Real-Time Face Pose Estimation from Single Range Images

Michael D. Breitenstein, Daniel Kuettel, Thibaut Weise, Luc Van Gool, Hanspeter Pfister

Computer Vision Laboratory, ETH Zurich
School of Engineering and Applied Sciences, Harvard

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Harvard School of Engineering and Applied Sciences

Motivation

Task: Locate Face and estimate Head Pose in Real-Time

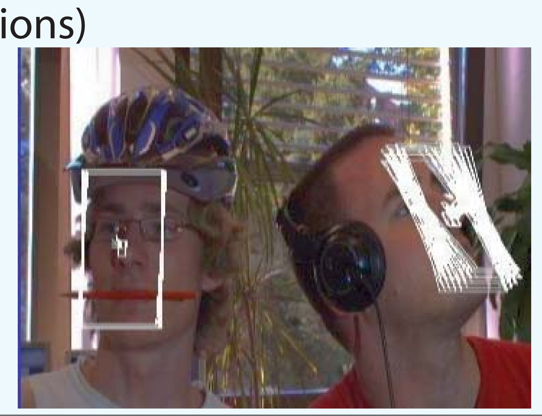
Often used during runtime or as a preprocessing step for:

face recognition
facial expression analysis
driver-attentiveness monitoring
human-computer interaction

Target: Robustness for real-world applications
⇒ Use range images to overcome limitations of 2D methods

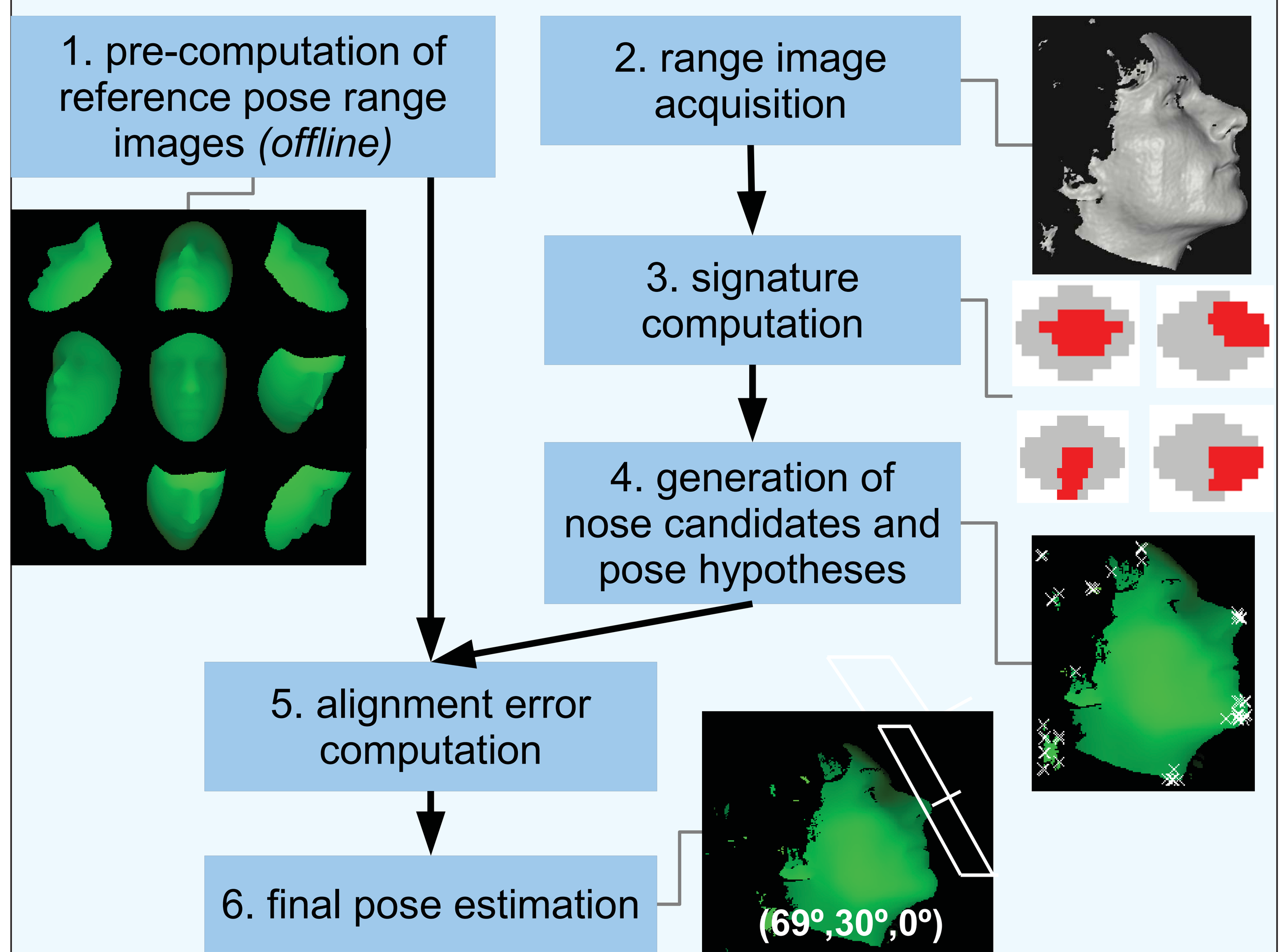
Requirements:

- Robustness to:
  - Large pose variations (±90°/ ±45°/ ±30°)
  - Facial variations (expressions, emotions)
  - Occlusions (glasses, hair, gestures)
  - Frame drop-outs (no tracking)
  - Multiple faces in the field of view



- No manual initialization or interaction
- For previously unseen persons
- Real-time

Method Overview



Contributions

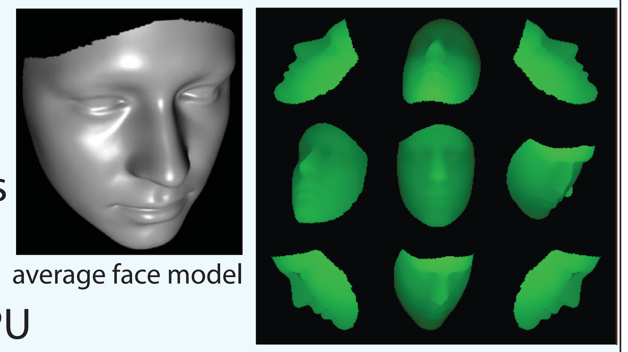
- 3D shape signature to find nose tip
- Error function to evaluate alignment of two range images
- Algorithms designed for highly parallel implementation on GPU

⇒ Parallel computations replace piecemeal analysis based on sophisticated feature extraction

Algorithm Details

1. Precomputation of Reference Pose Range Images

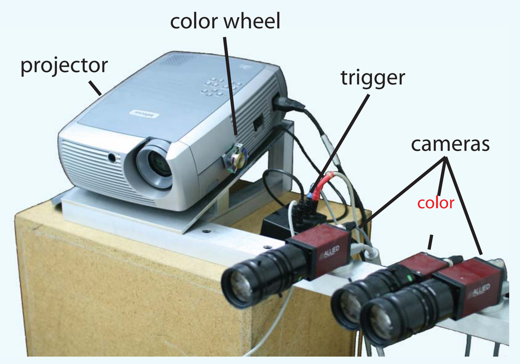
- Generate average 3D face model
  - Mean from 138 persons
- Render face model for many poses
  - With step sizes of 6°
- Store reference pose images to GPU



reference pose range images

2. Range Image Acquisition

- Stereo-enhanced structured-light scanner [Weise et al., CVPR'07]
  - Real-time (28 fps)



range scanner

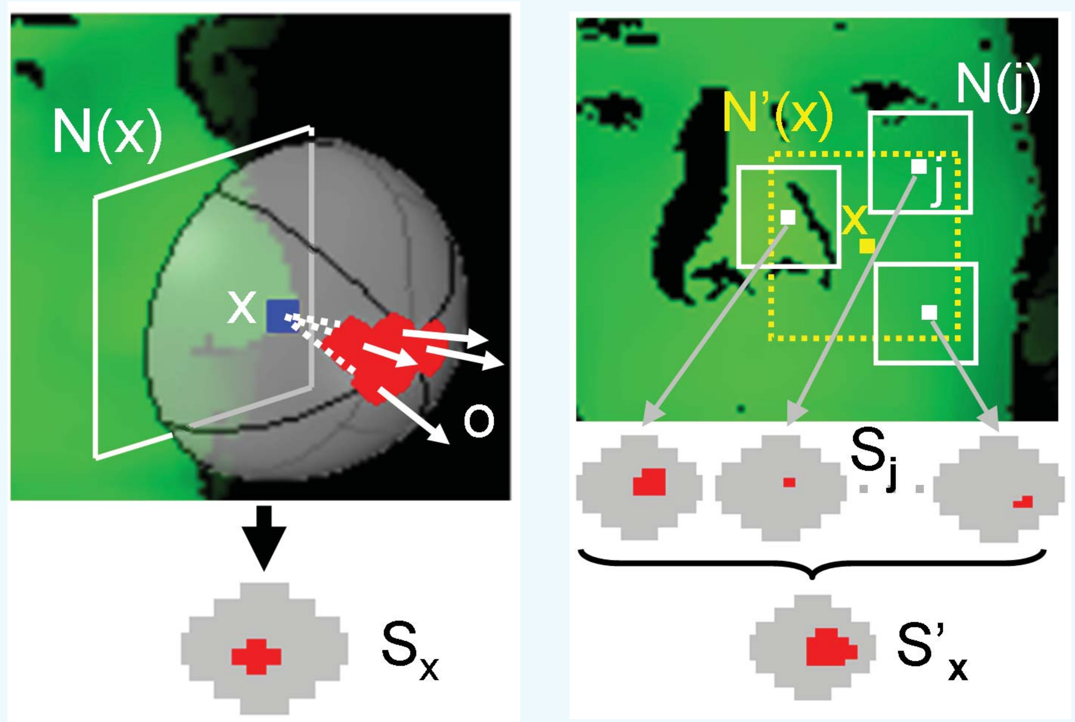

input range image

3. 3D Shape Signature Computation

Target: Find nose tip in range image for initial alignment of input and reference range images

⇒ Compute a signature that is:
- Characteristic for local shape (e.g. high curvature regions)
- Independent of head pose
- Able to distinguish different facial regions

- Single signature (matrix) for each pixel **x**:
  - Each cell corresponds to one orientation **o**
  - Cell marked iff **x** is a local directional maximum for **N** (= max. along **o** compared to pixels in neighbourhood **N**)
  - Computed for 56 orientations
- Signatures sparse ⇒ merge signatures in neighbourhood **N'**
  - Cell marked iff a pixel in **N'** is a local directional max. for **N**



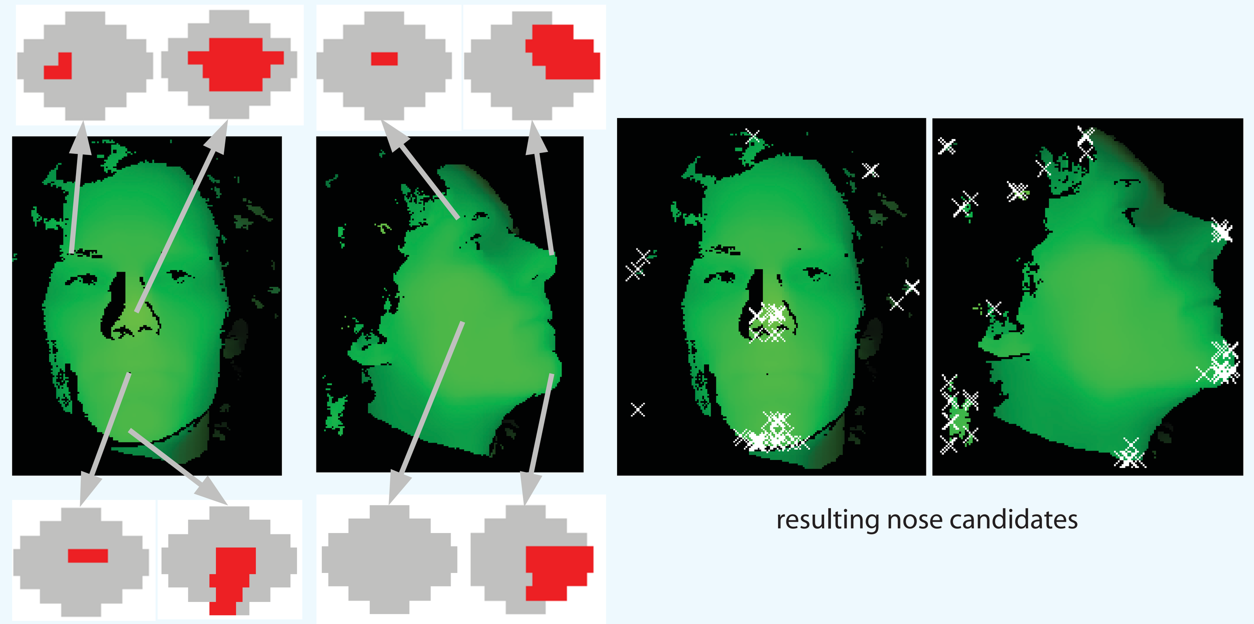
single signature   aggregated signature

Algorithm Details (cont.)

4. Generation of Nose Candidates and Pose Hypotheses

- ***Resulting signatures:***
  - ***Distinct for different facial regions***
  - ***Cover many adjacent cells for convex extremities (nose, chin)***
  - ***Look similar if head is rotated***
- Create nose candidates from pixels based on signatures:
  - **T** > 5 cells have to be marked
  - Pixel is representative for area (⇒ Single signature contains mean orientation of area)

⇒ Rough pose hypothesis = nose candidate + mean orientation


typical signatures for different facial regions   resulting nose candidates

5. Alignment Error Computation

- ***Target: Evaluate alignment of two range images M, I***
  - Nose and chin positions annotated in pose reference image **M**
  - Input image **I** translated to nose candidate position **x**
- Per-pixel error function:

$$e(M_{\boldsymbol{o}}, I_{\boldsymbol{x}}) = e_d(M_{\boldsymbol{o}}, I_{\boldsymbol{x}}) + \lambda \cdot e_c(M_{\boldsymbol{o}}, I_{\boldsymbol{x}}) + C$$

- Depth difference error term over foreground pixels of **I** and **M**
  - Does not penalize small overlaps between **M** and **I**

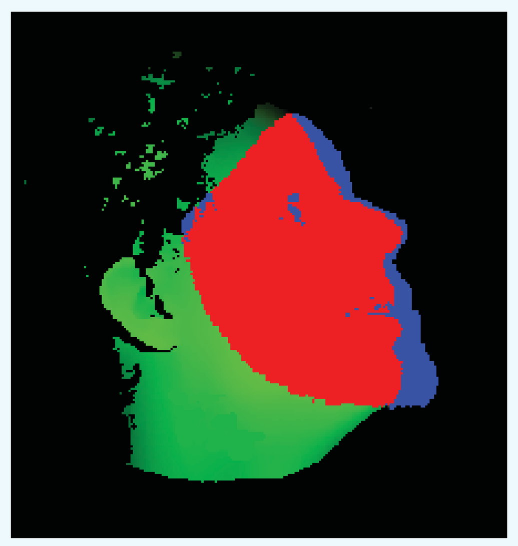
$$e_d(M_{\boldsymbol{o}}, I_{\boldsymbol{x}}) = \frac{\sum_{\boldsymbol{u} \in \mathcal{V}} (M_{\boldsymbol{o}}(\boldsymbol{u}) - I_{\boldsymbol{x}}(\boldsymbol{u}))^2}{|\mathcal{V}|}$$

- Coverage error term
  - Ratio of foreground pixels in **M** without correspondence in **I**

$$e_c(M_{\boldsymbol{o}}, I_{\boldsymbol{x}}) = \left(\frac{|\mathcal{V}^{-1}|}{|\mathcal{V}_{M_{\boldsymbol{o}}}|}\right)^2$$

- Constant **C** for additional robustness if no signature at chin

red = $\mathcal{V}$
blue = $\mathcal{V}^{-1}$
red + blue = $\mathcal{V}_{M_{\boldsymbol{o}}}$


one alignment example

Algorithm Details (cont.)

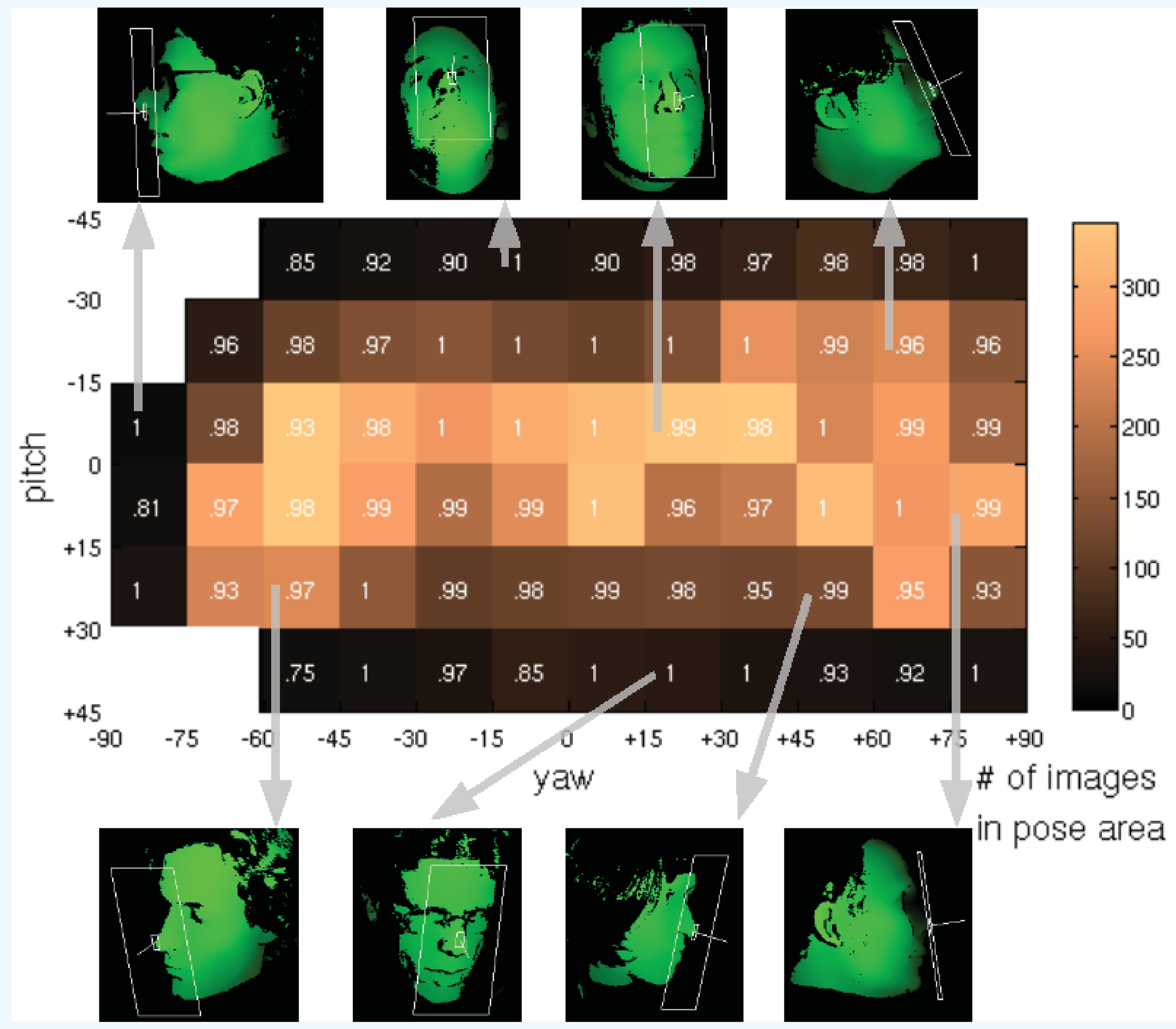
6. Final Pose Estimation

- ***Target: Parallel pose hypotheses evaluation:***
  - Select 5 rough pose hypotheses with smallest error
  - Augment with adjacent orientations from fine sampling (6° step)
  - Compute error of 125 pose hypotheses in parallel

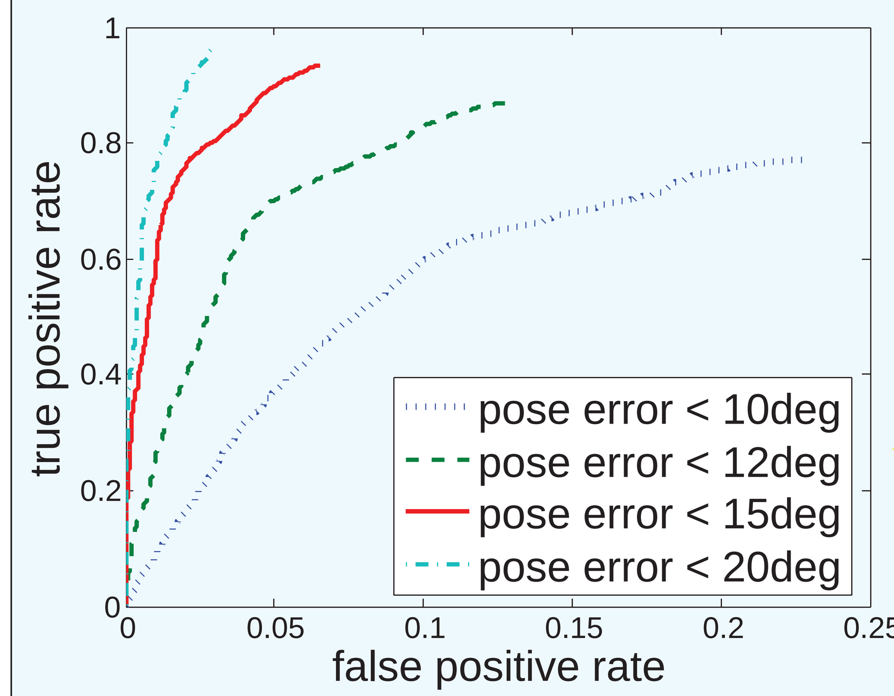
⇒ Pose hypothesis with smallest error = final pose estimation + confidence value

Results

- Robust:
  - Works for a very large pose range (±90°/ ±45°/ ±30°)
  - Robust to different variations (occlusions, facial expression)
  - 97.8% success rate for error < 15°
- Fast:
  - 55.8 fps (15 fps with range acquisition on same PC)
  - Necessary resolution only 32 x 32 pixels



pose estimation success rate dependent on pose area



Data set:
- 10'545 images,
- 26 people
- Poses (±90°/ ±45°)

Data available!

ROC curves of pose estimation performance for different error criteria