3D Shape Regression for Real-time Facial Animation

Chen Cao, Yanlin Weng, Stephen Lin, Kun Zhou

FaceWarehouse: a 3D facial Expression Database for Visual Computing

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, Kun Zhou

Zhejiang University MSRA

Presented by Shu Liang

(Black-on-white slides are Shu's)

Facial Animation

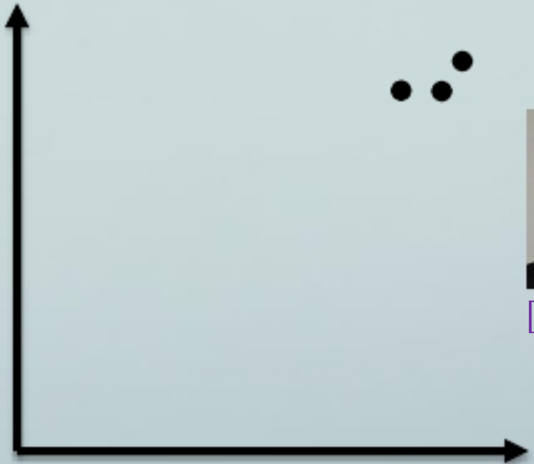
- Facial animation is widely used in films & games
- Performance-based facial animation

Avatar 2009

L.A. Noire 2011

Related Work

- Performance-based Facial Animation



Special Equipment

Facial Markers

[Huang et al. 2011]

Camera arrays

[Beeler et al. 2011]

Structured light

[Weise et al. 2009]

Related Work

- Performance-based Facial Animation

Single Camera

Optical Flow

[Vlasic et al. 2005]

ASM & AAM

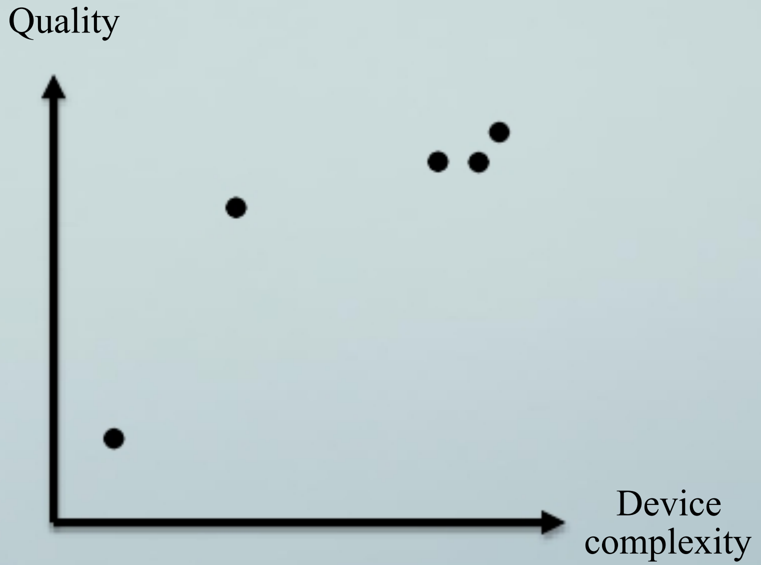
[Cootes et al. 1992-2001]

Regression-based alignment

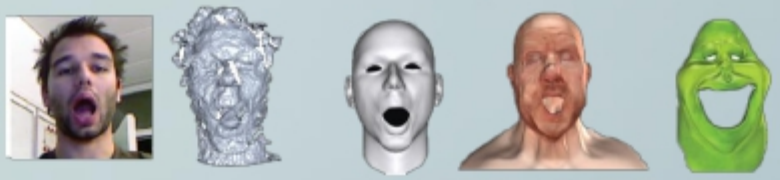
[Cao et al. 2012]

Related Work

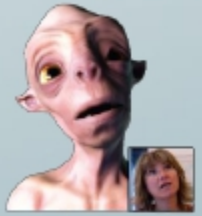
- Performance-based Facial Animation



Consumer RGBD Camera



[Weise et al. 2011]

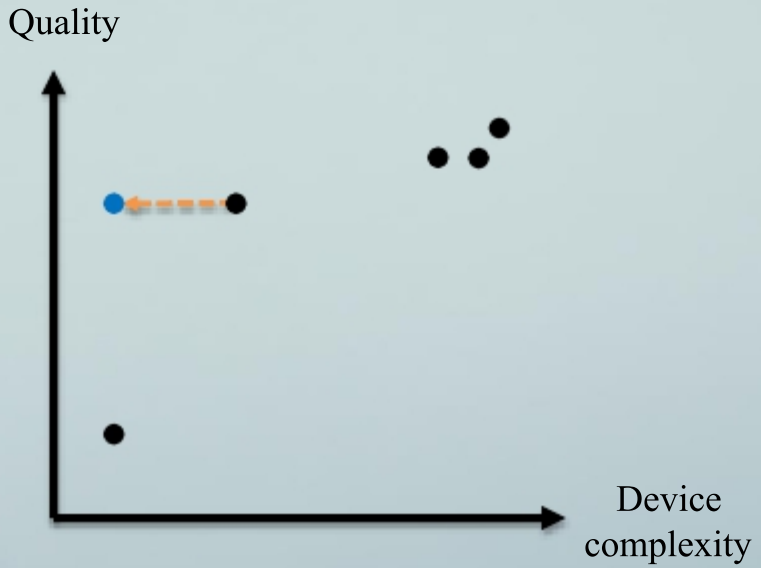


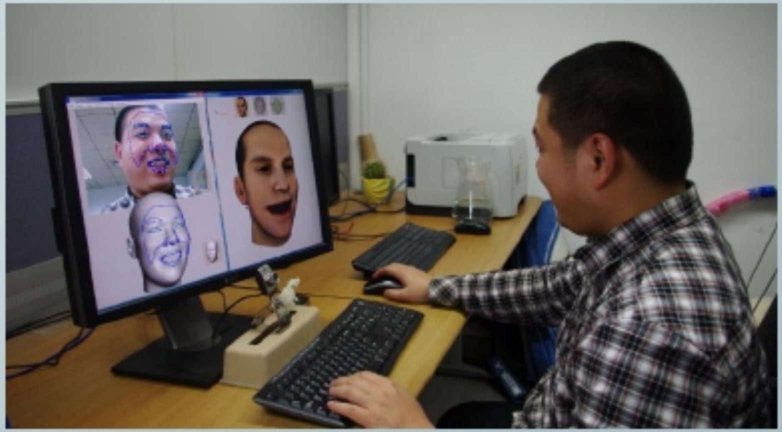
[Bouaziz et al. 2013]

[Li et al. 2013]

Our Goal

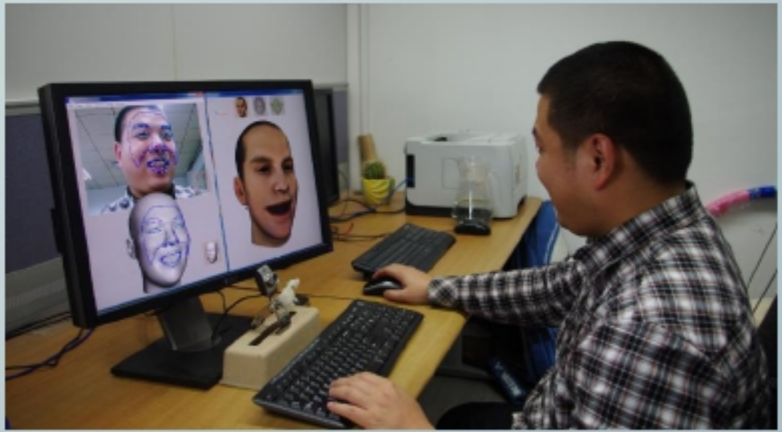
- Real-time facial animation for average users



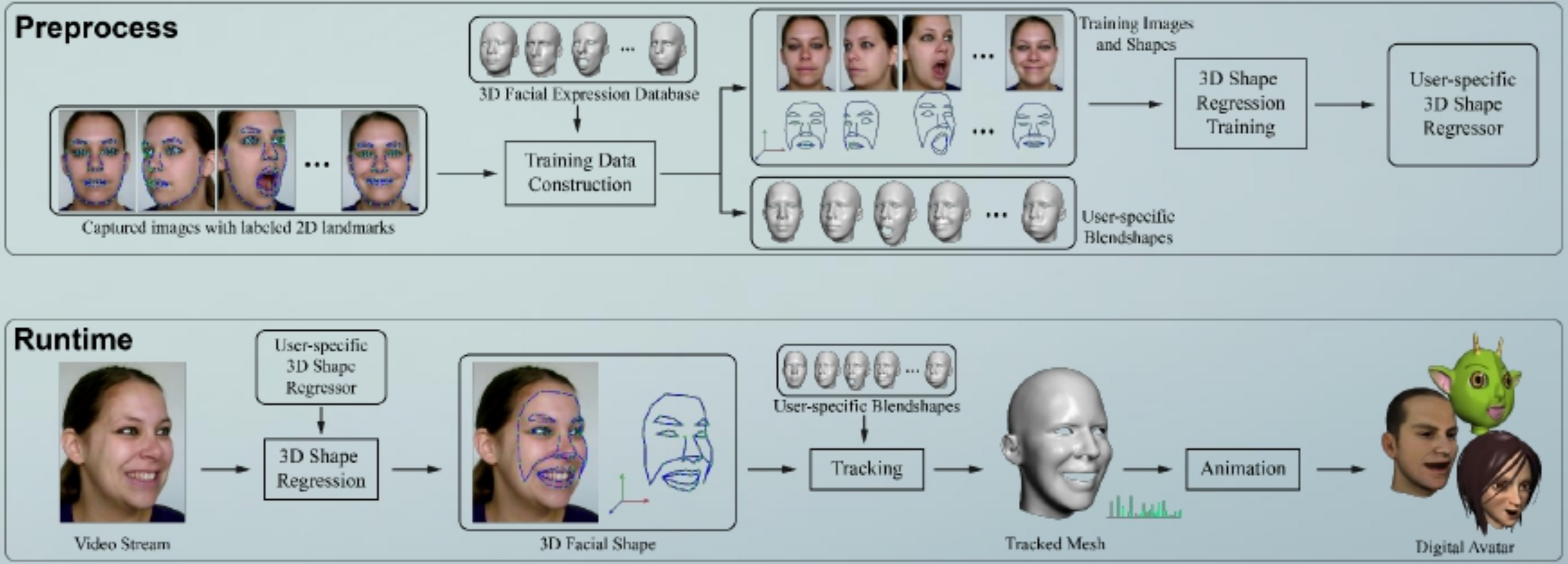


Our Goal

- Real-time facial animation for ordinary users
  - Single web camera
  - Robust
    - Fast motions
    - Large rotations
    - Exaggerated expressions
  - General environments
    - Indoors and outdoors
  - High performance
    - Mobile devices

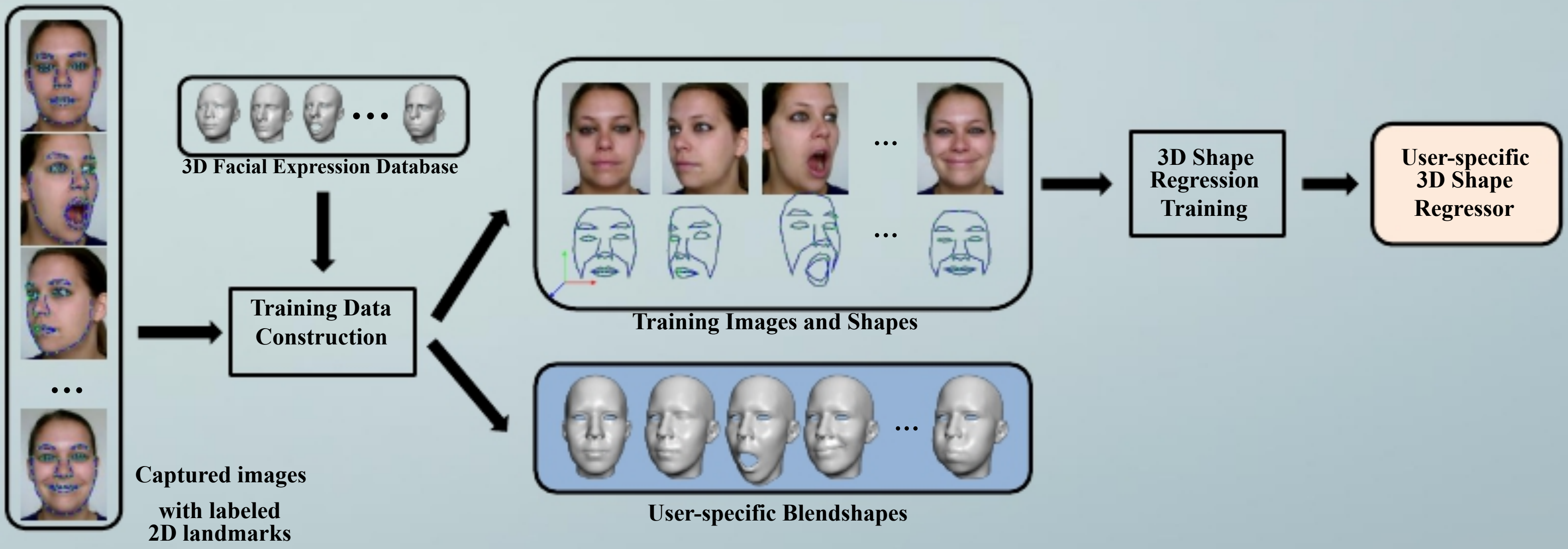


Our Pipeline



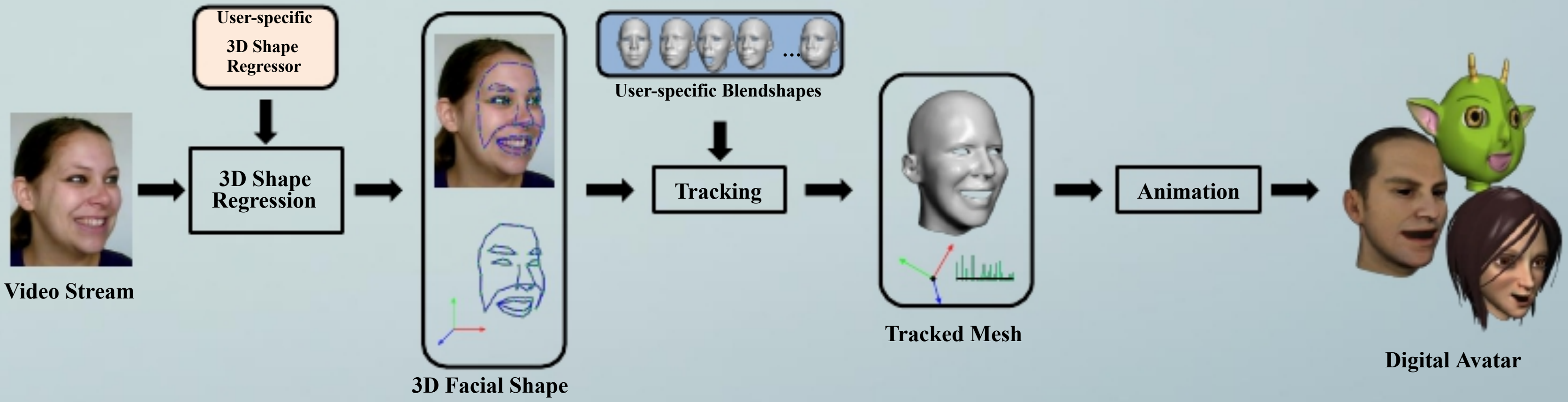
Our Pipeline

- One-time Preprocess



Our Pipeline

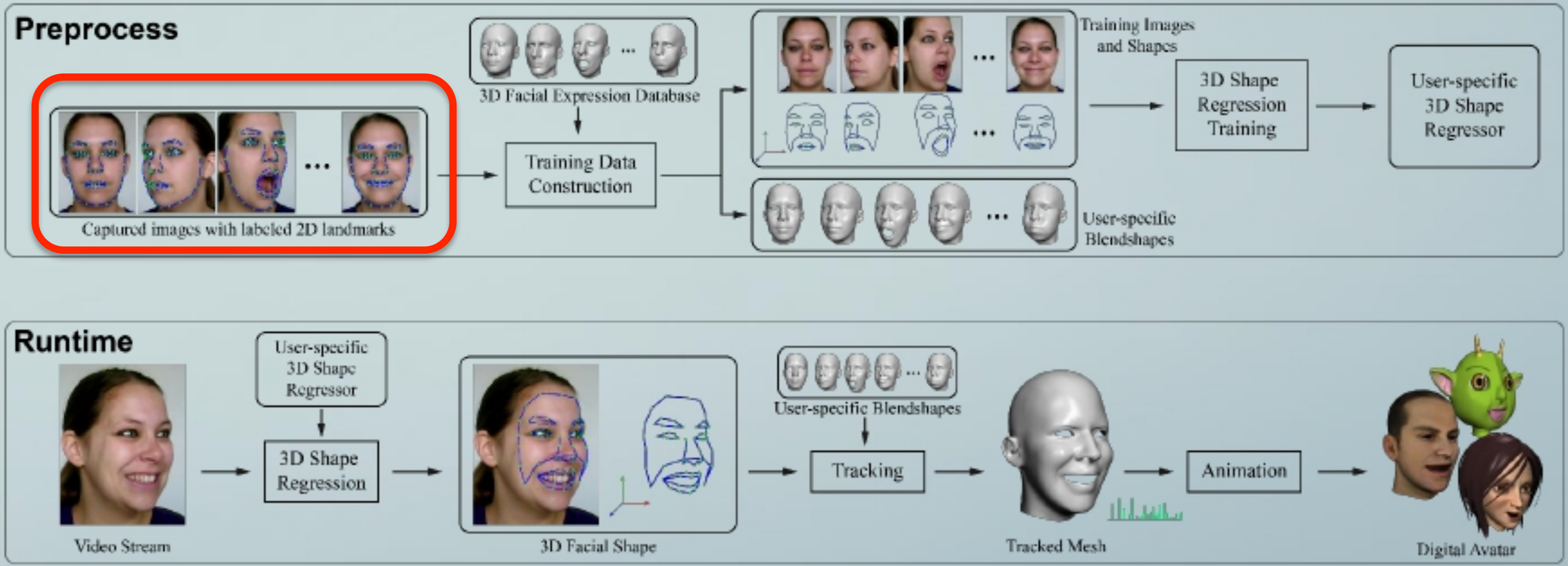
- Runtime computation



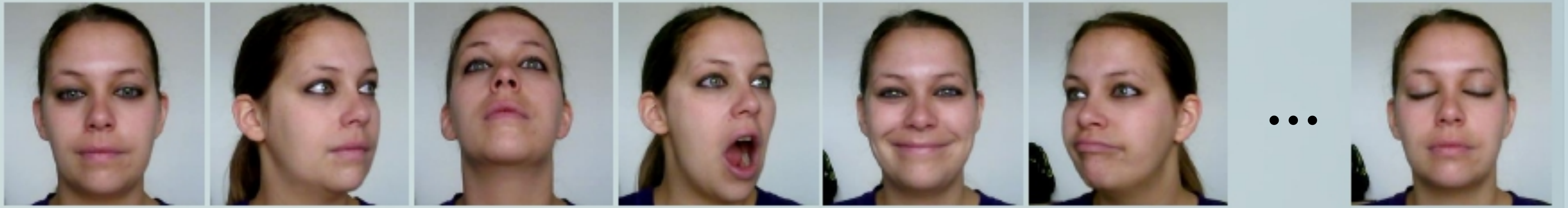
3D Face Shape Regression: Preprocess

- Data Collection
  - Image capturing & labeling
  - Blendshapes generation
  - Shape reconstruction
  - Training data generation
- Training

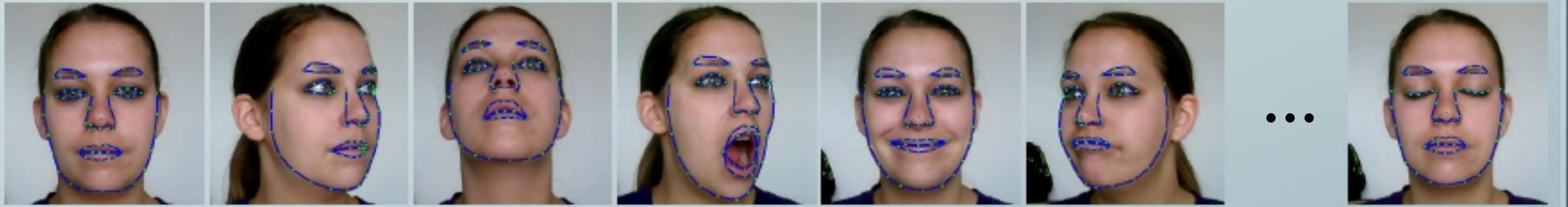
Our Pipeline



Preprocess: Image Capturing & Labeling



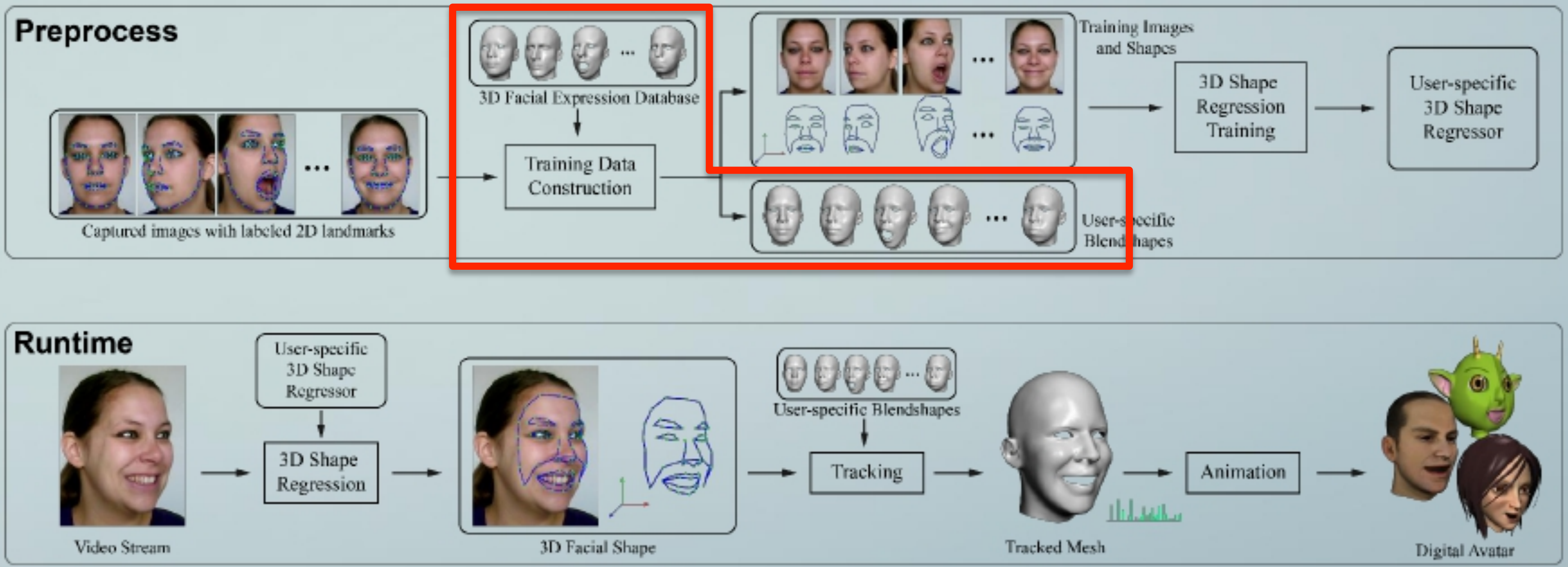
Captured Images



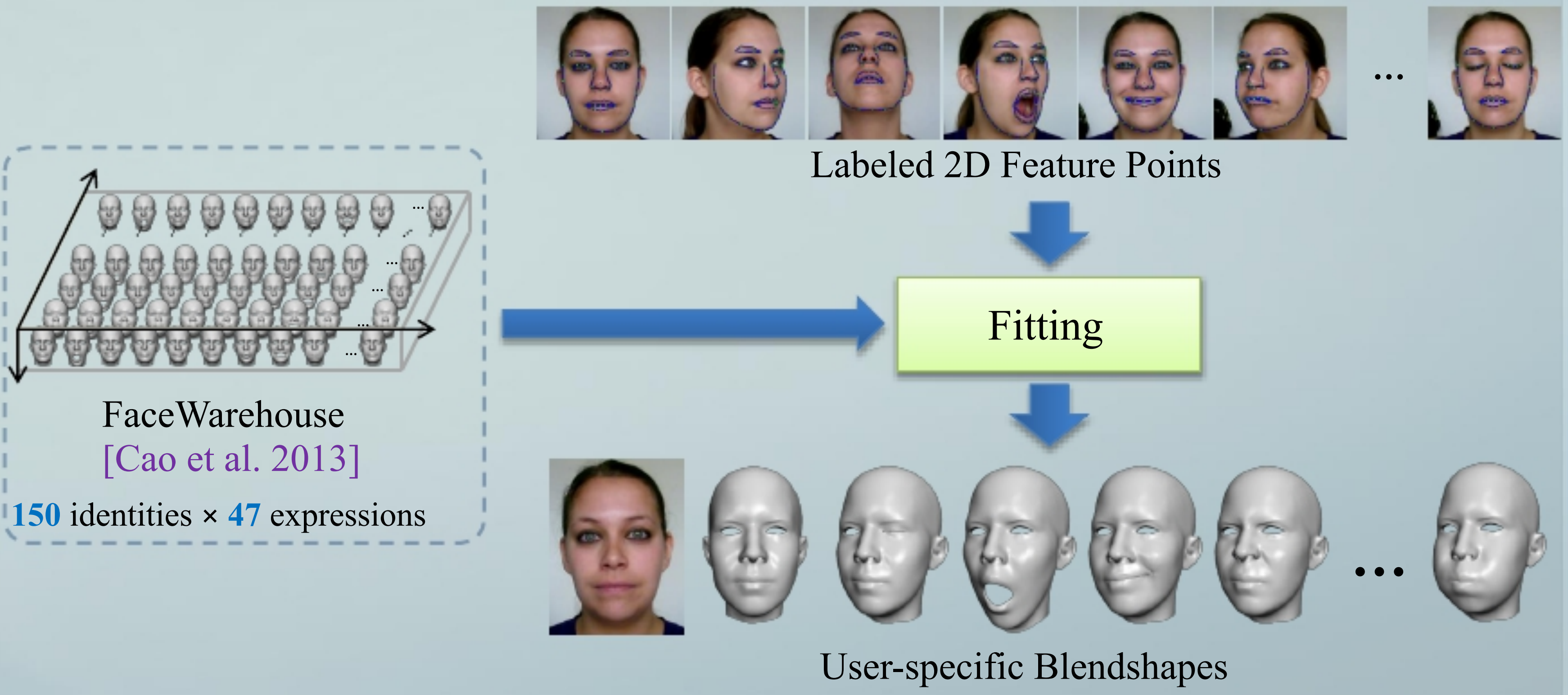
[Cao et al. 2012] + Manual Adjustment

Labeled 2D Feature Points

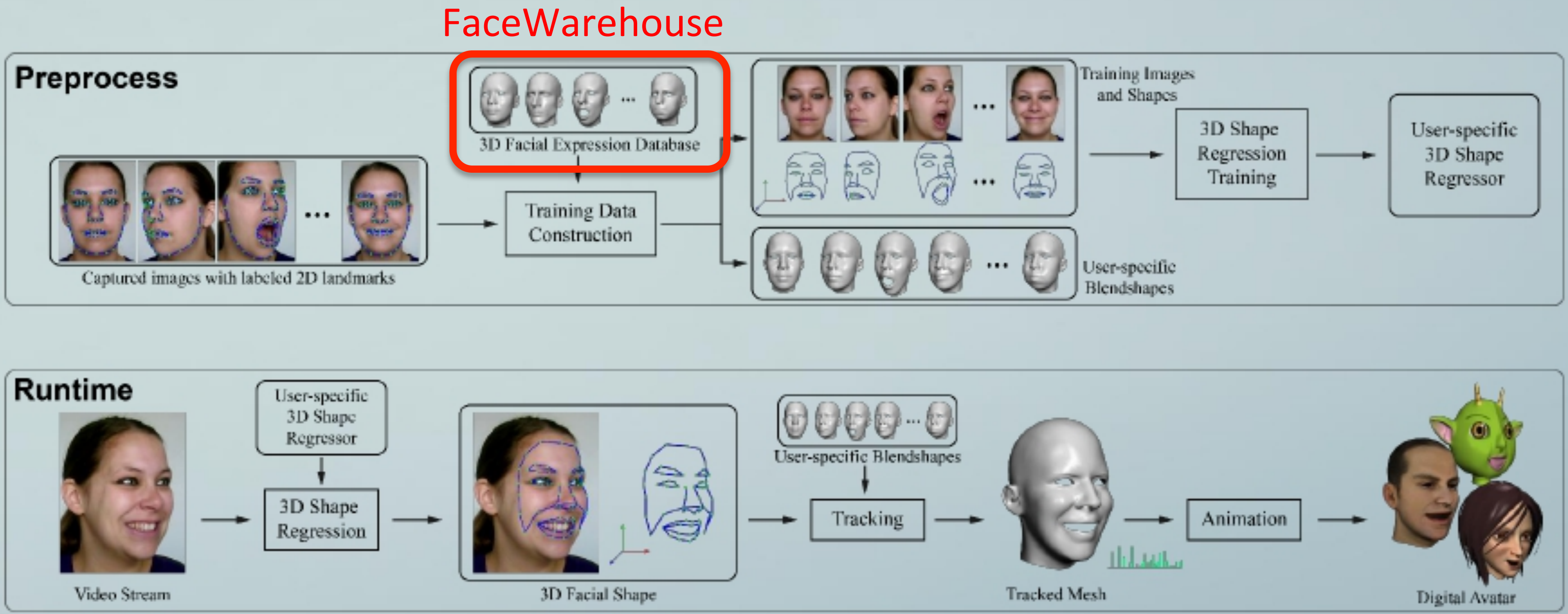
Our Pipeline



Preprocess: Blendshapes Generation

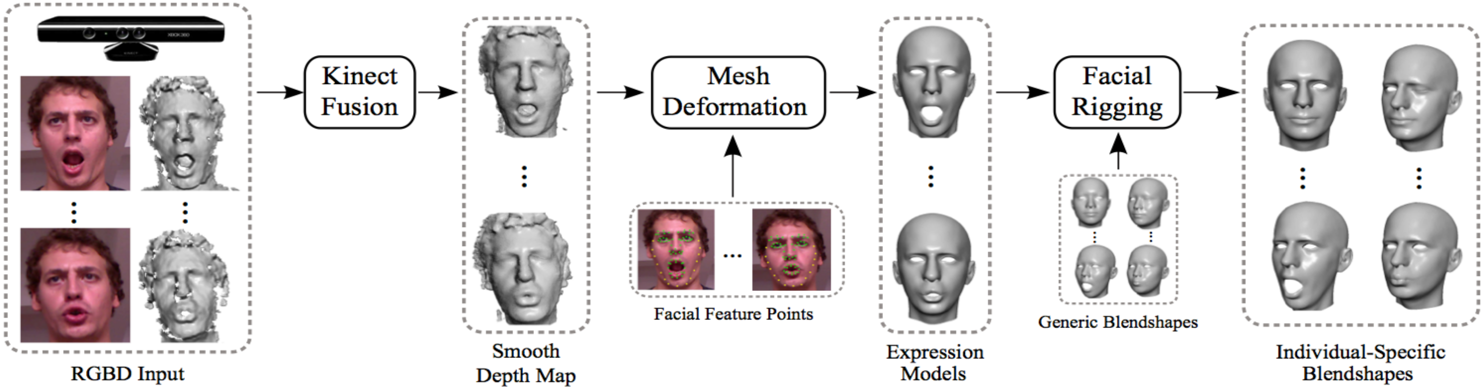


Our Pipeline



FaceWarehouse

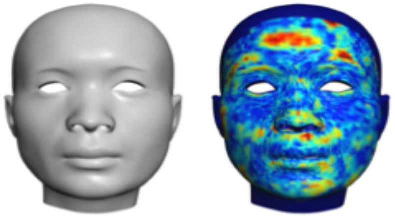
- RGBD images of 150 individuals captured by Kinect
- Aged 7-80 from various ethnic backgrounds
- Different expressions, one neutral and 19 other expressions.



FaceWarehouse

– Mesh deformation

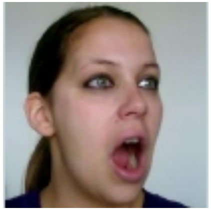
Neutral expression



[Blanz et. al 2004]

[Huang et. al 2006]

Other expressions



[Sumner et. al 2006] [Huang et. al 2006]

S_0, S_1, S_2 … S_{19} for 20 expressions.

FaceWarehouse

– Individual-specific expression blendshapes

- Example-based facial rigging algorithm:

An expression H of the person can be:

$$H = B_0 + \sum_{i=1}^{46} \alpha_i (B_i - B_0)$$

$\{B_1, B_2, \ldots B_{46}\}$ 46 FACS blendshapes

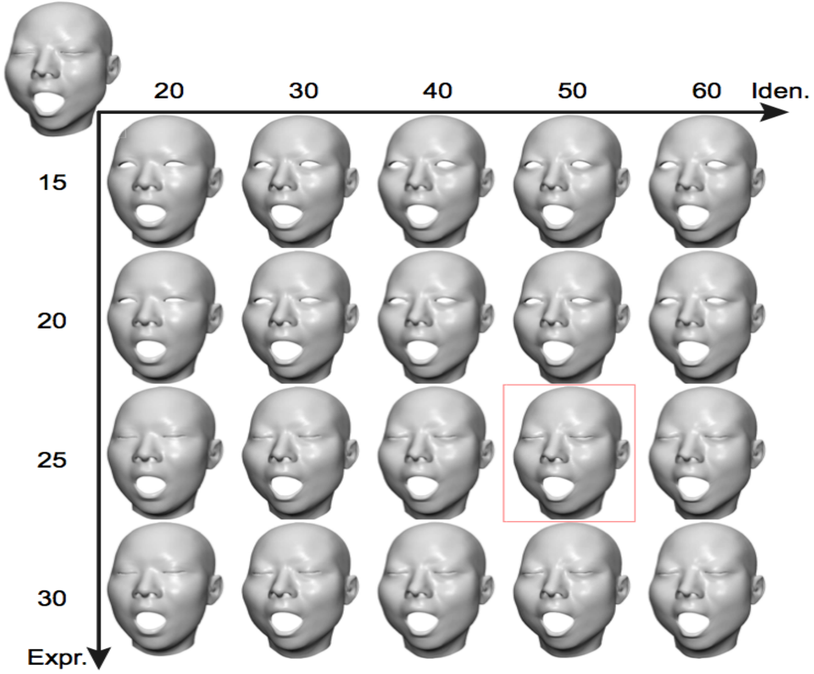
- Begins with a generic blendshape model $\mathbf{A} = \{A_0, A_1, \ldots, A_{46}\}$
- Optimized by minimizing the difference between S_j and linear combination of B_i with known weight for expression j, the difference between the deformation from B_0 to B_i and that from A_0 to A_i.

FaceWarehouse

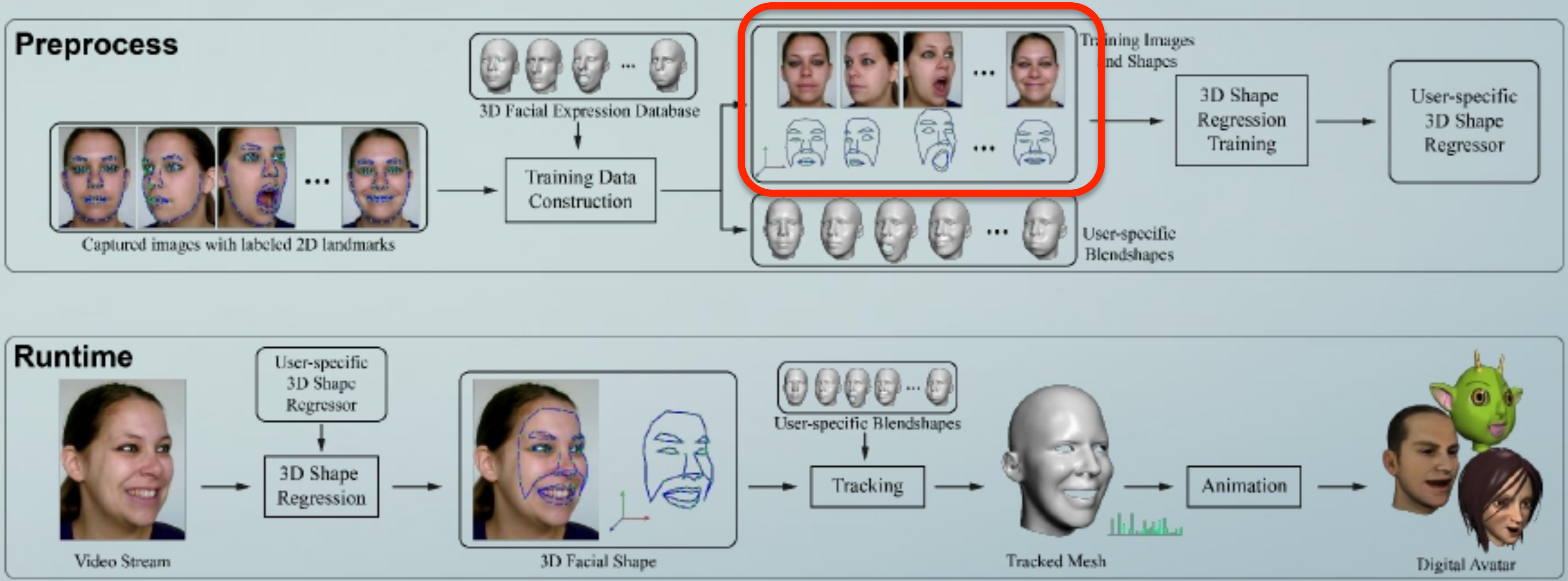
– Bilinear face model

A rank-three data tensor T. (11K vertices × 150 identities × 47 expressions）
Used N-mode SVD to decompose the tensor.

$$V = C_r \times_2 \mathbf{w}_{id}^T \times_3 \mathbf{w}_{exp}^T,$$



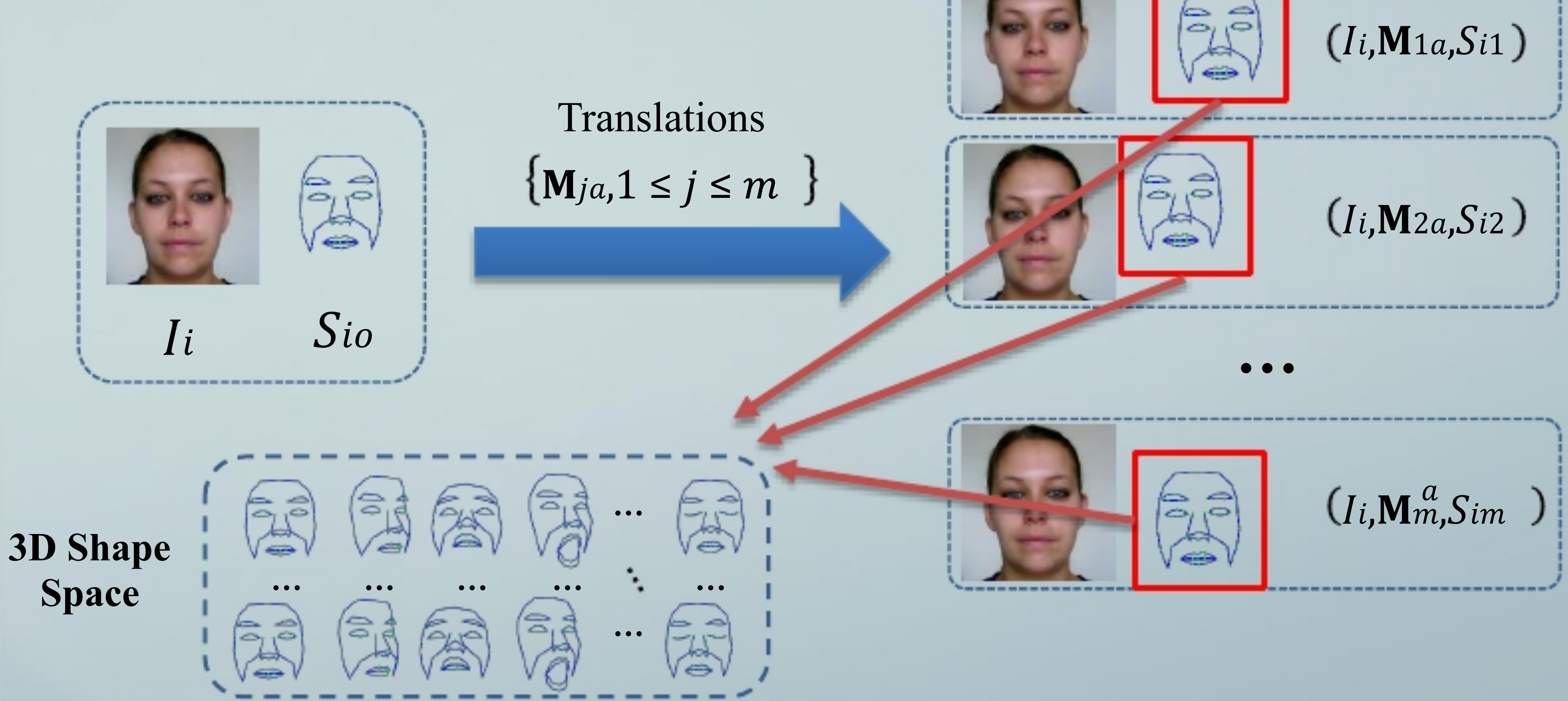
Our Pipeline



Preprocess: 3D Shape Reconstruction

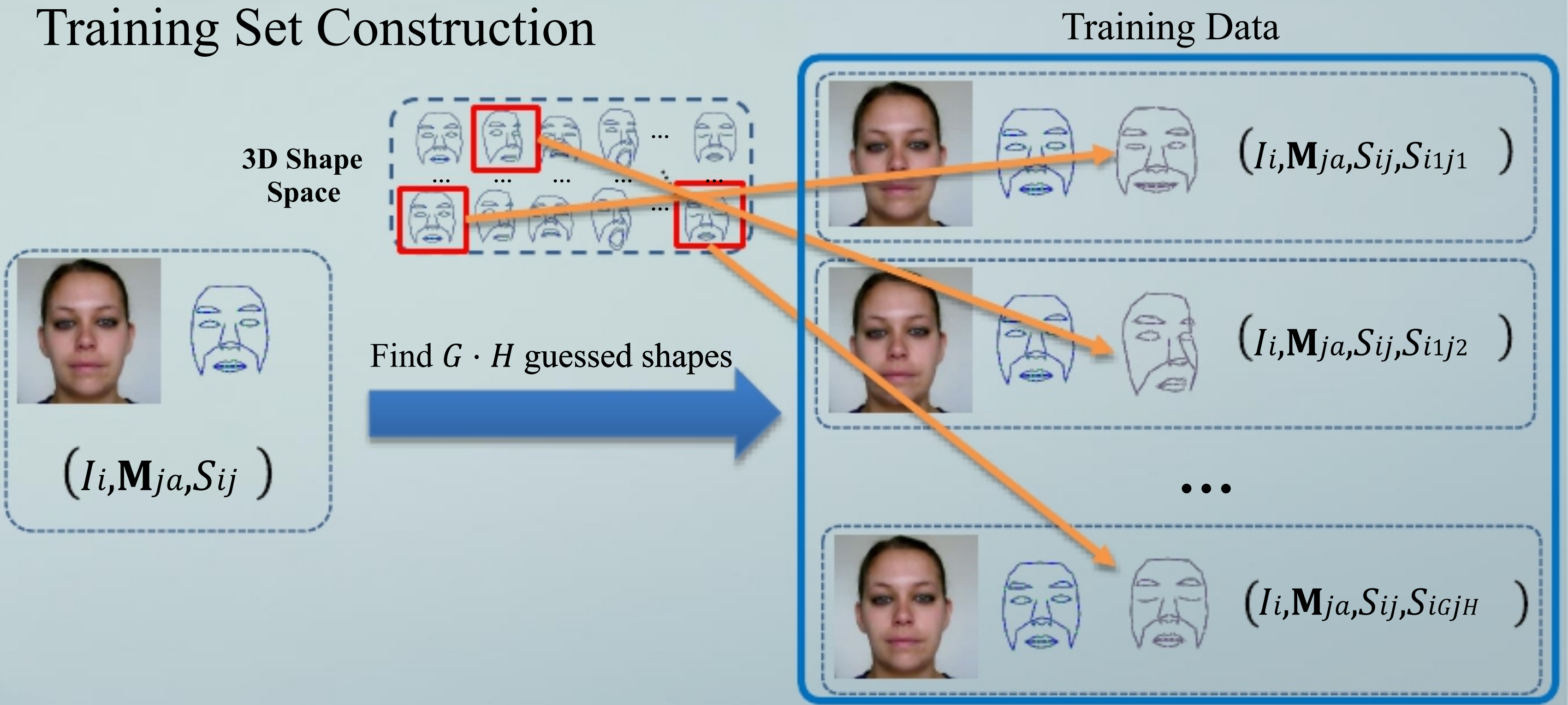
Preprocess: Training Data Generation

- Data Augmentation

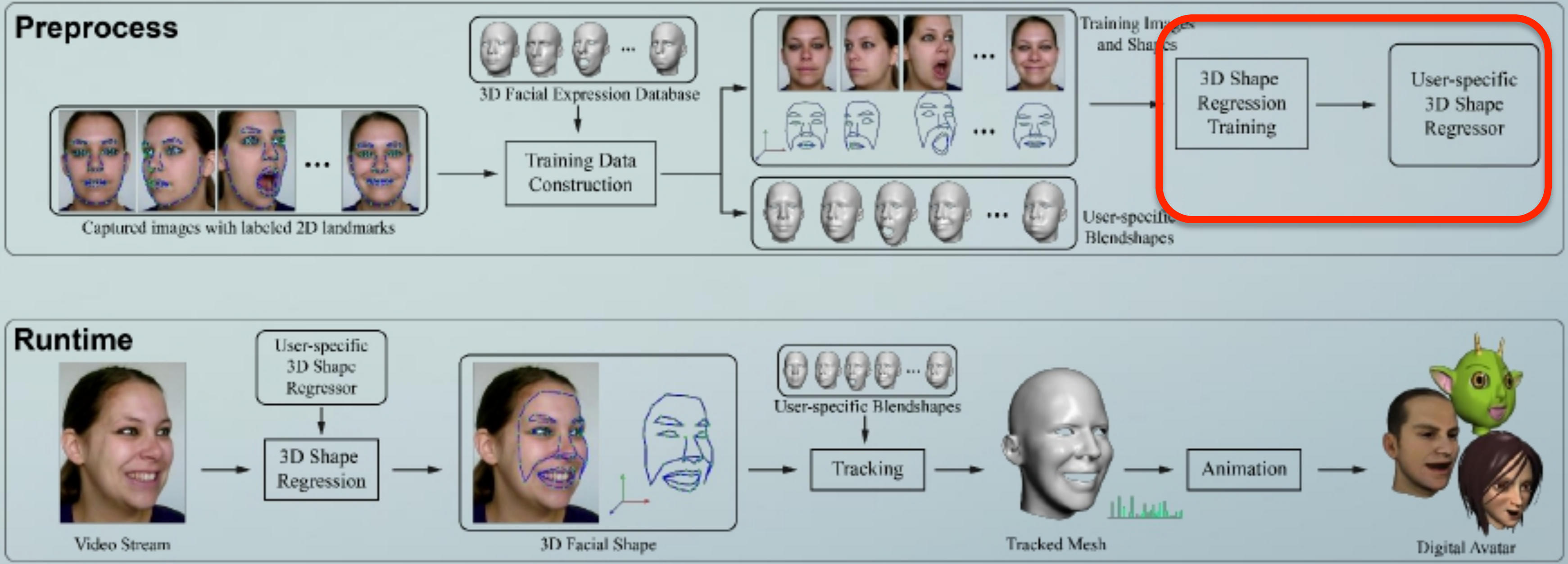


Preprocess: Training Data Generation

- Training Set Construction



Our Pipeline



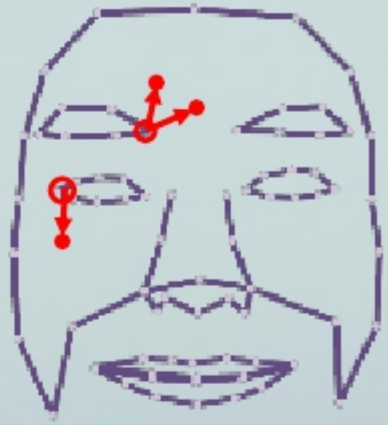
Preprocess: Training

- Appearance vector
- Primitive regressor: fern

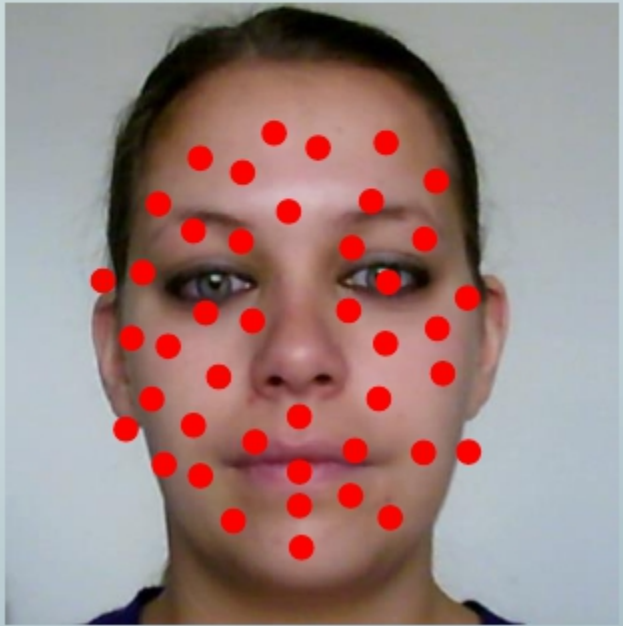
- Summary: two-level boosted regressor

Preprocess: Training

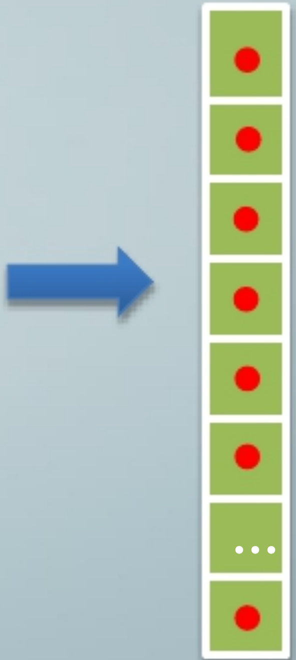
- Appearance vector



S_{ic}



$(I_i, \mathbf{M}_{ia})$



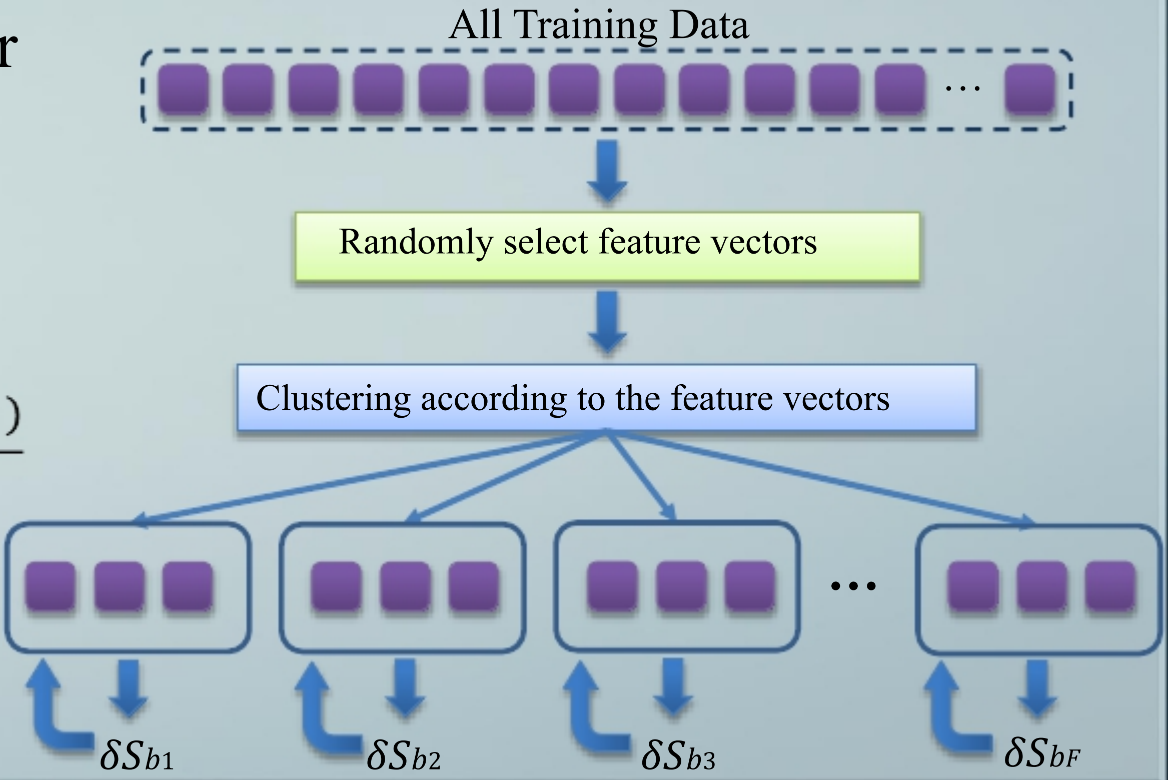
Appearance Vector

Preprocess: Training

- Primitive regressor

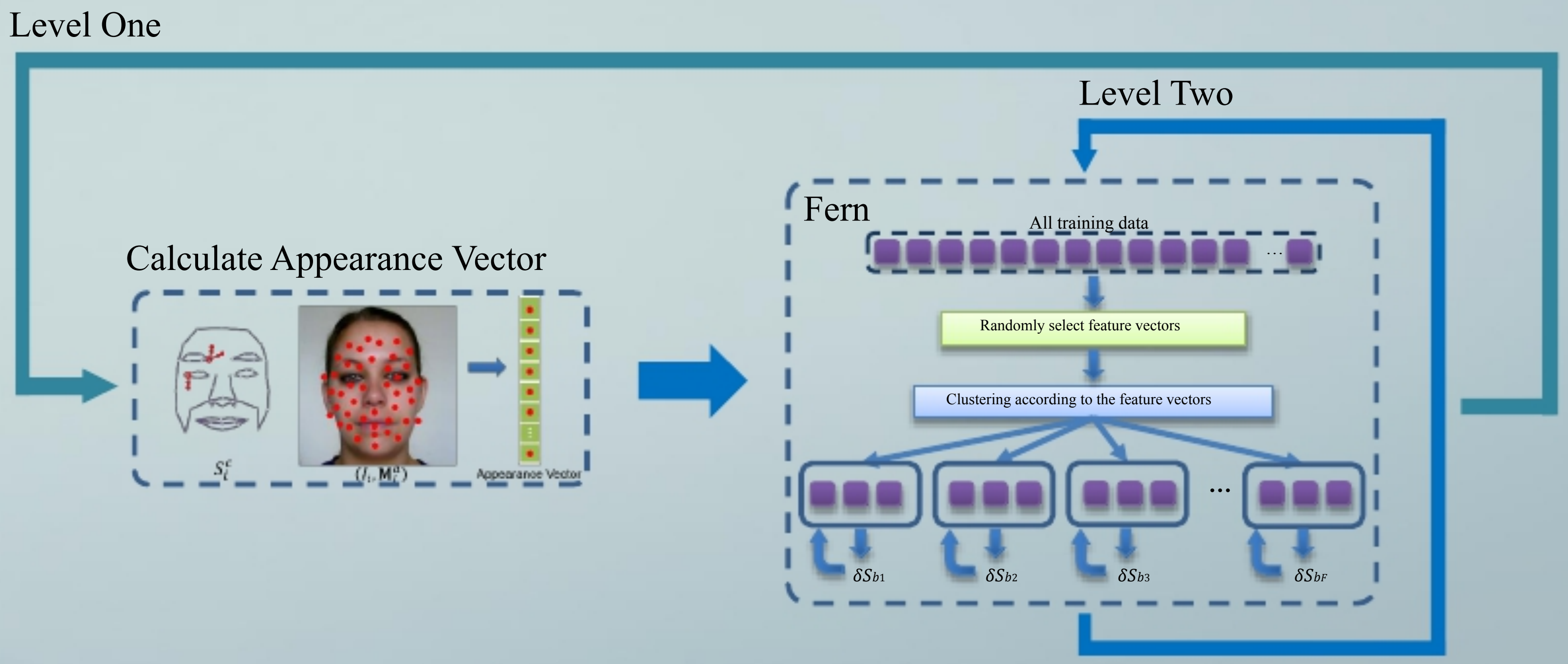
$$\delta S_b = \frac{1}{1+\beta/|\Omega_b|}\frac{{}_{i\in\Omega_b}(S_i - S_{ic})}{|\Omega_b|}$$

$$S_{ic} = S_{ic} + \delta S_b, i \in \Omega_b$$

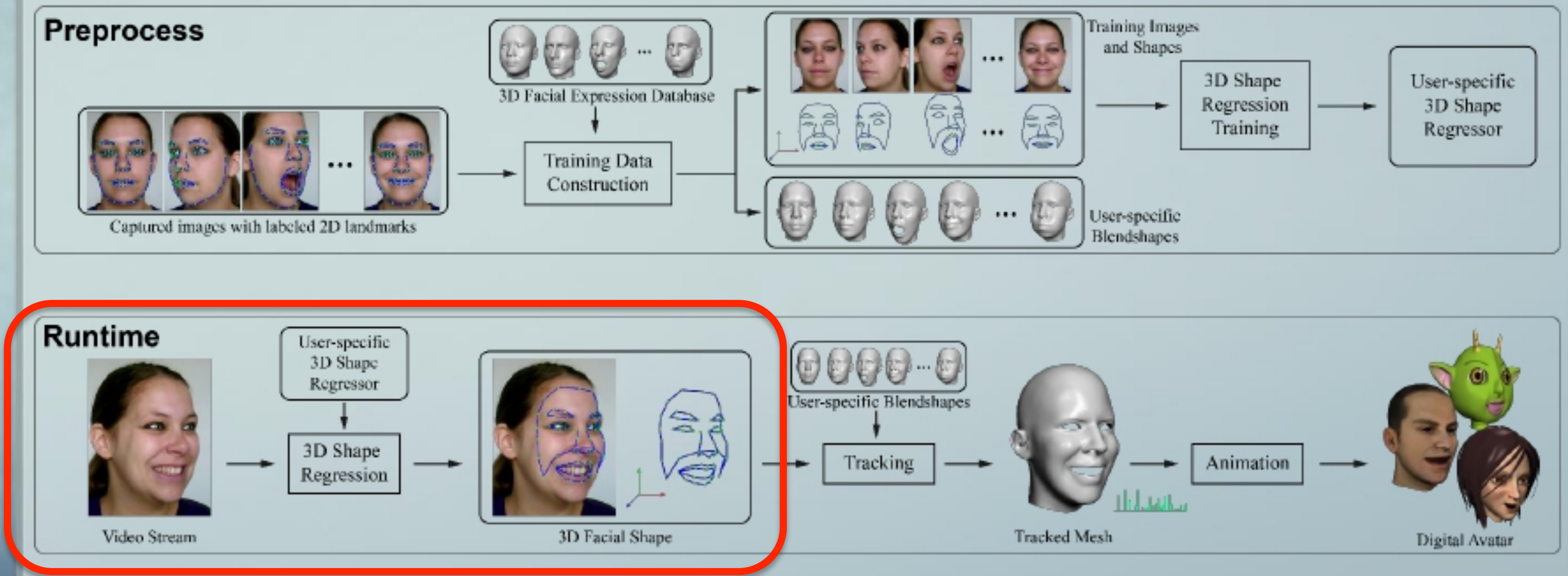


Preprocess: Training

- Summary: two-level boosted regression



Our Pipeline

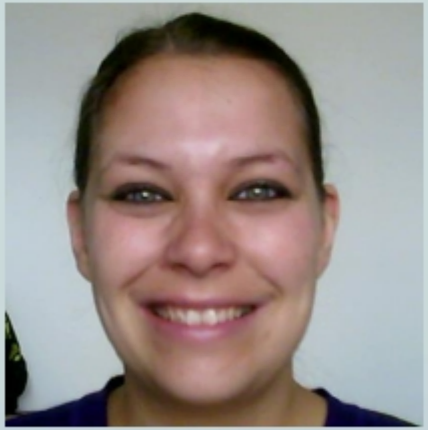


3D Face Shape Regression: Runtime

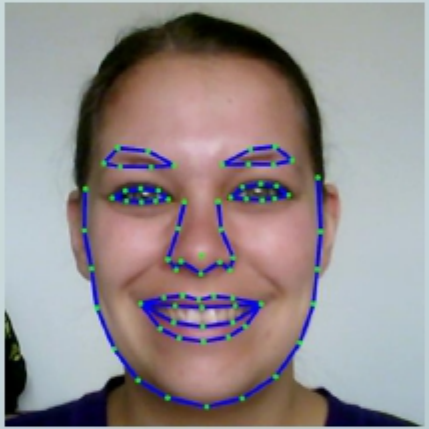
- Initialization: first frame
- Following frames
  - Find guessed shapes
  - Two-level boosted regression

Runtime: Initialization

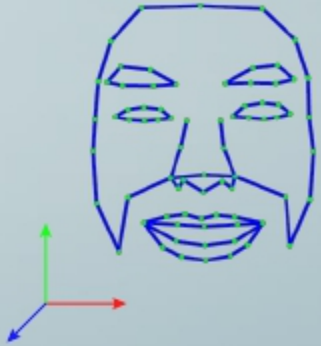
- First frame



Face Detection
[Viola and Jones 2001]



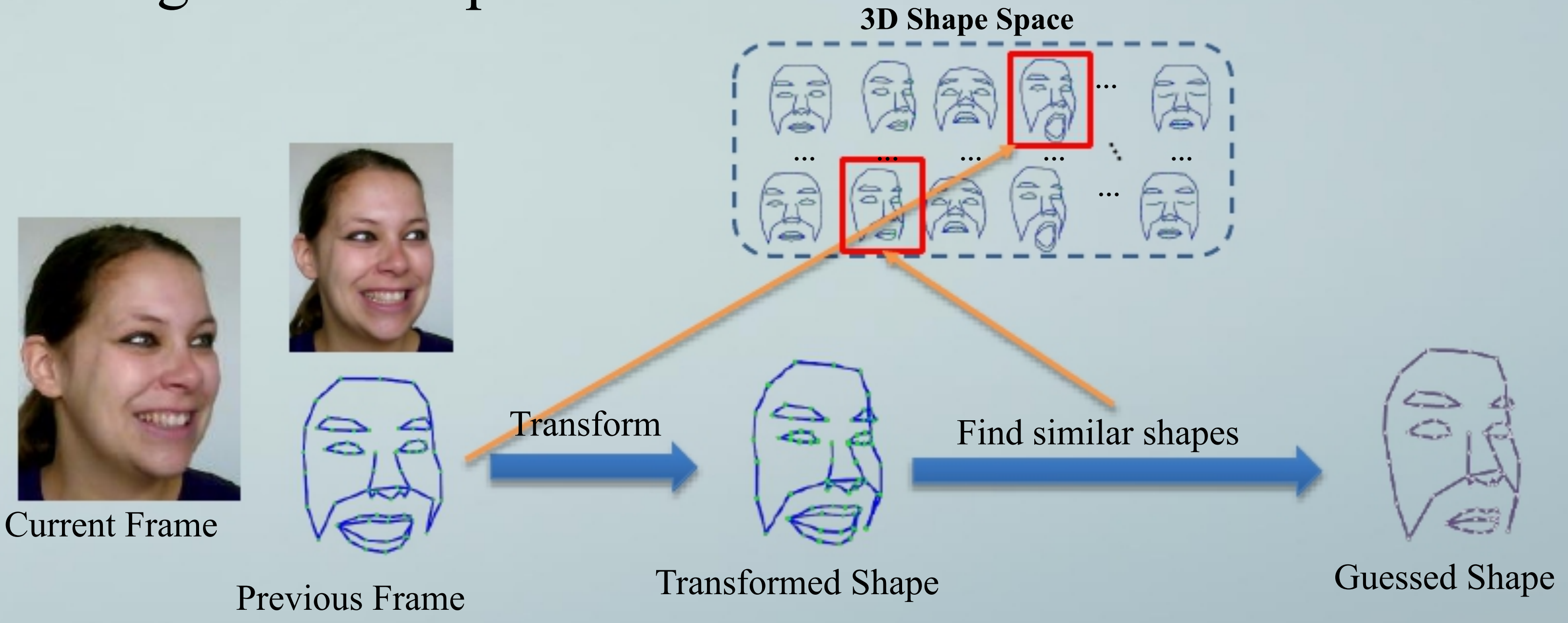
2D Feature Alignment
[Cao et al. 2012]



3D Shape Recovery

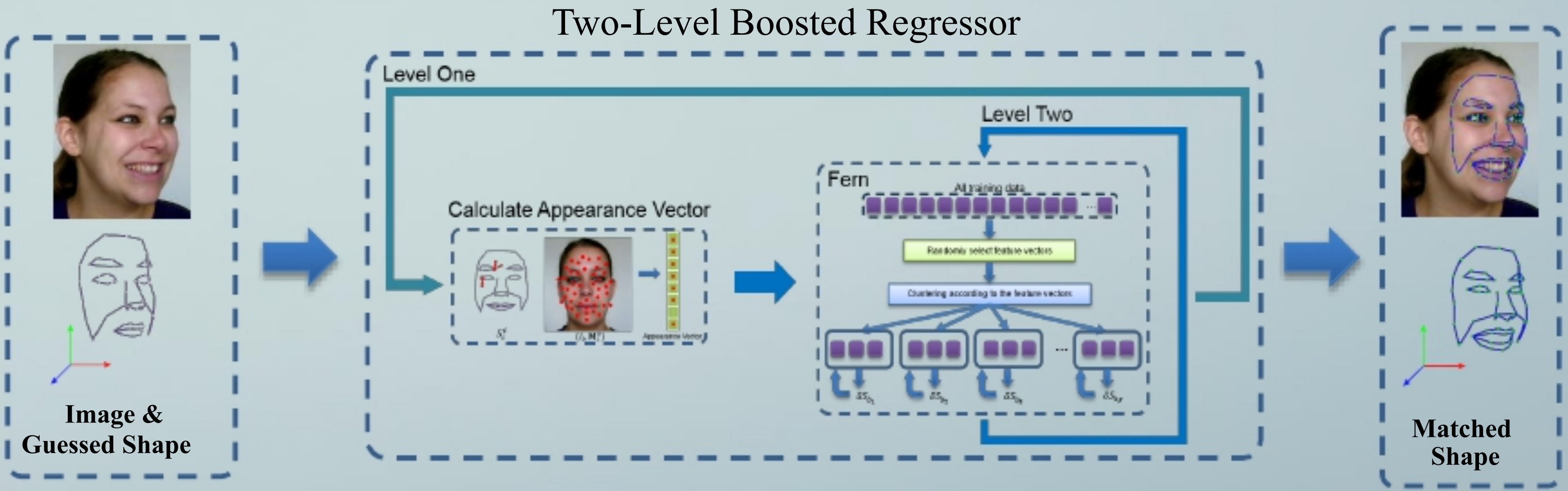
Runtime: Following Frames

- Find guessed shape

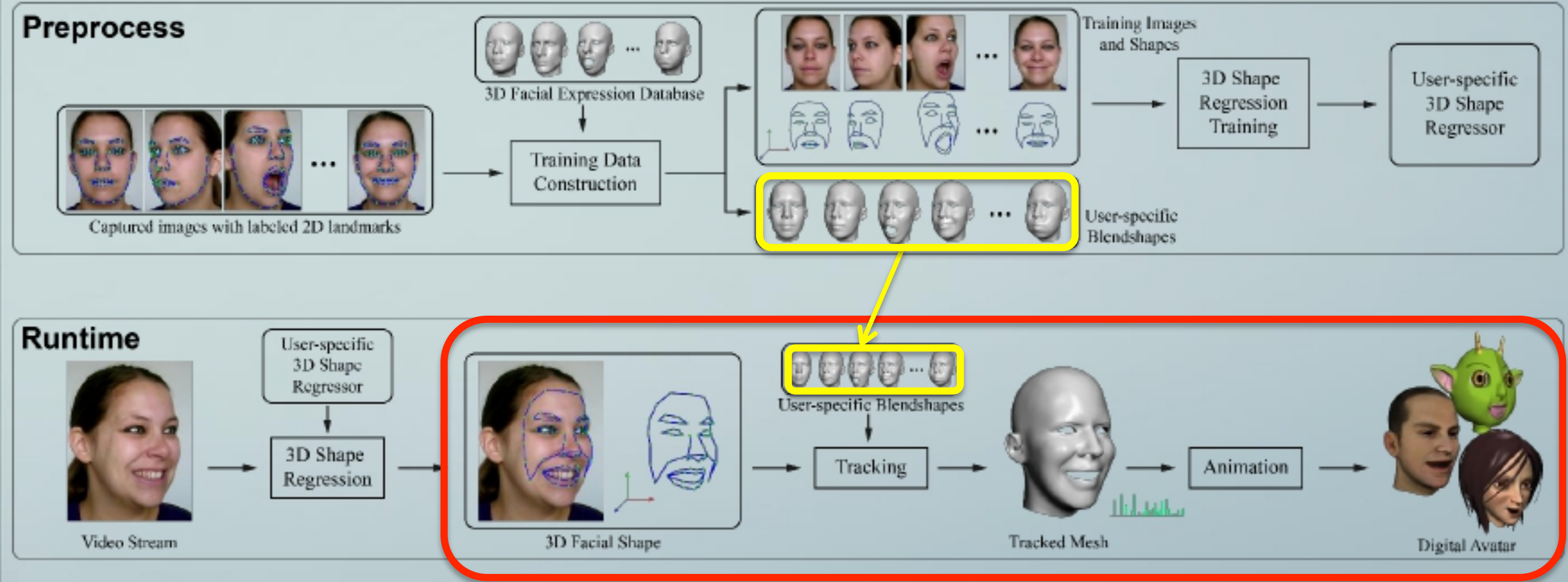


Runtime: Following Frames

- Two-level boosted regression

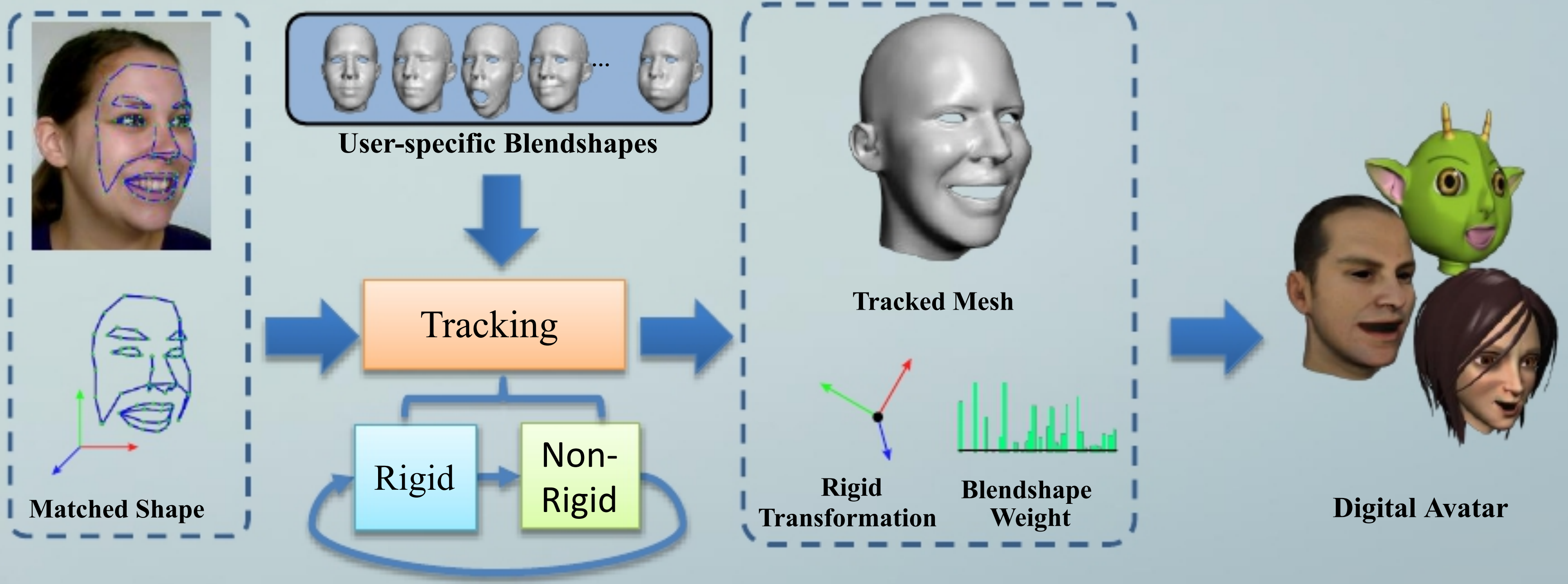


Our Pipeline



Tracking & Animation

Similar to [Weise et al. 2011]



Evaluation: Regressed shape vs. Kinect 3D vs. 2D vs. Optical Flow

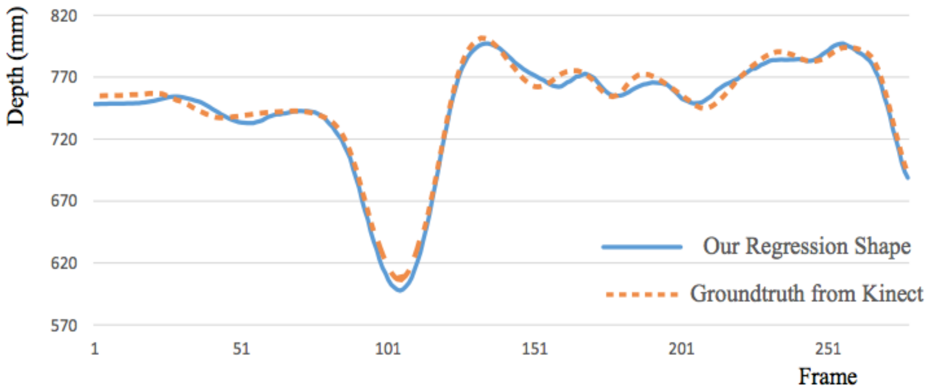


Figure 8: *Comparison of depth from 3D shape regression and ground truth from Kinect.*

RMSE	< 3 pixels	< 4.5 pixels	< 6 pixels
3D Regression	73.3%	80.8%	100%
2D Regression	50.8%	64.2%	72.5%
Optical Flow	20.8%	24.2%	41.7%

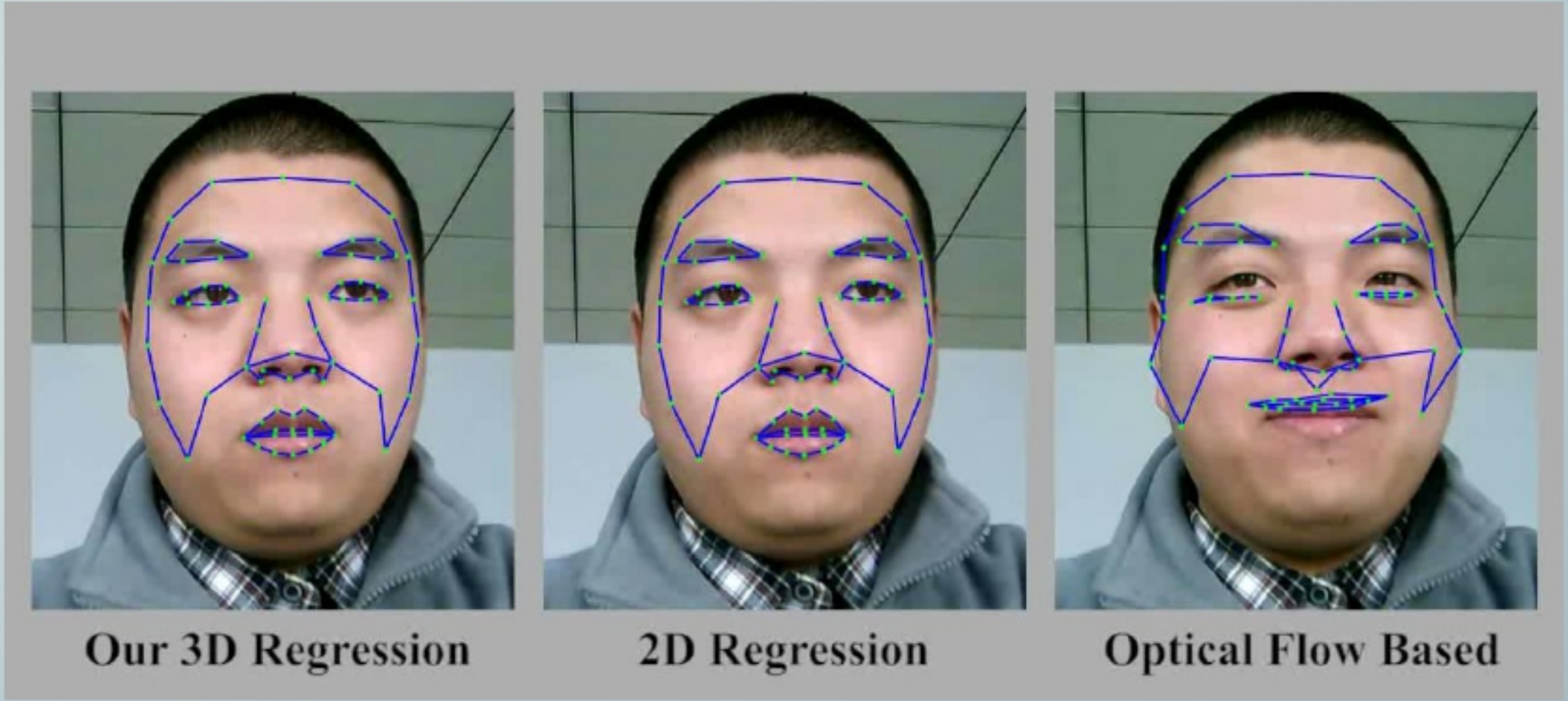
Table 1: *Percentages of frames with RMSE less than given thresholds for the tested video sequence.*

Live Demo

- Demo

Evaluation: Regressed shape vs. Kinect

Evaluation: 3D vs. 2D vs. Optical Flow



SIGGRAPH2013

More Results: Outdoor

SIGGRAPH2013

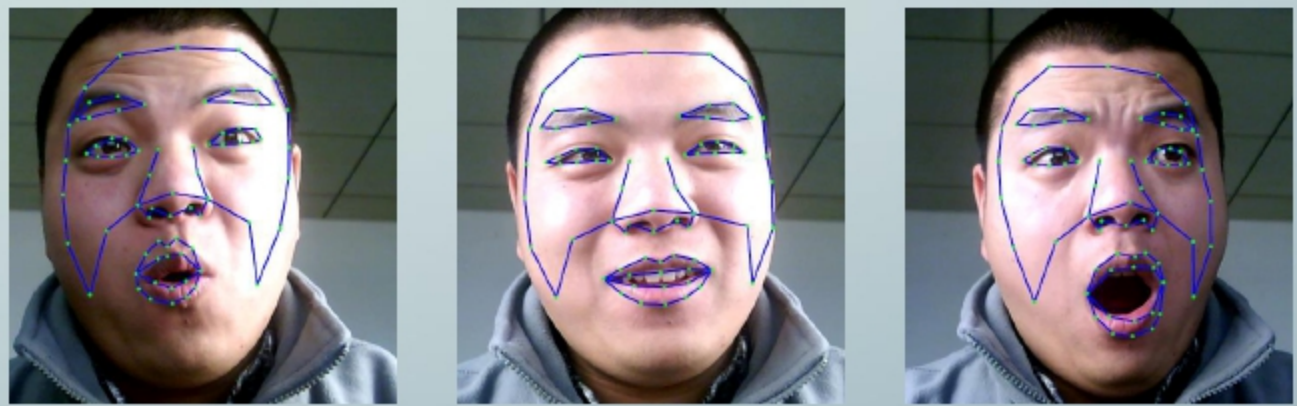
Our System on Mobile Device

Timings

- Preprocess: 45 mins
  - Capture: 10 mins
  - User interaction: 25 mins
  - Training: 10 mins
- Runtime: less than 15 ms
  - Regression: 5 ms
  - Tracking & Animation: 8 ms

Limitations

- Much training data
  - 60 head poses and facial expressions
- Dramatic lighting changes



Summary

- 3D facial performance capture from 2D video
  - Regression-based approach
  - Robust: fast motions, large rotations, exaggerated expressions
  - General environments: indoors and outdoors
  - High performance: real-time

- Future work
  - Handle lighting variations
  - Reduce training data

Acknowledgement

- Face capture: Marion Blatt, Steffen Toborg
- Anonymous reviewers
- Funding
  - NSFC (No.61003145 and No.61272305)
  - 973 program of China (No.2009CV320801)

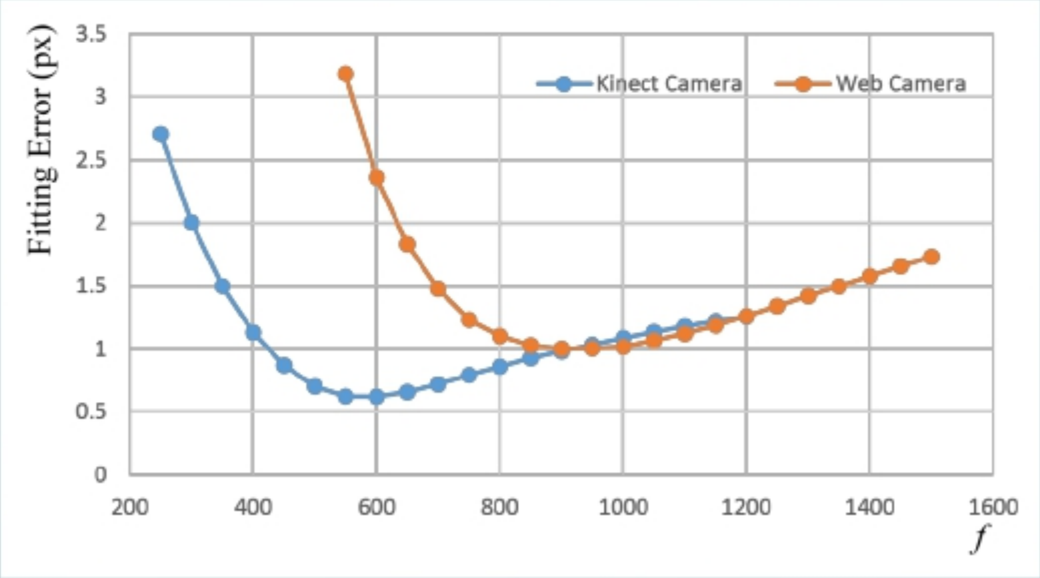
- FaceWarehouse Data: http://gaps-zju.org/facewarehouse/

Thank you!

Preprocess: Camera Calibration

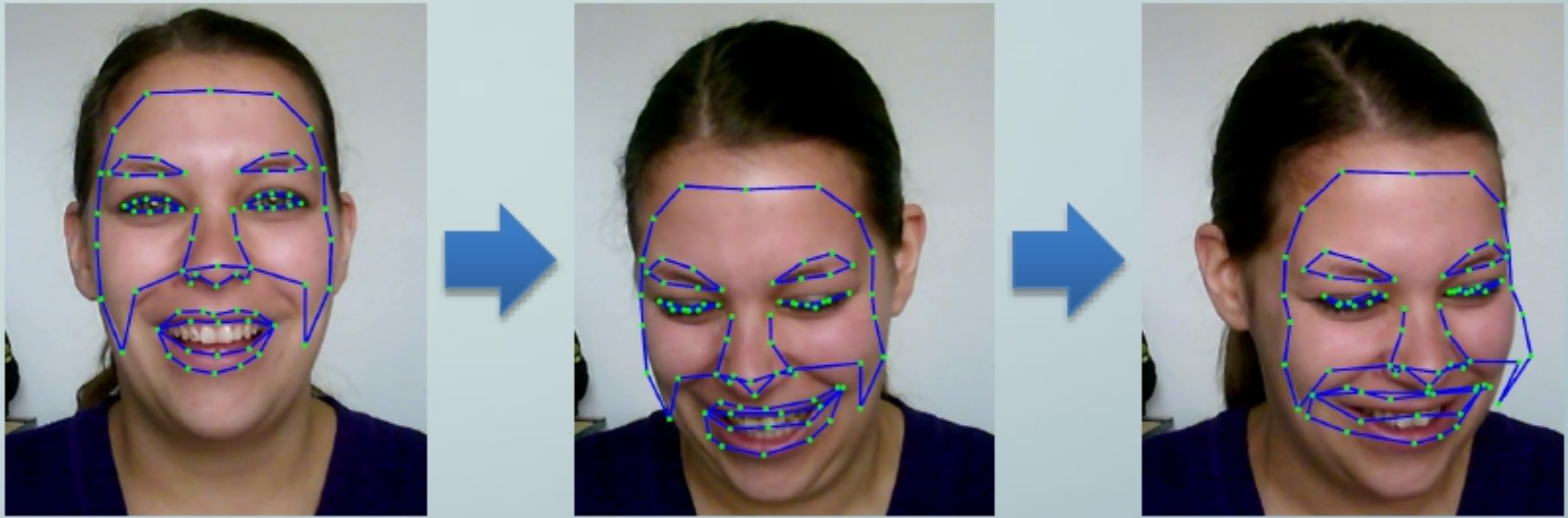
Blendshape Generations: $E = \sum_{i=1}^{n} \sum_{k=1}^{75} \left\| \Pi_{\mathbf{Q}} \left(\mathbf{M}_i \left(C_r \times_2 \mathbf{w}_{id}^T \times_3 \mathbf{w}_{exp,i}^T \right)^{(v_k)} - \mathbf{u}_i^{(k)} \right) \right\|^2$

$$\mathbf{Q} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$



Why not directly use previous shape?

- Error accumulation



Why not directly regress parameters?

- Expression coefficients in [0:1]
- Animation prior
  - Temporal coherence
- Rigid transformation & expression coefficients
  - Different spaces