3D Modeling and Tracking of Human Lip Motions

MIT MEDIA LABORATORY

Sumit Basu
sbasu@media.mit.edu

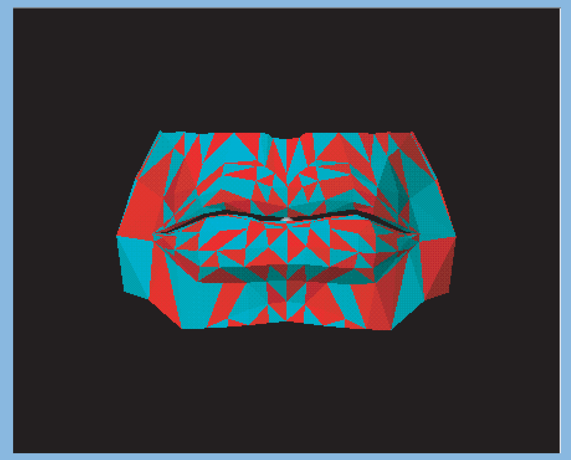
Nuria Oliver
nuria@media.mit.edu

Alex Pentland
sandy@media.mit.edu

TRAINING

The Problem

+ lips carry critical information about speech and expression
+ difficult to track due to complex, non-rigid deformations



The Approach

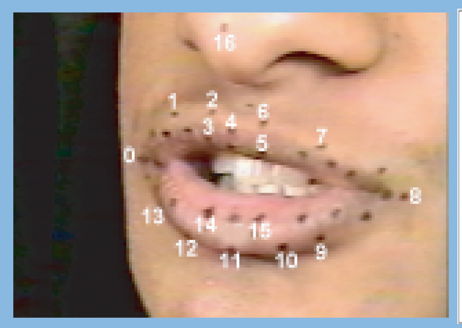
+ create a 3D, physically-based model
+ train the physics from observations
+ use the learned model for tracking

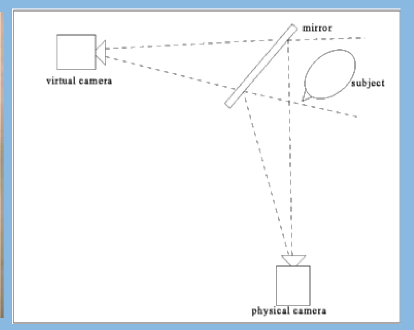
Initial Model

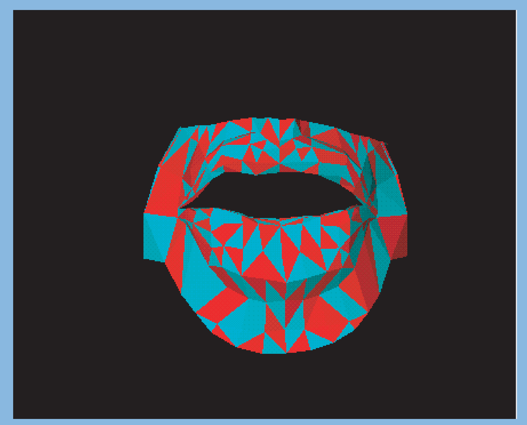
+ shape taken from graphics model
+ FEM with generic physical properties (uniform elastic material)
+ need to learn correct physics from observations

The Observations

+ 17 points marked on the lips, face
+ tracked in 3D from 2 views at 30 Hz
+ data contained full range of lip motions used for English







Deforming the Model

+ tracked points only constrained a few degrees of freedom
+ find overall deformation by minimizing overall strain given these constraints
+ this technique gives us the displacements for all model points

THE MODEL

Extracting the Model

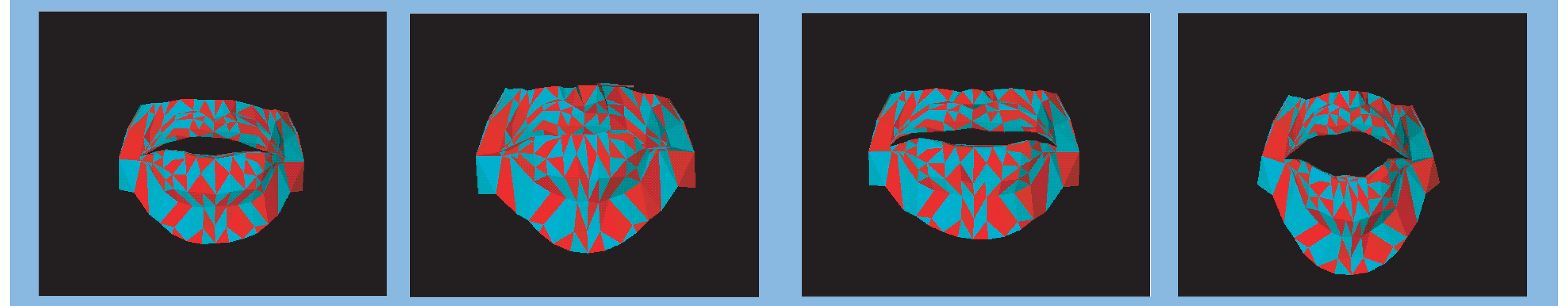
+ decompose deformations with Principal Components Analysis
+ top 10 modes cover 99.2% of variance
+ the mean and three modes are shown below

What the Model Means

+ relationship between covariance R and strain matrix K:

$$\mathbf{R} = \mathbf{K}^{-2}$$

+ connects physical and probabilistic interpretations of model
+ we thus have a probabilistic model for the subspace of permissible lip shapes (the modes and covariance) and a means of deforming it with forces



Data Used	3D Reconstruction Error
xy (16 points)	6.70e-3
yz (16 points)	7.13e-4
xyz (8 points)	1.10e-3

3D Reconstruction

+ test ability of model to reconstruct 3D shape from 2D data
+ using learned model, find MAP estimate of 3D point locations given subset of data
+ MSE per point is given in table at left in (model is 2x3x5 units)
+ these results imply we can accurately reconstruct 3D lip shape from 2D observations due to the restricted subspace of possible motions learned by the model

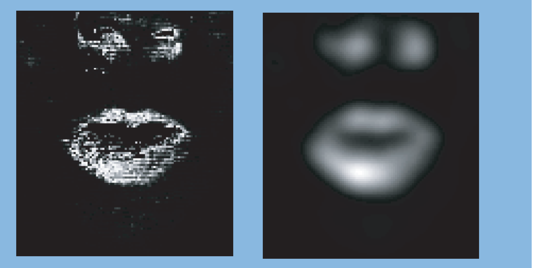
TRACKING

Tracking Lips in Raw Video

+ we no longer have 3D observations, but we have seen that we can accurately reconstruct the 3D shape from 2D data (given the head pose)
+ we now need to optimize the parameters p* (the coordinates in the learned subspace) given some general observations O using the following relation:

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

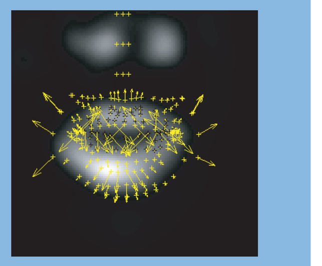
+ we use the lip/skin probability maps obtained with the system for our observations due to their robust nature. We then smooth the maps for gradient computations.



Iterating to a Solution

+ we find the posterior probability of the parameter values given the observation. This quantity is found by taking the scaled product of the observation probabilities:

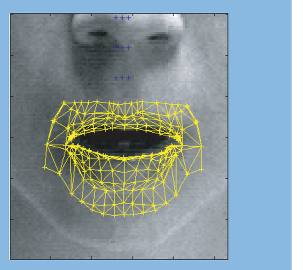
$$\log f(p|O) = \log[f(p)\prod_i f(O(x,y)|p)] = \log f(p) + \gamma \sum_i \iint_{face\ i} \log f(O(x,y)|p)$$



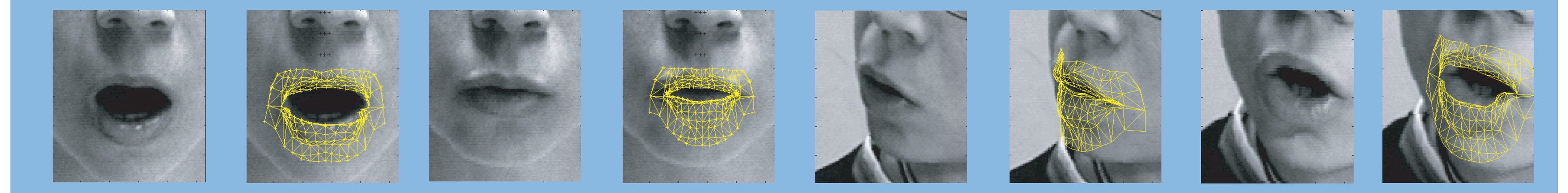
+ we then use gradient ascent to optimize the parameters. Because of the linearity of the learned subspace, the gradient can be computed with minimal computational cost:

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

from map gradients — modes — weighted inner product



+ we then iterate this process until the posterior probability converges to a local maximum. The images below show several examples of the final fit found by this technique.



FUTURE WORK

+ intraperson mode/eigenvalue variations
+ integrating with automatic 3D head-tracking work
+ learning paths through model's state space (i.e., constraints *within* the subspace)

APPLICATIONS

+ ilipreading/expression recognition
+ audio-visual speech recognition
+ video coding

MORE INFORMATION

+ for more information, and audio-visual sequences demonstrating the tracking results, see http://www.media.mit.edu/~sbasu/lips