Image-based Modeling and Rendering

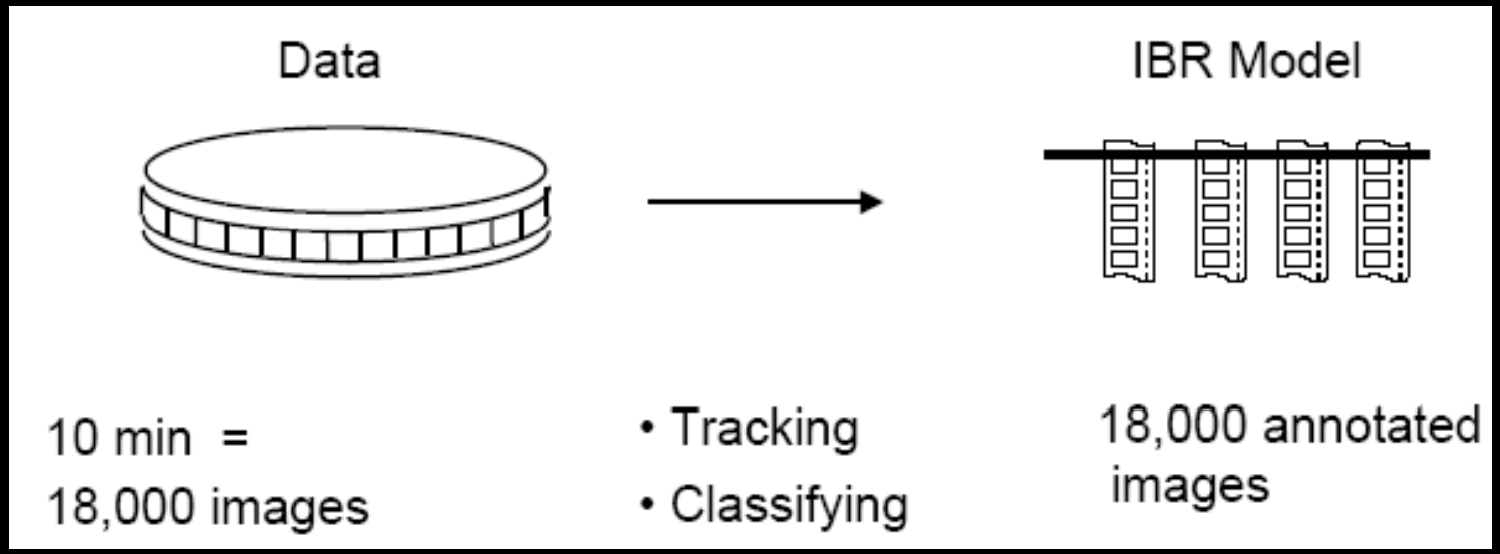
7. Advanced Topics in IBMR : Image-based Animation

National Chiao Tung Univ, Taiwan

By: I-Chen Lin, Assistant Professor

Introduction

- Synthesizing complex or highly deformable objects are still challenging.
  - Faces, plants, hairs, cloth, flame…..
- Are image-based approaches applicable?
- The challenge of IBMR in animation



Introduction (cont.)

- We've talked about video texture and transition.
- This talk will focus on animating people.

Ref:

• Image-based Modeling and Rendering, SIGGRAPH'99 course notes.

• C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: Driving Visual Speech with Audio ", Proc. SIGGRAPH'97, pp. 353-360.

• E. Cosatto, H.P. Graf, "Sample-based synthesis of photo-realistic talking heads", Proc. Computer Animation '98, pp. 103-110

• T. Ezzat, G. Geiger, and T. Poggio, Proc. ACM SIGGRAPH'02, pp.388-398.

• J. Carranza, C. Theobalt, M.A. Magnor, H.-P. Seidel, Proc. SIGGRAPH'03, pp.569-577.

Research about facial animation

- Topics for an animated face:
  - How to render a realistic face?
  - How to drive and deform a still face model?
  - How to retarget a motion sequence?
  - How to simulate non-linear *co-articulation effects* ?

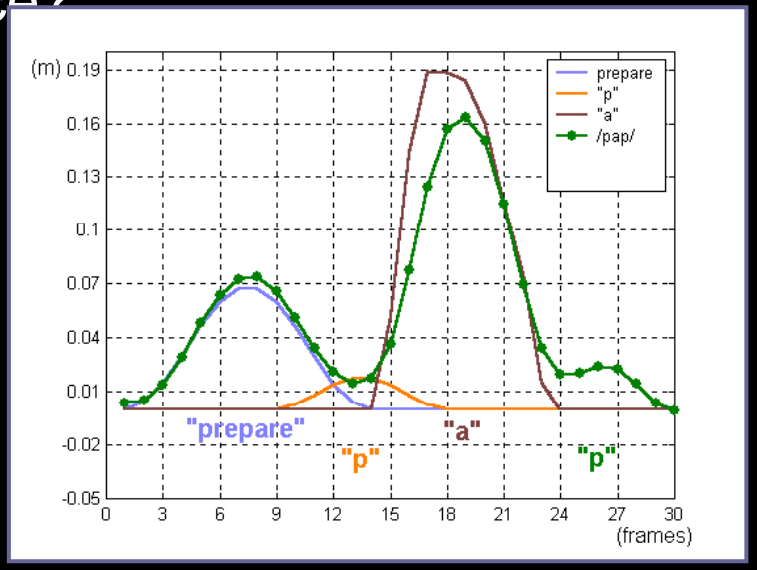
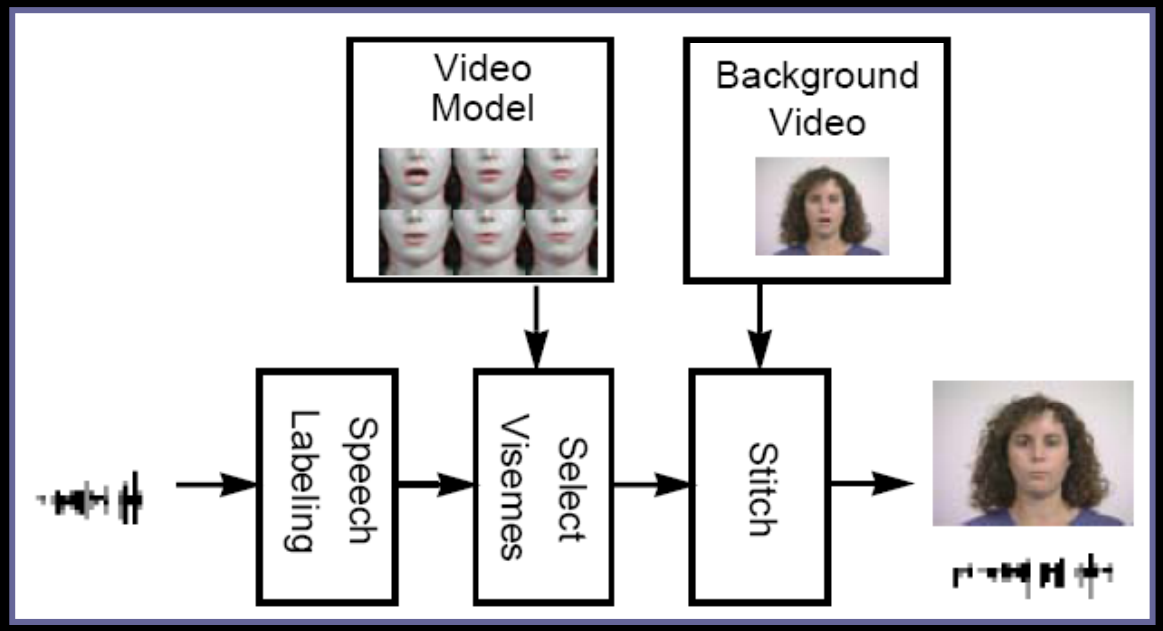


Image-based facial animation

- Parameterizing the input video sequences.
- Synthesizing animation by blending of image samples.
- With sufficient samples, the synthetic results are video-realistic.
  - Avoiding the problem of realistic rendering.
  - Limitation of lighting and view points.

Video rewrite

- C.Bregler et al., “Video Rewrite: Driving Visual Speech with Audio”, Proc. SIGGRAPH’97.
  - A speech-driven facial animation.



Video rewrite (cont.)

- From actual film footages and modify them to match the new speech.
  - In movie “Forest Gump”, using labor-intensive interaction.
  - Automatically piecing together from old footages.
- Idea: “Replace” the mouth according to phonemes.

… …

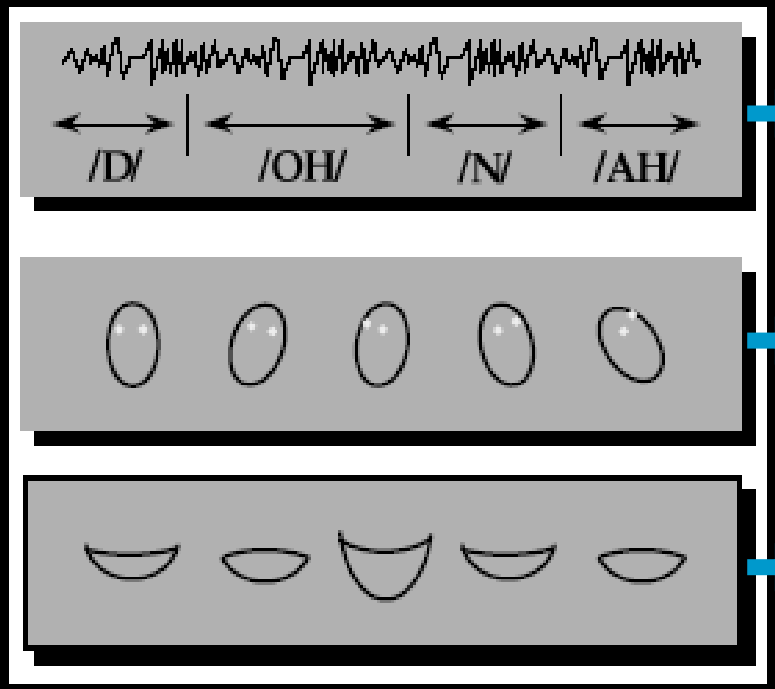
/ Y / /ɛ/ /s/

Analysis

Phoneme

Head pose

Mouth shape

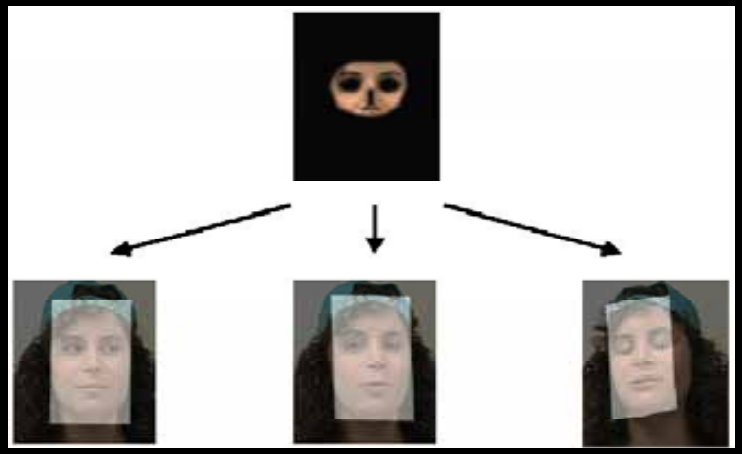


By speech recognition
(for viseme annotation)

By minimizing differences
(for global alignment)

By eigenpoints
(for local alignment)

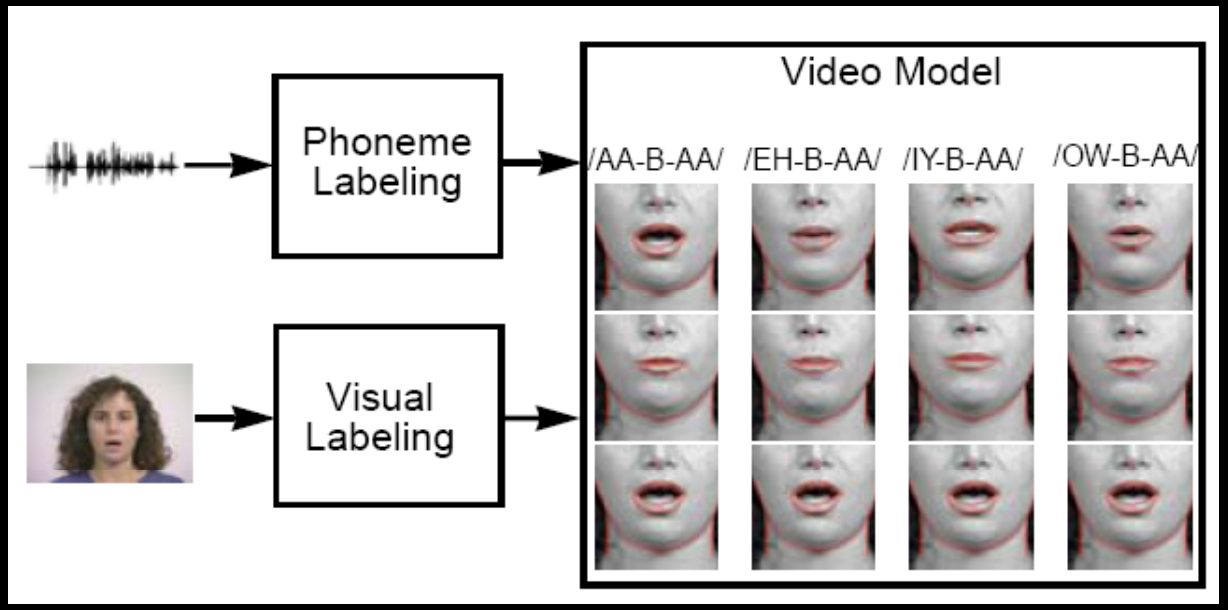
Head pose estimation



Triphone model

- To simulate co-articulation effects.
- HMM-based phonetic labelling.

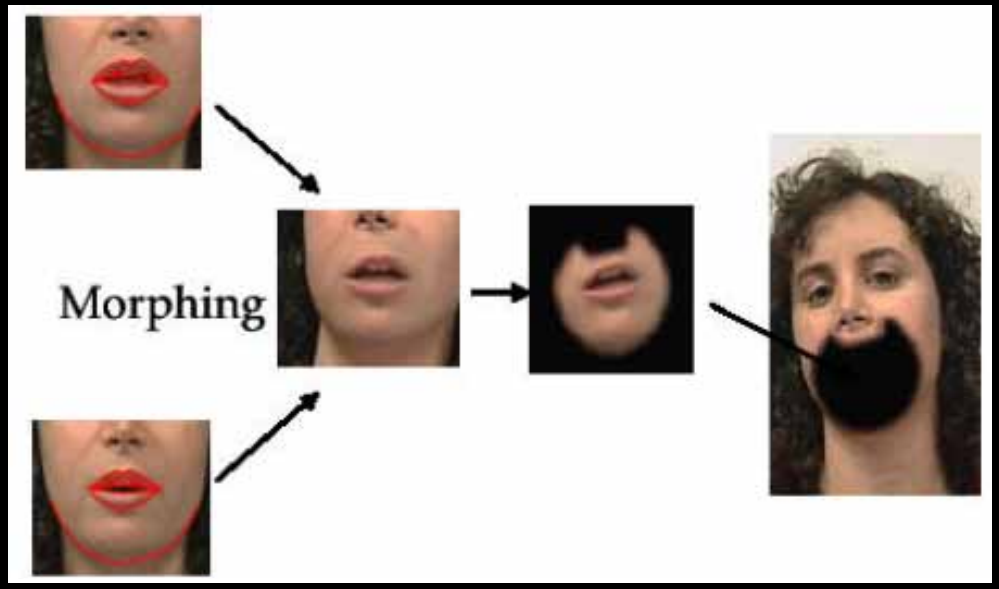
26 viseme classes.
Ten are consonant classes:/CH/, /K/, /T/….
Fifteen are vowel classes: /EH/, /ER/, /AH/
One class is for silence, /SIL/.



Synthesis

- Labeling the new soundtrack
- Selection of Triphone segments
  - Error $=\alpha D_p + (1 - \alpha) D_s$
    - D_p : phoneme-context distance
    - D_s : lip-shape distance
    - E.g. Teapot -> /T-IY-P/ + /IY-P-A/
    - Optimizing the overlapping (temporal offset and duration)
- Using dynamic programming to find the sequence of triphone segments

Synthesis (cont.)

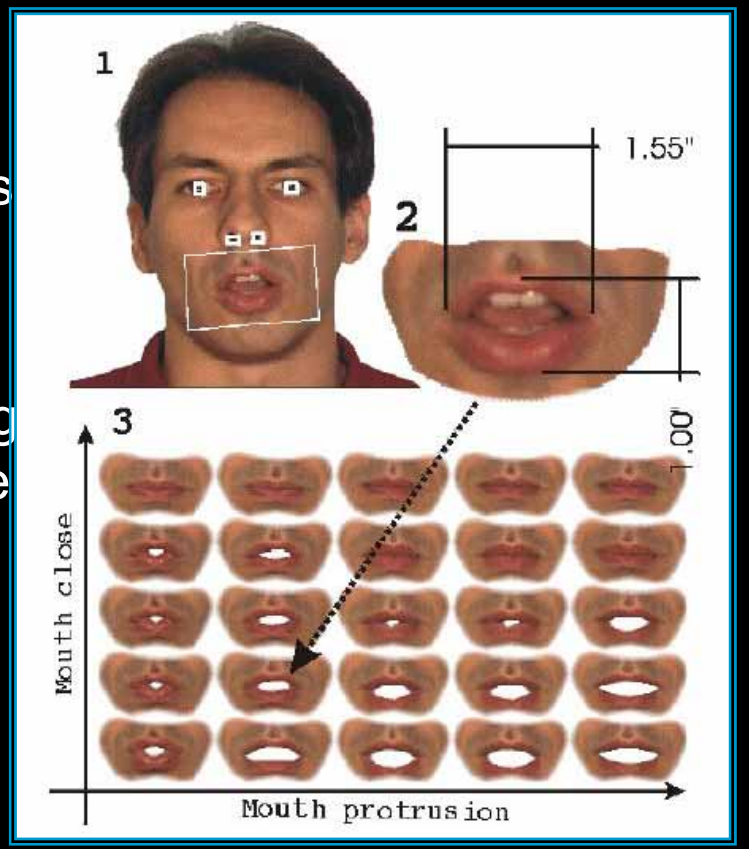


Results

- Video-realistic
  - 8 min of video (109 sentences)
  - 3500 triphone videos
- Realism depends on image samples.
  - Require a large sample data.
- Not transferable.
- Limited view directions and fixed lighting.

Improvement in data size

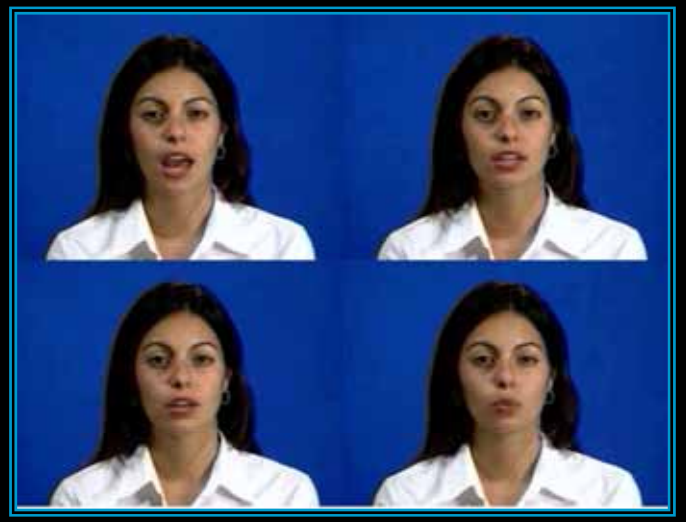
- E. Cosatto, H.P. Graf, "Sample-based synthesis of photo-realistic talking heads", Proc. Computer Animation ‘98, pp. 103-110.
- Further dividing samples into regions
- Instead of the triphone model, finding trajectory in the sample image space



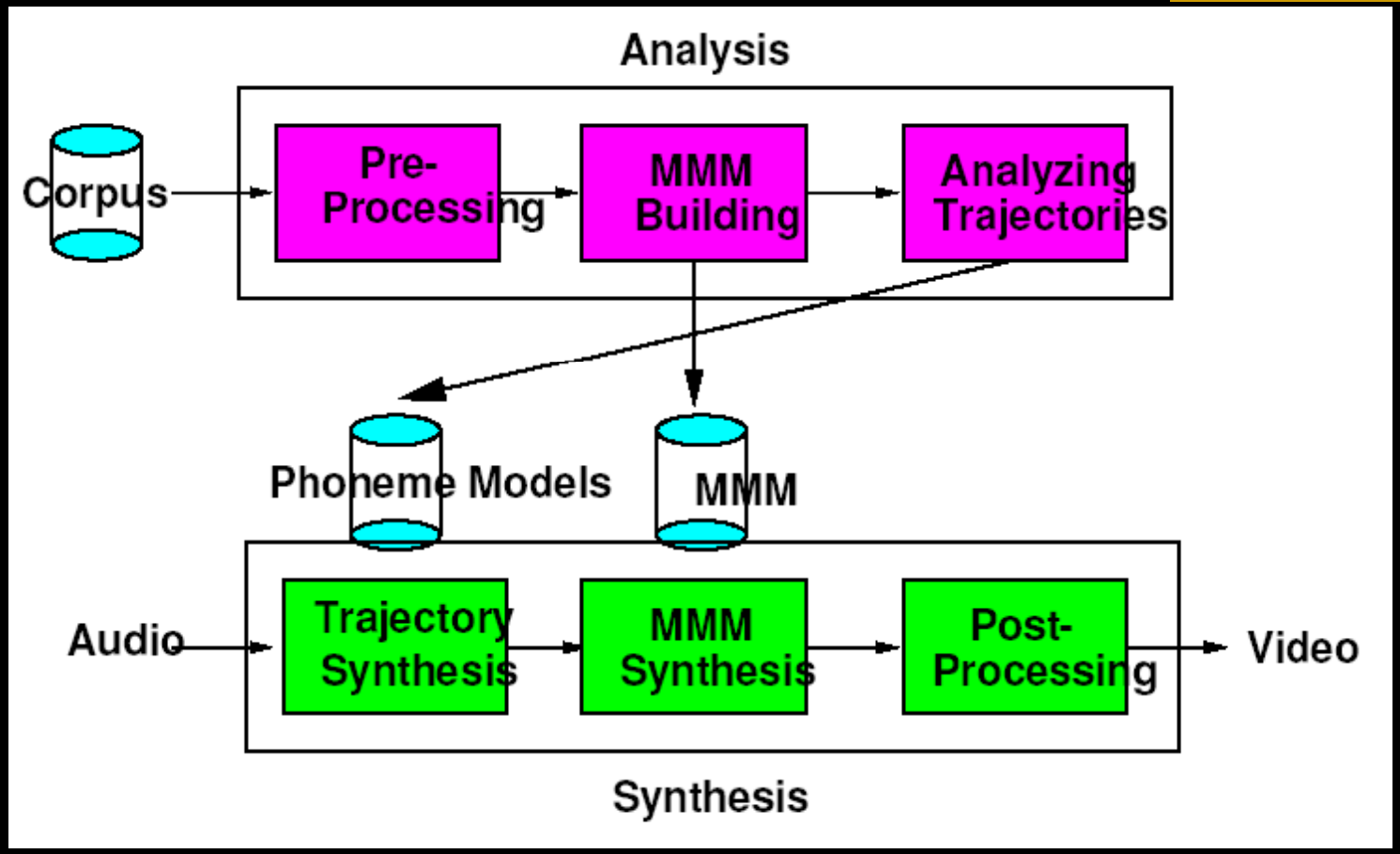
Trainable videorealistic speech animation

- T. Ezzat et al., “Trainable Videorealistic Speech Animation”, Proc. SIGGRAPH’02.

- Instead of image sample sequences, using a multidimensional morphable model (MMM).
  - Blending image prototypes.
    - No more large data.
    - Synthesizing novel expressions
  - Trajectories of optical flows



Overview


Analysis
Corpus
Pre-Processing
MMM Building
Analyzing Trajectories
Phoneme Models
MMM
Audio
Trajectory Synthesis
MMM Synthesis
Post-Processing
Video
Synthesis

Corpus record

- 640x480-pixel^2 resolution; NTSC 29.97 fps.
- 15-min corpus record (about 30000 frames).

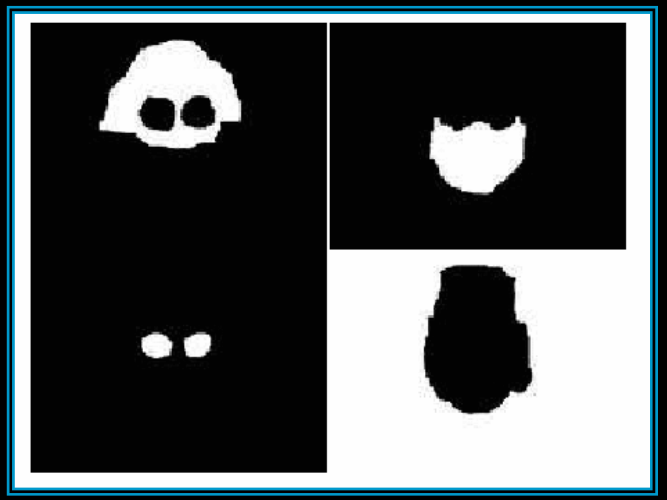
- 152 1-syllable words (e.g. “bed”).
- 156 2-syllable words (e.g. “dagger”).
- 105 short sentences.

- Uttering all sentences in a neutral expression.

Preprocessing

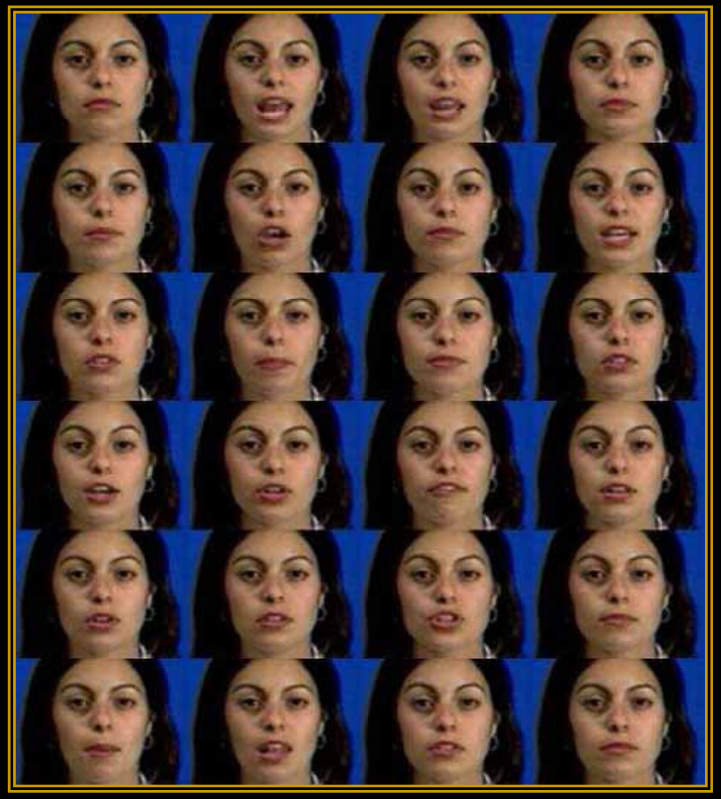
- Alignment of phonemes in the audio by the *CMU Sphinx system*.
- Face segmentation by masks.
- Normalization – removal of head movement.
  - Assume that a face is planar and estimate perspective plane movement.

The head, mouth, eye, background masks



Multidimensional morpable model

- Combining image prototypes to synthesize new, unseen lip configuration.
- Separating appearance variation and shape variation of a mouth.
  - 46 prototype images
  - 46 optical flow correspondences



24 of the 46 image prototypes

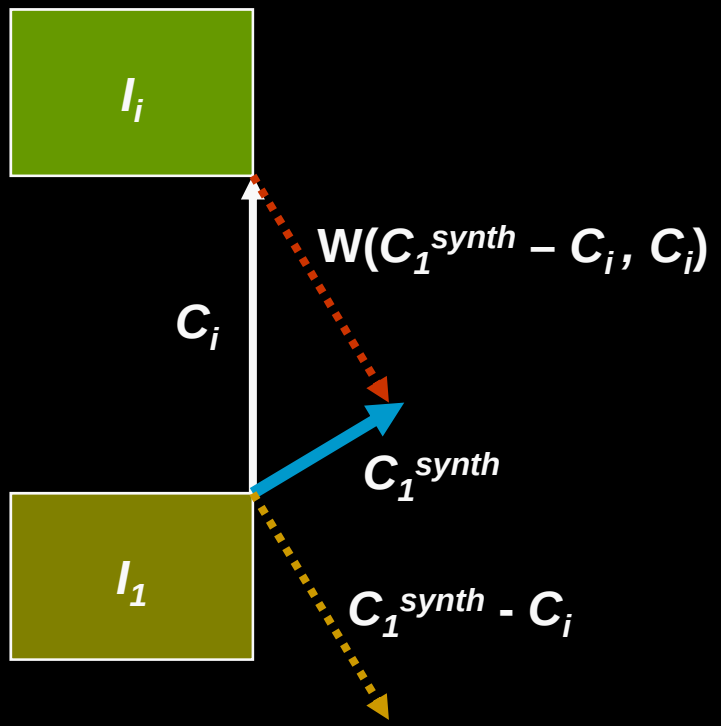
The top left frame is the reference image

Prototype images

- Multidimensional parameter space (α, β)
  - α : 46-dim. vectors as weights for mouth shape (optical flow).
  - β : 46-dim. vectors as weights for mouth texture.
- How to find the prototype images?
  - First, using EM-PCA to reduce dimensions. (for memory and computation efficiency)
  - Using k-means clustering to select prototype images.
  - Using flow concatenation (indirect optical flow) for correspondences.

Synthesis in MMM

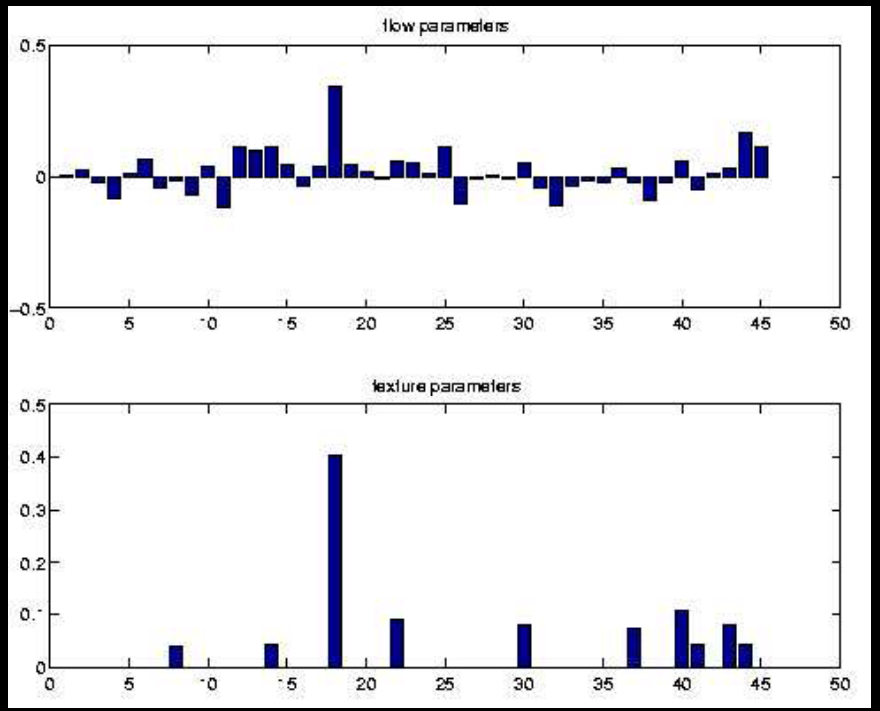
- Prototype flow C_i; synthetic flow C_i^{synth} (compared to C_i):
  - $C_1^{synth} = \Sigma\ \alpha_i C_i$
  - $C_i^{synth} = W(C_1^{synth} - C_i,\ C_i)$
  - $I_i^{warped} = W(I_i,\ C_i^{synth})$



- Synthesizing a new face:
  - $I^{morph}(\alpha,\beta) = \Sigma\ \beta_i\ I_i^{warped}$

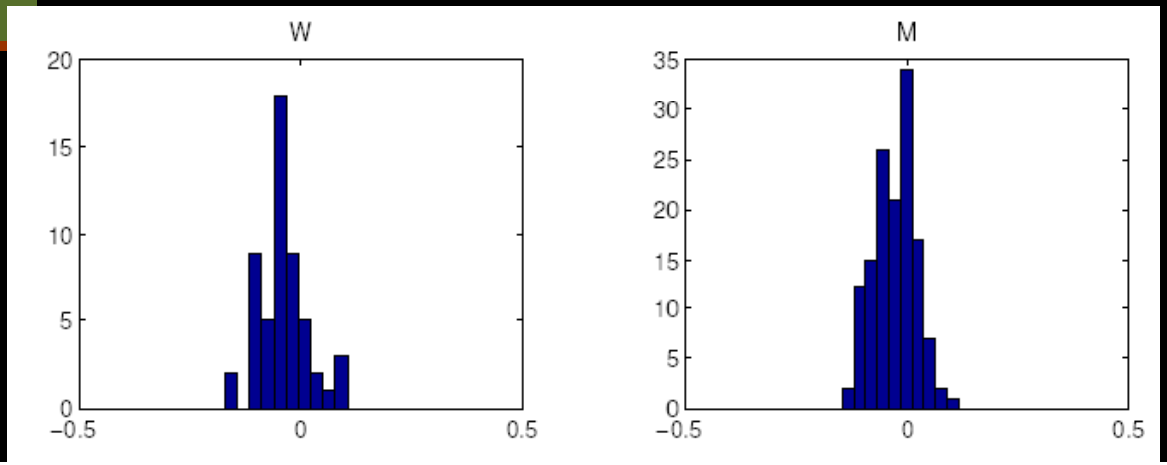
Analysis in MMM

- Analysis of (α,β) from a novel image.
- Analysis-by-synthesis and gradient-descent-based minimization.
  - Analyzing optical flow first:
    $\alpha = \arg\min \| C_{novel} - \Sigma \alpha_i C_i \|$.
  - and then texture:
    $\beta = \arg\min \| I_{novel} - \Sigma \beta_i I_i^{warp} \|$
    where $\beta_i > 0$ and $\text{sum}(\beta_i) = 1$.



Trajectory synthesis

- To map from a phone stream $\{p_t\}$ to a trajectory
  $y^t = (\alpha^t, \beta^t)$.
  "one"=(\w\,\w\,\w\,\w\, \uh\,\uh\,\uh\,\uh\,\uh\,uh\, \n\,\n\,\n\,\n\,\n\)
- Representing each phoneme *p* as a multidimensional Gaussian $(\mu_p, \sum_p)$.
  - Estimating Gaussian parameters for flow and texture separately.



Histograms of the α_1 *parameter for \w\ and \m\ .*

Trajectory synthesis (cont.)

- Synthesizing a trajectory by minimizing an *objective function E:*

$$E = \underline{(y-\mu)^T D^T \Sigma^{-1} D(y-\mu)} + \underline{\lambda\, y^T W^T W y}.$$

target term *smoothness*

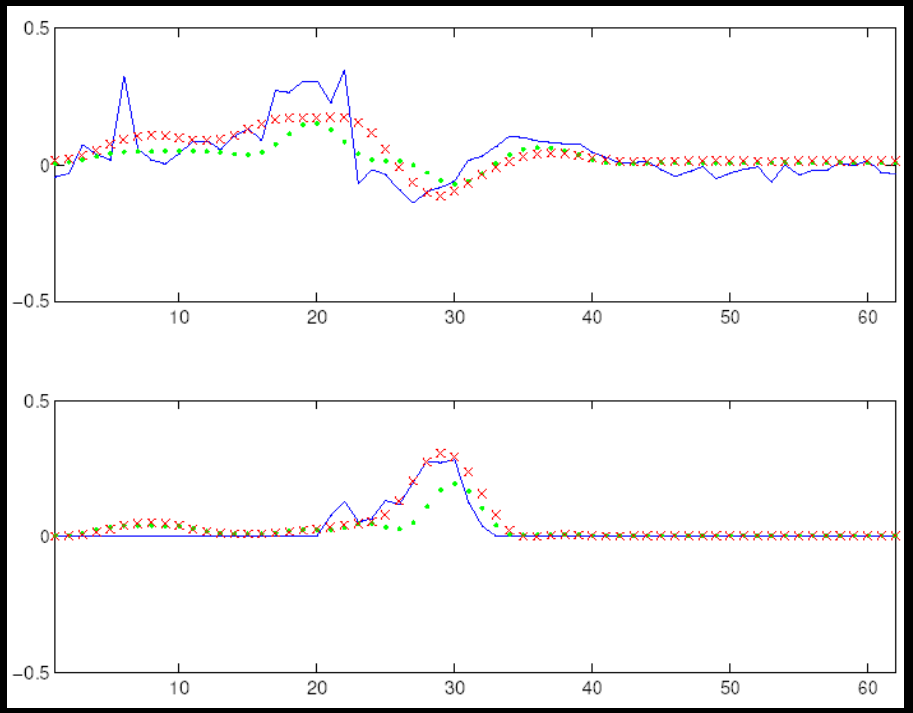
$$y = \begin{bmatrix} y_t \\ \vdots \\ y_T \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_{P_t} \\ \vdots \\ \mu_{P_T} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{P_t} & & \\ & \ddots & \\ & & \Sigma_{P_T} \end{bmatrix}$$

$$D = \begin{bmatrix} \sqrt{I - \frac{D_{P_1}}{T}} & & & \\ & \sqrt{I - \frac{D_{P_2}}{T}} & & \\ & & \ddots & \\ & & & \sqrt{I - \frac{D_{P_T}}{T}} \end{bmatrix} \quad W = \begin{bmatrix} -I & I & & & \\ & -I & I & & \\ & & & \ddots & \\ & & & & -I \quad I \end{bmatrix}$$

Trajectory synthesis (cont.)

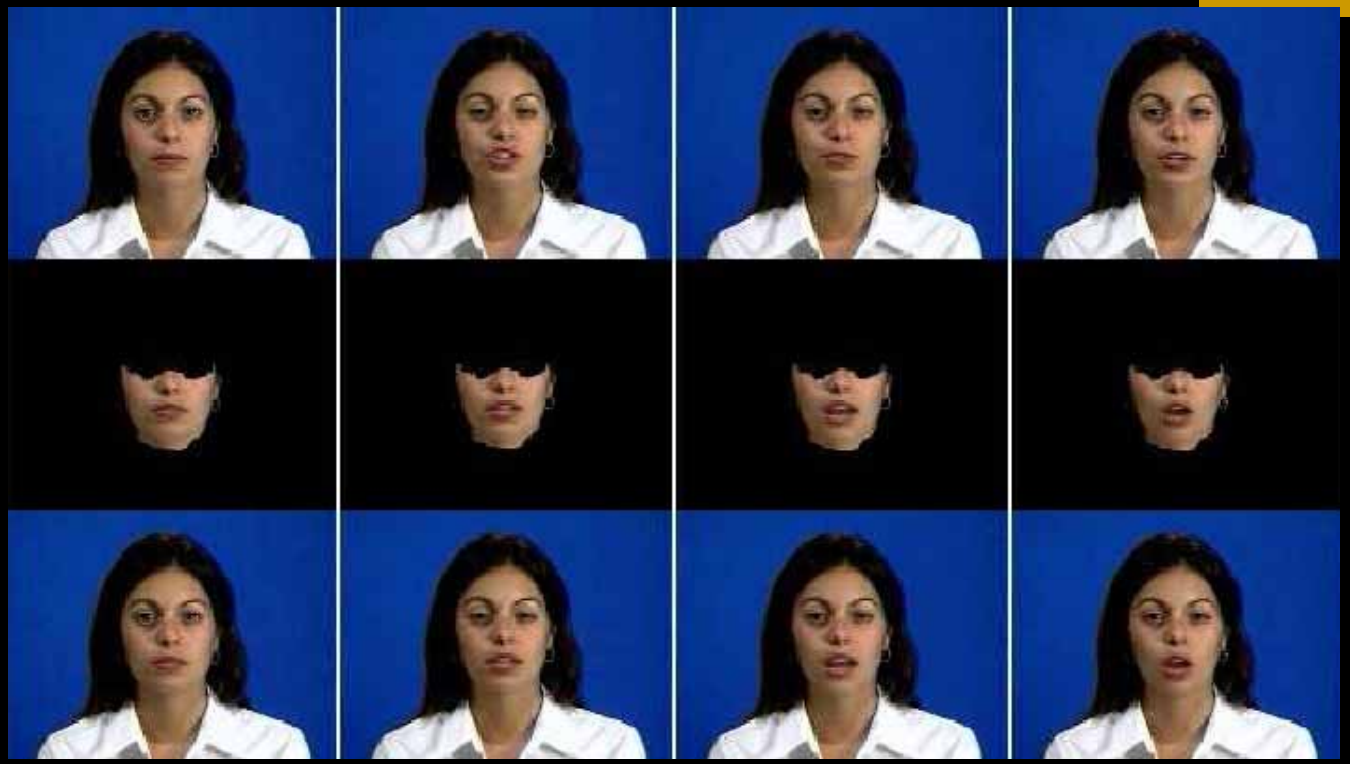
- Co-articulated phonetic dynamics emerge implicitly through the interplay between the magnitude of Σ_p for each phoneme.
- The μ_p and Σ_p are trained by *gradient descent learning.*

Min $E_2 = (z-y)^t(z-y)$



Trajectories of α_{12} and β_{28}. Blue: analyzed one; green: untrained one; red: trained one.

Post-processing



The background composition process.

Top: video clips.

Middle: a generated sequence.

Bottom: The final composite sequence.

Future work

- Limitation in head poses, viewpoints and lighting $\rightarrow$ Real-time 3D scanner.
- Faithfulness of MMM trajectory $\leftrightarrow$ “Voice puppetry” HMM trajectory (SIGGRAPH’99)
- Emotion dynamics.

The blurred tongue and teeth.

Top: original images. Bottom: synthetic images.

Free-Viewpoint Video of Human Actors

- J. Carranza et al., “Free-Viewpoint Video of Human Actors”, Proc. SIGGRAPH’03, pp.569-577.

- Input: multi-view synchronized video of an actor without marker.
- Interactive re-rendering the appearance from any viewpoint.

Free-Viewpoint Video of Human Actors (cont.)

- How to generate free-viewpoint video?
  - Stereo vision (triangulation)
    - The correspondence matching problem
  - Light fields
    - Dense views are necessary.
  - Visual hull techniques
    - The concave surface problem.
    - Dense views for a better result.
  - Geometry + Image based (hybrid) approach
    - Prior information about the target.

Strategies

- Goal: automatic and robust motion capture without markers.
  - Reconstruction based on silhouette images.
  - Prior information: a generic human body model.

  - Recording multi-view video

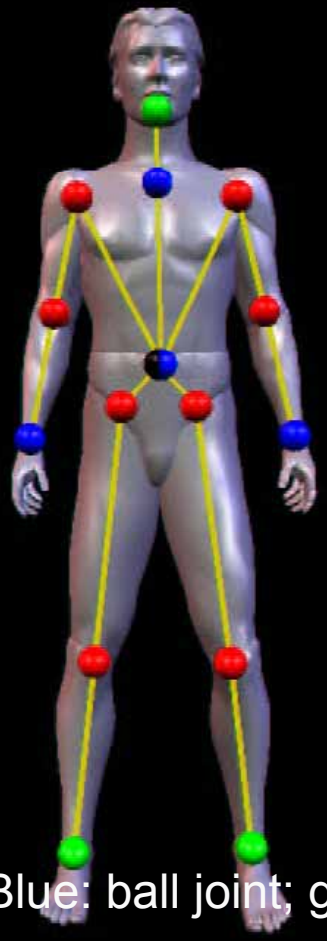
• Synchronization with external trigger.

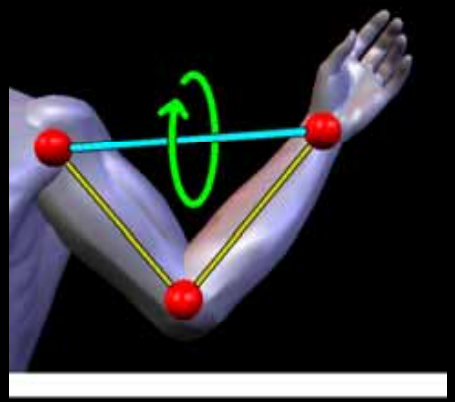
320x240 15fps or

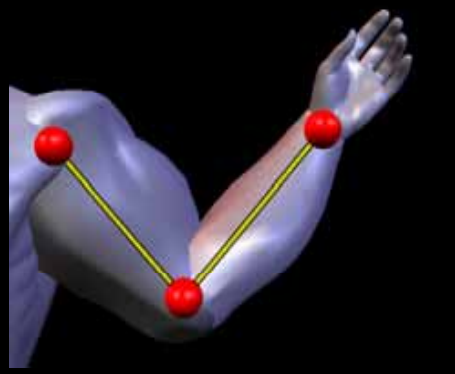
640x480 10fps

Strategies (cont.)

Silhouette image



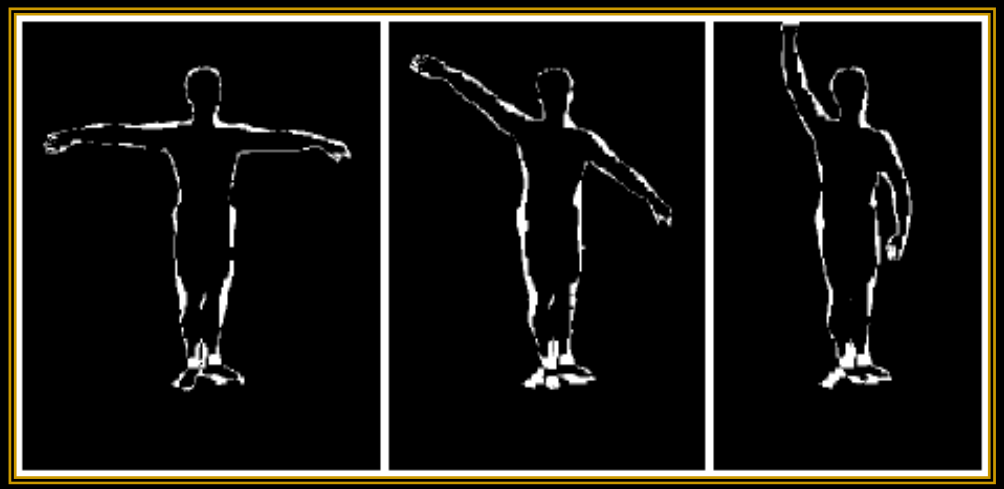




Blue: ball joint; green: hinge joint; red: 4DOF

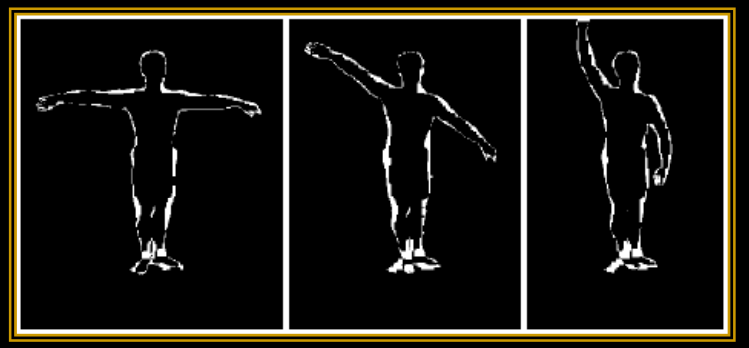
Marker-free motion capture

- Estimating the motion sequence by
  - Maximize the overlap between projected model silhouettes and input silhouette images.



Marker-free motion capture (cont.)

1. Initialization
   - Fitting global model position
   - Scaling the torso, limbs, etc.



2. Motion estimation
   - To avoid being trapped in local minima, adopt a sequence of optimizations.

   - Starting from the previous pose.
   - Estimating global T & R of the root.
   - Estimating rotations of the head and hip.
   - Independent optimizations of 2 arms and 2 legs.

Marker-free motion capture (cont.)

3. Adapting the model shape to silhouettes.

Texture generation

- Model silhouettes ≠ image silhouettes
- An intuitive solution: locally deform the model to fit the silhouette.
  - Noisy boundaries
- Alternative:
  - Remove one layer of boundary pixels.
  - Augment the foreground color.
  - Weighted combination of multi-view pixels according to view angles and visibility.

$$\omega_i' = \frac{1}{\left(\max_i(\omega_i) + 1 - \omega_i\right)^{\alpha}}$$

Texture generation (cont.)

Incorrect texture projection (shown in red) is solved through a modified visibility calculation.

Results