of the subspace that they span.

None of the methods mentioned above accurately register the lips. While Bradley *et al.*[5] do track the inner lip contour explicitly, this represents the occluding boundary from the camera's viewpoint, not a fixed line on the lips themselves.

Lip tracking is a long-studied problem and active contour models [14] have been used in this area for many years. In this framework a contour is optimized according to a cost that incorporates local image features as well as a smoothness term. An extension to active contours that has also been applied to lip tracking is the Active Shape Model (ASM) [8], a linear model of the lip shape learned from a small set of hand labelled images. This increases robustness and convergence speed over active contours but limits the result to lie within the linear subspace spanned by the model. A further extension of the ASM is the Active Appearance Model (AAM) [7] which takes texture information into account as well as shape, Nguyen and Mil-

gram [20] demonstrate this technique applied to lips. Eveno *et al.* [11] use a modified version of active contours, "jumping snakes", that allows for initialisation far from the target edge. Colour can also be used in lip tracking, for example in [15], and feature based approaches have also been proposed, such as [21].

3 Image Based Lip Tracking

This section presents our method for inner and outer lip contour tracking. The inner lip contour follows the occluding boundary of the lips and the outer lip contour lies between the lip region and its surrounding face region. To track the lips the probability that any given pixel is on one of the contours is calculated and an active contour is then fit to this probability map.

3.1 Contour registration

When only local descriptors are used the boundary between the lower lip and the teeth in one frame can appear nearly identical to the boundary between the tongue and teeth in another frame, making correct classification of the inner lip contour challenging. To overcome this problem a descriptor with a larger support region is used to take advantage of the surrounding structure. Boosted Edge Learning (BEL) [10] uses a small set of training images in which edges have been labelled to train a classifier for edge/non-edge pixels by applying the Probabilistic Boosting Tree classification algorithm to low-level image features taken from a wide aperture around points on the labelled edges. A separate BEL model is learned for both the inner and outer lip contours by training on a small number of labelled frames from the sequence to be tracked.

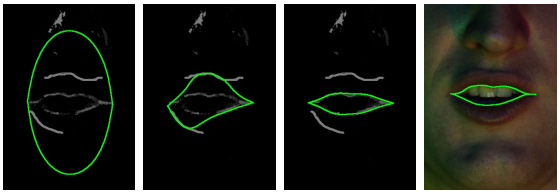
Once probability maps have been estimated for the inner and outer contours, they are tracked sequentially using an active contour [14] based approach. Each contour is defined by N_c points, each with a 2D image position $\mathbf{a}_i$. The outer contour is registered with the additional constraint that it must coincide with the inner lip contour at the mouth corners.

Based upon the mouth position of the previous frame an estimate of the location of the mouth corners is made. The contour is initialised as an ellipse passing through these corner points with its major axis oriented vertically and a factor 1.5 times longer than its minor axis to ensure that the starting position of the contour is outside of the true contour, even when the mouth is open Fig. 2(a). The contour is then contracted by iterating between the following two steps until convergence:

1. For each point on the contour find the inward facing normal from its two neighbours. Search along a line in this normal direction starting from a distance 3δ outside the contour until a point in the probability map with value greater than τ is reached or until the search has moved a distance δ inside the contour.

2. Smooth the contour by averaging each point's position with that of its neighbours n times.

The effect of these steps is to contract the contour, allowing it to find edges with high probability in the edge image only if they form a nearly completely closed contour. The smoothing step is essential and ensures that small gaps in the true lip contour are bridged. A sample optimisation can be seen in Fig. 2.



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

Figure 2: Active contour convergence. *Convergence given an edge probability map. Two additional lines have been added to the map to show the robustness of the proposed approach to false edges. The contour does not move inside the inner lips despite gaps in the edge map.* **(a)-(c)** *Iterations 0, 40, 100.* **(d)** *Final position after optimisation from section 3.2.*

The threshold τ can be set over a very wide range due to the near binary nature of the probability maps produced by the BEL stage, in all experiments it was set to 0.5. The step length δ, coupled with the number of smoothing iterations n determines the ability of the contour to move past partial edges. In all experiments presented here δ was set to 2 pixels and n to 20.

3.2 Contour optimisation

The procedure outlined above is highly robust but results in the active contour settling on the outside edge of the true contour instead of at its maximum. A refinement step is used to improve the final result.

For each point on the contour a line search is carried out along a line perpendicular to the contour to find the maximum in a smoothed version ($\sigma = 2$) of the probability map. For point i the maximum has a value v_i and occurs at image point $\mathbf{m}_i$. An optimisation is then carried out to minimise the energy

$$E = E_{int} + \sum_{i=1}^{N_c} v_i \|\mathbf{a}_i - \mathbf{m}_i\|^2, \tag{1}$$

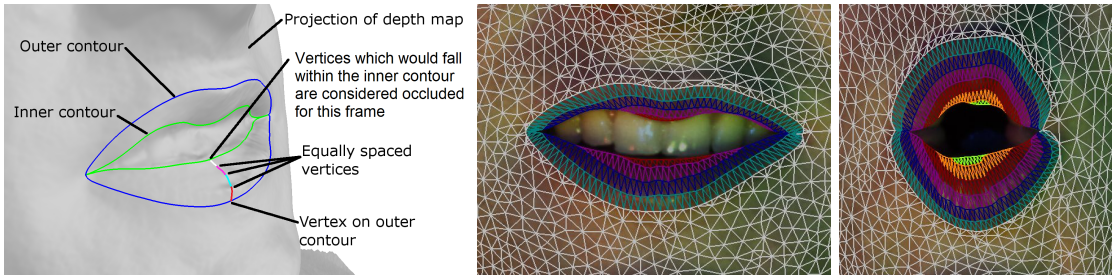
where E_{int} is the active contour's internal energy [14]:

$$E_{int} = \sum_{i=1}^{N_c} \alpha_i \|\mathbf{a}_i - \mathbf{a}_{i+1}\|^2 + \sum_{i=1}^{N_c} \beta_i \|\mathbf{a}_{i-1} - 2\mathbf{a}_i + \mathbf{a}_{i+1}\|^2. \tag{2}$$

Modulo arithmetic is used for the indices since the contour is closed. α_i and β_i are the same for all points apart from the two mouth corners at which β_i is set to zero since a discontinuity in gradient is expected.

4 Face registration

To register a 3D face model over time we use the approach of [23] with a modified regularisation term. An initial mesh, M_0, is built in the first frame of the sequence and is defined by N_v vertices, where each vertex i has an initial position $\mathbf{p}_{0i}$ and a set of n_i neighbours whose indices form the set η_i. As the mesh is deformed over time the position $\mathbf{p}_i$ of each vertex i must be found. At each frame an input depth map is given along with the previous mesh and an optical



(a) (b) (c)

Figure 3: 3D Lip Representation. **(a)** *Formation of lip mesh. Outer and inner contours are projected onto the depth map and vertices are placed between them in steps of constant geodesic distance.* **(b)** *and* **(c)** *The number of vertices used to represent the lips changes depending on the visible lip surface.*

flow field from the last frame. The weighted sum of three energies is minimised; a depth energy E_d, a flow energy E_f and a regularisation energy E_r:

$$E = w_d E_d + w_f E_f + w_r E_r. \quad (3)$$

The depth and flow terms are the same as those in [23] and enforce that the deformed mesh is consistent with both the observed depth maps and the flow field from the previous frame.

We use a regularisation term which enforces regularisation in Laplacian coordinates:

$$E_r = \sum_{i=1}^{N_v} \frac{1}{n_i} \left\| \sum_{j \in \eta_i} (\mathbf{p}_i - \mathbf{p}_{0i}) - (\mathbf{p}_j - \mathbf{p}_{0j}) \right\|^2 . \quad (4)$$

We found this regularisation to perform better on the sequences tested than the original regularisation term in [23] as it does not penalise uniform stretching.

The L-BFGS algorithm is used to minimise (3). The above framework provides reasonable tracking over the majority of the face, however around the lips the optical flow term is not reliable due to rapid movement along with occlusions and disocclusions.

4.1 3D Lip registration

2D lip tracking can be used to constrain the registration process described above. The two contours tracked in section 3 have two different physical meanings and are used in different ways.

The outer lip contour corresponds to a fixed line defined as the boundary between the lips and the surrounding face region. This means that points on this contour correspond to the same points on the mesh at all times. This can be enforced by adding an extra term E_l to equation (3) which forces the set of vertices V_l, that lie on the boundary between the lip and face region, to coincide with the tracked outer contour:

$$E_l = \sum_{i \in V_l} \|\mathbf{p}_i - \mathbf{l}_i\|^2 , \quad (5)$$

where $\mathbf{l}_i$ is the projection onto the depth map of the point on the tracked contour corresponding to vertex i. This term is added to the weighted sum in equation (3) to give:

$$E = w_d E_d + w_f E_f + w_r E_r + w_l E_l. \quad (6)$$

The inner lip contour corresponds to the occluding boundary of the lips and as such does not correspond to a fixed point. Due to the occlusion and disocclusion occurring around the lips optical flow estimation can be unreliable leading to poor registration. To simplify the registration problem in this region we assume that the lips do not stretch in the direction perpendicular to the lip contours and we do not include the vertices on the lips in the optimisation of equation (6). Instead, the inner and outer lip contours are projected onto the depth map and vertices are placed starting from the outer contour going towards the inner contour in steps of constant geodesic distance. This means that as the lips become partly occluded, the number of lip vertices changes as can be seen in Fig. 3.

5 Experiments

The proposed method was applied to a 2000 frame sequence of a single speaker. The BEL classifiers were trained on 13 hand labelled frames from the sequence, chosen to cover a wide range of lip shapes. Depth maps were provided through a combination of multi-view stereo and photometric stereo as in [1]. Optical flow was estimated using the implementation of Liu [18]. Both depth maps and flow fields had a resolution of 1200×1600. The weightings in equation (6) were set to $E_d = \frac{1}{\epsilon^2}$, $E_f = 1$, $E_r = \frac{1}{\epsilon^2}$, $E_l = \frac{1000}{\epsilon^2}$, where ϵ was the average edge length in the mesh. Qualitative results are best seen dynamically in the supporting video, however several stills are shown in Fig. 4. Throughout the 2000 frame sequence there was only a single tracking failure which was recovered in the next frame.

To quantitatively analyse the accuracy of the system the inner lip contour was labelled in every hundredth frame and the average distance of each vertex on the inner lip from this contour was measured. Fig. 5 shows the improvement by including explicit lip tracking; the average error over the 20 labelled frames was reduced from 7.7 pixels to 1.7, with the standard deviation of the results falling from 1.7 to 0.53 pixels. We compared the proposed method to an Active Appearance Model (AAM) based approach for 2D lip tracking. An AAM was trained on the same images that were used for the BEL models and used to track the lip region. It can be seen from Fig. 5 that the average AAM tracking error is slightly higher (a mean of 2.7 pixels with standard deviation of 0.62 pixels). More importantly, AAM tracking was less robust, and the system had to be manually reinitialised four times throughout the sequence.

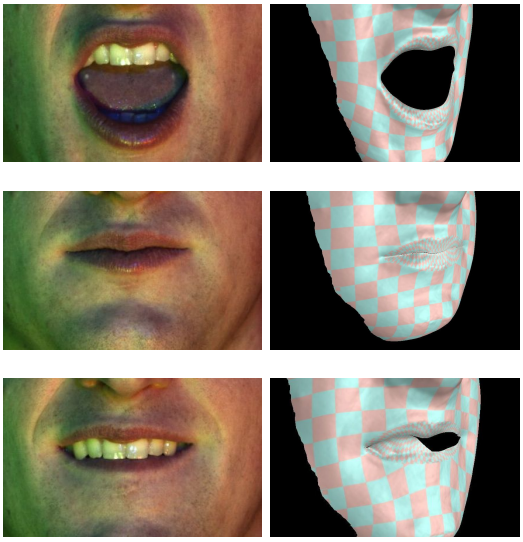


Figure 4: Face registration results. *First column: input image. Second column: resulting reconstruction. The top two rows demonstrate successful registration. The third row shows the only frame in the 2000 frame sequence in which tracking fails.*

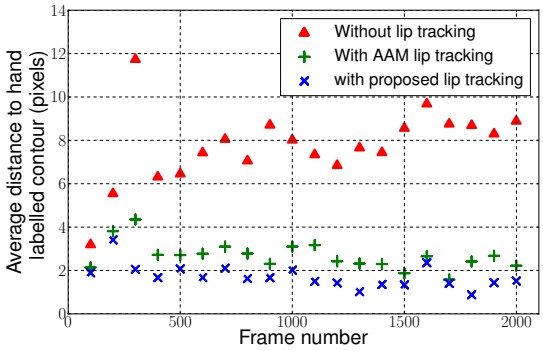


Figure 5: Registration error. *Mean distance of projections of 3D vertices on the inner lip contour to a hand labelled contour. Explicit lip tracking significantly reduces error and the proposed approach outperforms tracking using an AAM.*

Similar results were achieved when testing on a second subject. The BEL classifier had to be retrained using training samples from this new subject.

6 Conclusions and future work

A practical method for registering lip motion in 3D has been proposed. BEL has been shown to be an effective means of classifying points belonging to both the inner and outer lip contours and a robust active contour fitting approach was demonstrated using the BEL probability maps. The contours tracked in this way were used to deform a mesh whilst taking account of the occlusions and disocclusions that occur during speech at the inner lip boundary. Registration was successfully demonstrated on a 2000 frame sequence and outperformed tracking using an AAM.

In the future it would be interesting to apply the same technique to other structures on the face which form closed contours such as the eyes or eyebrows and also to build a person independent BEL classifier and observe its performance.

References

[1] R. Anderson, B. Stenger, and R. Cipolla. Color photometric stereo for multicolored surfaces. In *ICCV*, pages 2182–2189, 2011.

[2] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In *ACM TOG*, volume 30, pages 75:1–75:10, 2011.

[3] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross. Multi-scale capture of facial geometry and motion. *ACM TOG*, 26(3):33, 2007.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, pages 187–194, 1999.

[5] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. *ACM TOG*, 29(4), 2010.

[6] G. Brostow, C. Hernandez, G. Vogiatzis, B. Stenger, and R. Cipolla. Video normals from colored lights. *PAMI*, 2011.

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *ECCV*, 2:484–498, 1998.

[8] T. Cootes and C. Taylor. Active shape models-smart snakes. *BMVC*, 1992.

[9] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *CVPR*, pages 231 –238, 1996.

[10] P. Dollar, Z. Tu, and S. Belongie. Supervised Learning of Edges and Object Boundaries. *CVPR*, (12):1964–1971, 2006.

[11] N. Eveno, A. Caplier, and P. Coulon. Accurate and quasi-automatic lip tracking. *IEEE T Circ. Syst. Vid.*, 14(5):706 – 715, 2004.

[12] Y. Furukawa and J. Ponce. Dense 3D motion capture for human faces. In *CVPR*, 2009.

[13] G. Fyffe, T. Hawkins, C. Watts, W. Ma, and P. Debevec. Comprehensive facial performance capture. *Eurographics*, 2011.

[14] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *IJCV*, 1(4):321–331, 1988.

[15] R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. *ICCV*, pages 370 –375, 1998.

[16] M. Klaudiny and A. Hilton. Towards optimal non-rigid surface tracking. *ECCV*, pages 743–756, 2012.

[17] M. Klaudiny, A. Hilton, and J. Edge. High-detail 3D capture of facial performance. *3DPVT*, 2010.

[18] C. Liu. Beyond pixels: Exploring new representations and applications for motion analysis. *Doctoral Thesis, MIT*, 2009.

[19] W. Ma, A. Jones, J. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *SIGGRAPH*, 27(5), 2008.

[20] Q. Nguyen and M. Milgram. Semi adaptive appearance models for lip tracking. *ICIP*, pages 2437 –2440, 2009.

[21] Y. Tian, T. Kanade, and J. Cohn. Robust lip tracking by combining shape, color and motion. *ACCV*, 2000.

[22] T. Weise, H. Li, L. Van Gool, and M. Pauly. Face/Off: live facial puppetry. In *SIGGRAPH*, pages 7–16, 2009.

[23] L. Zhang, N. Snavely, B. Curless, and S. Seitz. Spacetime faces: high resolution capture for modeling and animation. *SIGGRAPH*, 23(3):546–556, 2004.