Sumit Basu, Nuria Oliver, and Alex Pentland. "Coding Human Lip Motions with a Learned 3D Model." In Proceedings of the Int'l Workshop on Very Low Bitrate Video Coding (VLBV '98). Urbana, Illinois. October, 1998.

Coding Human Lip Motions with a Learned 3D Model

Sumit Basu, Nuria Oliver, and Alex Pentland
Perceptual Computing Section, MIT Media Laboratory
E15-383, 20 Ames Street, Cambridge, MA 02139 USA
{sbasu,nuria,sandy}@media.mit.edu

Abstract

The lips are a critical factor in spoken communication and expression. Accurately tracking and synthesizing their motions from arbitrary head poses is essential for high-quality video coding. Our approach is to build and train 3D models of lip motion to compensate for the limited information available during tracking. We use physical models as a prior and combine them with statistical models, showing how the two can be smoothly and naturally integrated into a synthesis method and a MAP estimation framework for tracking. Because the resulting description has a small number of parameters, it is ideal for coding as well. We show how our methods allow us to accurately recover the 3D lip shape from raw 2D video data and resynthesize this shape with a small number of parameters.

1 Introduction

It is well-known that lips play a significant role in spoken communication. Summerfield's classic 1979 study [6] showed how presence of the lips alone (without tongue or teeth) raised word intelligibility in noisy conditions from 22.7% to 54% on average and up to a maximum of 71%. Not only can the lip shape be used to reduce noise and enhance intelligibilty for human/machine speech understanding, but also as a significant feature for expression understanding. As result, it is critical to accurately capture and resynthesize lip motions for high quality video coding.

However, to realize this in practice, it is necessary to robustly and accurately track the lips in 3D. Why is 3D so critical? In natural conversation and expression, we move our heads constantly, both in translation and rotation. If we cannot contend with this simple fact, we will never reach the unconstrained interfaces we desire. Computer vision techniques have been developed that accurately track the head's 3D rigid motions, leaving the formidable task of tracking the remaining 3D non-rigid deformations.

In this paper, we develop a method for successfully facing this difficult problem. One of the most vexing issues surrounding the lip tracking problem has always been the poor quality of the available data – contours, color, flow, etc., are all obscured at some point or other by lighting, the speed of motion, and so on. Our approach is thus to build and rely on strong models of the lip shape to correct for anomalies in the data. In essence, our model learns the permissible space of lip motions. The incoming data from the video stream is then *regularized* by this model – we find the *permissible* lip shape that could best account for the data. In this way, we remain robust to the unavoidable noise in the raw features. To build and train this model, we start by giving a lip-shaped mesh generic physical characteristics using the Finite Element Method (FEM). This acts a physically based "prior" (i.e., locally elastic behavior) on how things move. We then train this model with 3D data of real lip motions and blend the physical prior with the statistical characteristics of this data. Finally, we use this physical-statistical model in a MAP estimation framework to find the locally most probable lip shape that can account for the incoming data. Along the way, we have developed a full-fledged synthesis model as well – by moving the model through the permissible lip space, we can generate images of the 3D lips in motion. In addition, this space is parametrized by only ten parameters (which have a great deal of interframe correlation), allowing for efficient coding of the lip shape.

Through this method, we have been able to robustly and accurately track lip shapes in 3D from arbitrary head poses in a video stream. We will demonstrate our results with an illustration of the learned lip subspace, numerical figures on reconstruction accuracy, examples of static fits of the model, and audio-visual sequences demonstrating the tracking and synthesis in action.

1.1 Background

In looking at the prior work on lip modeling and tracking, there are two major groups of models. The first of these contains the models developed for analysis, usually intended for input into a combined audio-visual speech recognition system. The underlying assumption behind most of these models is that the head will be viewed from only one known pose. As a result, these models are only two-dimensional. Many are based directly on image data [5]; others use such low level features to form a parametrized description of the lip shape [1]. Some of the most interesting work done in this area has been in using a statistically trained model of lip variations (such as [4]). However, since these are 2D models, the changes in the apparent lip shape due to rigid rotations have to be modeled as complex changes in the lip pose. In our work, we begin by extending this philosophy to 3D.

The other category of lip models includes those designed for synthesis and facial animation. These lip models are usually part of a larger facial animation system, and the lips themselves often have a limited repertoire of motions. To

their credit, these models are mostly in 3D. For many of the models, though, the control parameters are defined by hand. A few are based on the actual physics of the lips: they attempt to model the physical material and musculature in the mouth region [7]. Unfortunately, the musculature of the mouth is extremely complicated and has proved to be very difficult to model accurately. Even if the modeling were accurate, this approach would still result in a difficult control problem.

We hope to fill the gap in these approaches with our learned 3D model, which can be used for both analysis and synthesis.

2 The Model

In the following section, we give a brief description of the choice of the model shape and the physics used. A more detailed account of the finite element method and the training method for our model is given in [2].

The underlying representation of our initial model is a mesh in the shape of the lips. At the initial stage, before any training has occurred, we have no learned notion of the lip shape. We thus simply extract the region surrounding the mouth in a Viewpoint Data Labs model of the human head and make a few minor changes to aid the physical modeling steps ahead. The final model has 336 faces and 204 nodes, resulting in 612 degrees of freedom (three per node).

Similarly, we have no real idea what the inherent degrees of freedom of the lips are. However, we do know something about how the lip material behaves, namely that it acts in a locally elastic way. When one portion of the lips is pulled on, the surrounding region stretches with it. We express this notion mathematically in our model by using the Finite Element Method (FEM). We use this method to give this initial mesh the properties of a generic elastic material – i.e., we treat the mesh as if it were formed from a rubber sheet. The resulting first-stage model is a "physical prior" for our training stages to come.

3 The Observations

To train this model to have the correct 3D variations of the lips, it was necessary to have accurate 3D data. Seventeen points were marked on the face with ink: sixteen on the lips and one on the nose. The placement of these points is shown in figure 1. The points were chosen to obtain a maximally informative sampling of the 3D motions of the lips. Once the points were marked, two views of the points were taken by using a camera-mirror setup to ensure perfect synchronization between the two views. The points were tracked over 150 frames at a 30Hz frame rate using supervised normalized correlation. It was attempted to have as great a variety of lip motions within this brief period as possible. The two views were then used to reconstruct the 3D point locations. Finally, the points were transformed into a head-aligned coordinate system to prevent the rigid motion of the head from aliasing

Figure 1: Locations of marked points on the face

with the non-rigid motions of the lips. See [2] for further details on these methods. Methods to continue the training using other forms of input data (lipstick, unmarked but clean data, etc.) are discussed in [3].

4 Training the Model

In order to relate the training data to the model, the correspondence between data points and model nodes had to be defined. This was a simple process of examining a video frame containing the marked points and finding the nodes on the lip model that best matched them in a structural sense. The difference between the observed point locations and their current locations in the model was then the displacement goal.

4.1 Reaching the Displacement Goals

The issue was then how to reach these displacement goals. The recorded data points constrained only 48 degrees of freedom (16 points on the lips with three degrees of freedom each) out of 612. We need the physically correct solution for the rest: we want to pin down the constrained points and let the other points go to their equilibrium locations.

Mathematically, this idea translates to the constraint of minimum strain. We wish to use the physics of the model to smooth out the regions where we have no observation data by minimizing the strain in the model. Fortunately, in the finite element framework, this solution can be found analytically and with little computation. Details of this method can be found in [2].

Using this method, we find the displacement of all the model nodes for all the frames. We then find the 10 linear modes that account for the greatest amount of variance in the input data by performing principal components analysis (PCA) on the sample covariance matrix. We can then reconstruct the modal covariance and $\mathbf{K}^{-1}$ matrices using these modes. We thus have a parametric description of the subspace of lip shapes (the modes) and a probability measure for the subspace (the modal covariance matrix).

Frontal and profile views of the the mean displacement ($\bar{u}$) and some of the first few modes are shown in figure 2 below. Though we are only using the first ten modes, it was found that these account for 99.2 percent of the variance in the data. We should thus be able to reconstruct most shape variations from these modes alone.

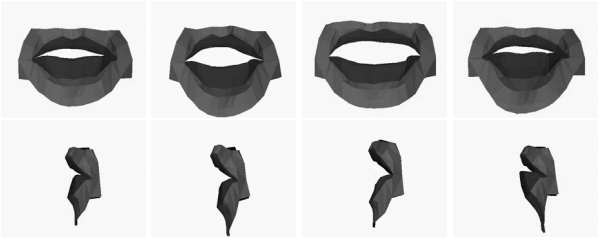

Figure 2: Front and side views of the mean displacement and some characteristic modes

5 Tracking the Lips in Raw Video

At this point, we have a parametric model of the permissible lip shapes and a probability model for the resulting subspace. The remaining task is to fit this model to the raw video stream in the absence of special markings on the lips or face. As we have stated from the beginning, our approach will be to find the lip shape within the learned subspace that best accounts for the incoming data. Statistically, this means finding the parameters with the highest *a posteriori* probability given the observations and our prior model. Intuitively, though, it simply means balancing the potentially noisy data from our observations with our learned notion of what shapes are permissible.

Any of a number of features (or a combination thereof) could be used as observations in our framework – color classification, optical flow, contours, tracked points, etc. For this implementation, we have chosen to use only the color content of the various regions, as it is a robust and easily computable candidate. Of course, this feature will not directly give us any kind of shape information – it will only give us the probabilities of each pixel belonging to the color classes, $f_{model} = f(color|model)$. From our statistical perspective, though, it is clear how this data should be used. We wish to find the set of parameters p^* for our model that maximizes its posterior probability given the observations:

$$p^* = \arg\max_p f(p|O) = \arg\max_p \frac{f(O|p)f(p)}{f(O)} \qquad (1)$$

we can neglect the denominator in the last expression, since it will be the same for all p. We can also maximize the log of the quantity instead of the original form, since the logarithm is a monotonic function. This leaves us with

$$p^* = \arg\max_p \big(\log f(O|p) + \log f(p)\big) \qquad (2)$$

Another piece of information we have is the color class of each point on our model. As shown in the figures above, the model contains the lips and some surrounding skin, and we know *a priori* which triangular faces belong to which class. If we now project the model in state p into the camera view, we can compute the term $f(O(x,y)|p)$ for each point in the visible surface of the model. This value is simply the probability of the observed color value at (x,y) belonging to the same class as the point in the model that is projected onto it. To find the overall probability of the model in this state, we simply take the product of the observations, which becomes a sum under the logarithm.

In order to apply these ideas to our tracking problem, we first train models of the color classes for the skin and lips. Next, we compute the probability maps for the image (i.e., 2D maps whose entries are the probability values of the given class). The model is then initially positioned based on the rigid pose and geometry of the head.

From this initial fit, we compute the gradient of the optimization function in the parameter space and take a step in this direction, iterating this process to climb to a local maximum of the posterior probability. The gradient we seek is:

$$\frac{d\log f(p|O)}{dp} = \frac{d\log f(O|p)}{dp} + \frac{d\log f(p)}{dp} \qquad (3)$$

We then use this results to take a step in the direction of the overall gradient. We continue this ascent process until we have converged to a local maximum, which typically occurs in less than twenty iterations. The computations necessary for these calculations can be minimized by using a few simplifications, as is shown in a forthcoming paper [3].

6 Results

In this section, we demonstrate the reconstruction and tracking capabilities of our method. We first show numerical results that demonstrate the capability of our model to accurately reconstruct 3D lip shapes from 2D data. We then go on to show examples of using the tracking method described above to capture the lip shape from a 2D video stream and reconstruct the 3D shape. We show this both with example fits in static frames and with audio-visual sequences. We also discuss the advantages of the modal form of our model.

6.1 Reconstruction Capabilities

As we have previously discussed, one of the major arguments behind the 3D representation was that we could use a small number of observations from any viewpoint to find a good estimate of the model shape. This is because we have learned the subspace of permissible lip shapes. Without the model, the 2D observations would leave far too many degrees of freedom unconstrained. With the model, as we will show, we can accurately reconstruct all degrees of freedom. We demonstrate this by reconstructing the 3D shape using only x-y (frontal view) data and only y-z (side view) data.

The mean-squared reconstruction errors per degree of freedom were then found for two cases of 2D observation scenarios and are shown in the table below. The results are given in the coordinate system of the model, in which the model is 2.35 units across, 2.83 units wide, and 0.83 units deep. The table shows the reconstruction error using only the first ten modes. Note that in both cases (see table 6.1), the reconstruction is quite accurate in terms of mean-squared error. This shows

that the ten learned modes are a sufficiently strong characterization to accurately reconstruct the 3D lip shape from 2D data.

Data Used	3D Reconstruction Error
xy (frontal)	6.70e-3
yz (profile)	7.13e-4

6.2 Tracking and Reconstruction Results

In this section, we show several examples of using our algorithm to estimate the 3D lip shape. The figures below (figures 3, 4, 5, and 6) show some other frames with the initial image, the final converged fit, and the profile view of the estimated model. The audio-visual sequences these frames are taken from, along with the tracking and reconstruction views, are available at <http://www.media.mit.edu/~sbasu/lips>.

Figure 3: Initial image, final fit, and 3D reconstruction

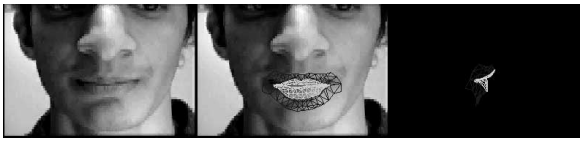

Figure 4: Initial image, final fit, and 3D reconstruction

It is worth noting here the flexibility that we gain from having a modal representation. As we have already described, the first few modes account for the greatest variation in lip shape, whereas the last few contribute the least. The more modes we use, the more accurately we can fit the shape. The fewer modes we use, the more robustly we can reject noise, since we only move along the directions of the greatest variation. The modal representation thus gives us the powerful capability of moving smoothly between high accuracy (many modes) and high robustness (few modes), allowing us to adapt to the quality of the available data and the bandwidth of the communication channel.

7 Conclusions and Future Directions

We have presented a method for estimating and reconstructing the 3D shape of human lips from raw video data. This method began with a physical model with generic physical properties – a rubber sheet in the shape of the lips. We then used 3D observations to train this intial model with the true variations of human lip shapes. This model fit naturally into a MAP estimation framework, which we then used for tracking and resynthesis of 3D lip shapes, all with a small (and flexible) number of parameters. We have shown through static and video examples how we can accurately track the observations in raw data, and have also demonstrated the ability of our model to accurately reconstruct 3D lip shapes from sparse 2D data.

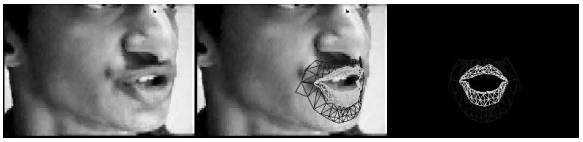

Figure 5: Initial image, final fit, and 3D reconstruction

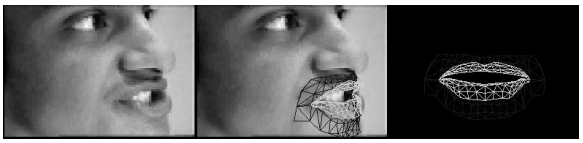

Figure 6: Initial image, final fit, and 3D reconstruction

8 Acknowledgements

We would like to thank the NSF and the La Caixa Foundation for their support in the form of graduate fellowships. Thanks also to Viewpoint Data Labs, who provided the model from which the initial shape of the lips was extracted.

References

[1] A. Adjoudani and C. Benoit. "On the Integration of Auditory and Visual Parameters in an HMM-based ASR". In David Stork and Marcus Hennecke, editors, *Speechreading by Man and Machine*, pages 461–472. Springer, 1995.

[2] Sumit Basu. "A Three-Dimensional Model of Human Lip Motion". Master's thesis, MIT Department of Electrical Engineering and Computer Science, 1997.

[3] Sumit Basu, Nuria Oliver, and Alex Pentland. "3D Lip Shapes from Video: A Combined Physical-Statistical Model". *Speech Communication*, 26:131–148, 1998.

[4] Christoph Bregler and Stephen M. Omohundro. "Nonlinear Image Interpolation using Manifold Learning". In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Info. Proc. Sys. 7*, pages 401–408, Cambridge, MA, 1995. MIT Press.

[5] Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile. "2D Deformable Models for Visual Speech Analysis". In David Stork and Marcus Hennecke, editors, *Speechreading by Man and Machine*, pages 391–398. Springer, 1995.

[6] Quentin Summerfield. "Use of Visual Information for Phonetic Perception". *Phonetica*, 36:314–331, 1979.

[7] K. Waters and J. Frisbie. "A Coordinated Muscle Model for Speech Animation". In *Proceedings of Graphics Interface '95*, pages 163–170, Ontario, Canada, May 1995.