Face Pose Estimation from Range Images

Jeppe Jensen and Carsten Høilund

October 16th 2008

Outline

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

Outline

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

Definition of Range Images

- Range images are grayscale images of the original scene where the intensity of a pixel corresponds to the depth at that point.
- Range images are also called depth images or depth maps.

(a) Original scene.

(b) Depth image. [Ballinger, 2007]

Range Image Acquisition

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

Range Image Acquisition

- Methods to acquire depth include but are not limited to [Carsten Høilund and Jeppe Jensen and Simon J. K. Pedersen, 2008]:
  - Time-of-flight
  - Structured Light
  - Shading
  - Motion
  - Stereo Vision
  - 3D Modelling

Time-of-flight

- Time-of-flight measures the delay between emitting a pulse and receiving the reflection.

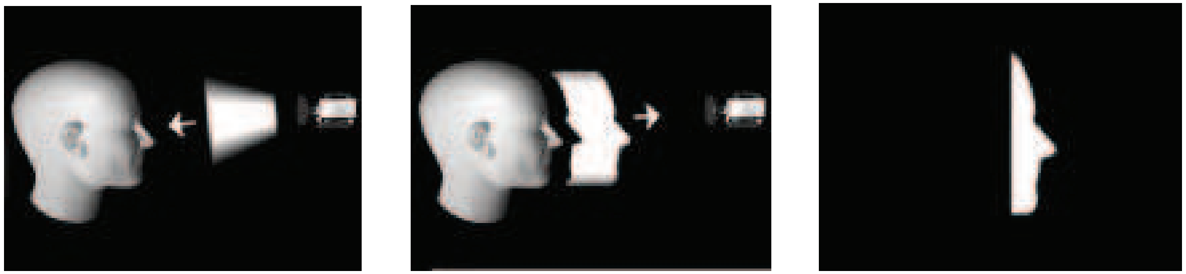


Figure 1: 3D camera using infrared light. [Gvili et al.,]

Structured Light

- Structured light works by projecting one or more patterns on an object and calculating the depth information from the distortion of the pattern on the objects surface.

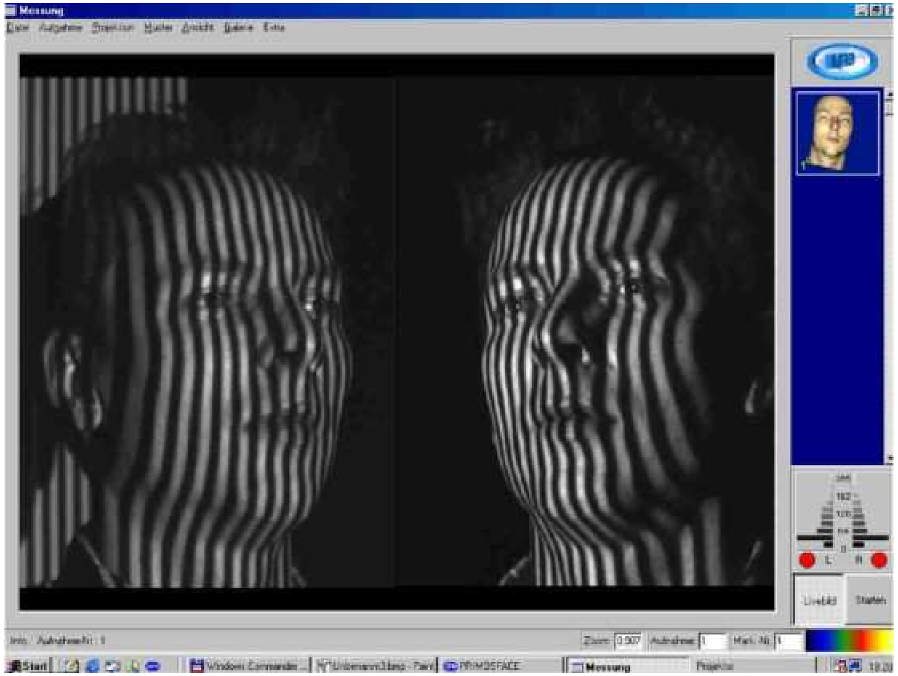


Figure 2: Multiview structured light. [Wikipedia, 2008b]

Shading

- Shading is related to structured light but is a passive technique since it utilizes existing light.

Figure 3: Notice the shading on the ball. [About.com, 2008]

Motion

- The motion parallax between stationary objects when moving the camera hints at their difference in depth.

Figure 4: See original image for 3D effect. [Heresy, 2007]

- Stereo vision is similar to the vision of humans (and other creatures) and uses the disparity between objects and their features when viewed from different angles to infer depth information of a scene.

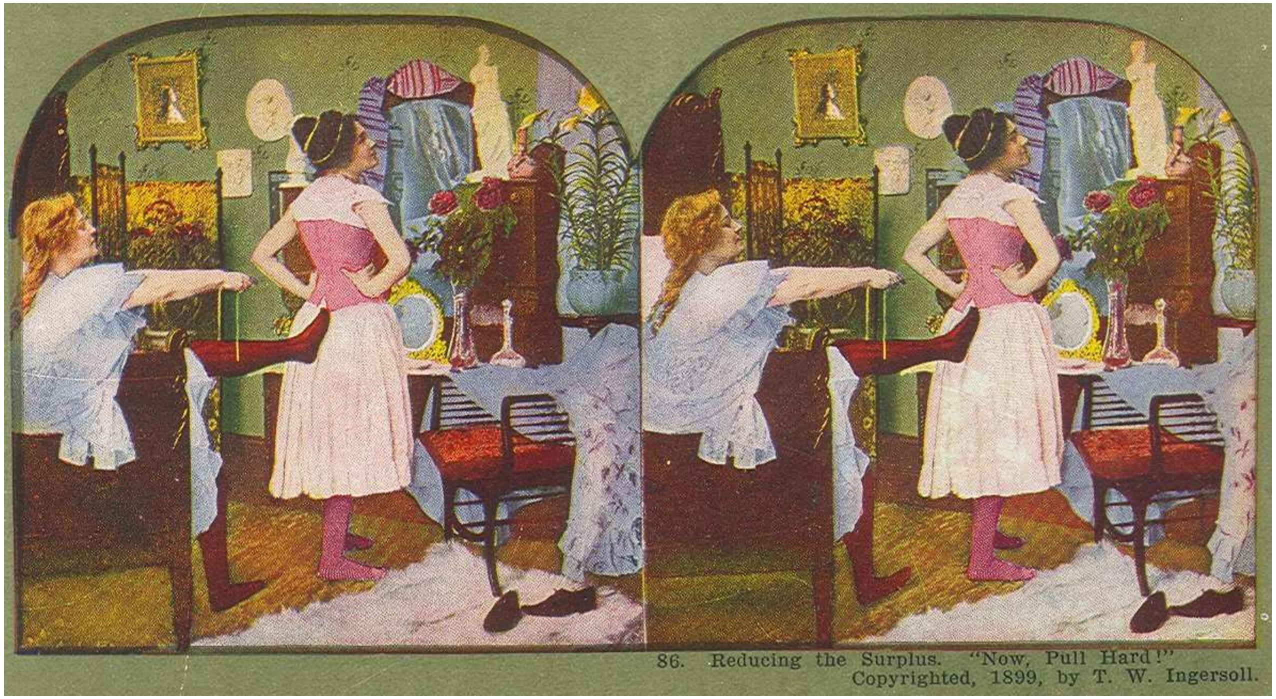


Figure 5: Two images of the same everyday scene seen from slightly different angles. [Wikipedia, 2008a]

3D Modelling

- In 3D modelling the scene is manually created and is thus different than the other approaches. It is quite common – especially in augmented reality with simple scenes.

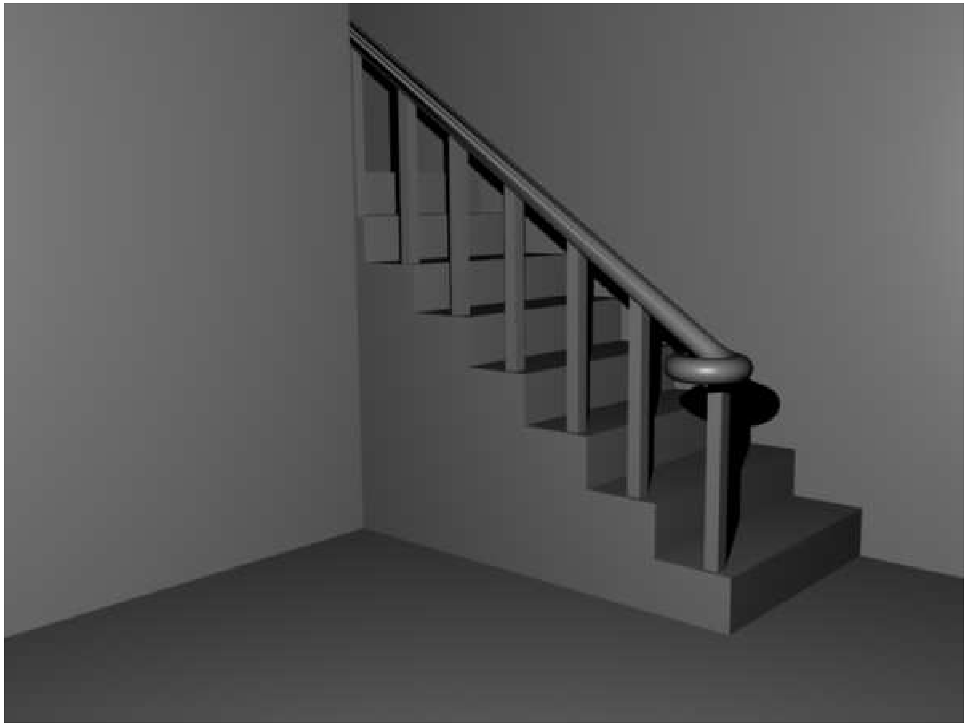


Figure 6: Hand modelled simple scene. [Xedium3D.com, 2008]

Applications

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

Applications

- Applications of depth information are plentiful and include:
  - 3D Scanning of Objects
  - Augmented Reality
  - Navigation
  - Quality Control
  - Face Pose Estimation

3D Scanning of Objects

- Create a digital model of a real life object

Figure 7: Scanned version of a real life figurine. [Weise et al., 2007]

Augmented Reality

- Enhancing reality with additional, visual information.

Figure 8: What the bright future might bring. [Boucher, 2005]

Navigation

- 3D information of the path ahead allows robots to navigate

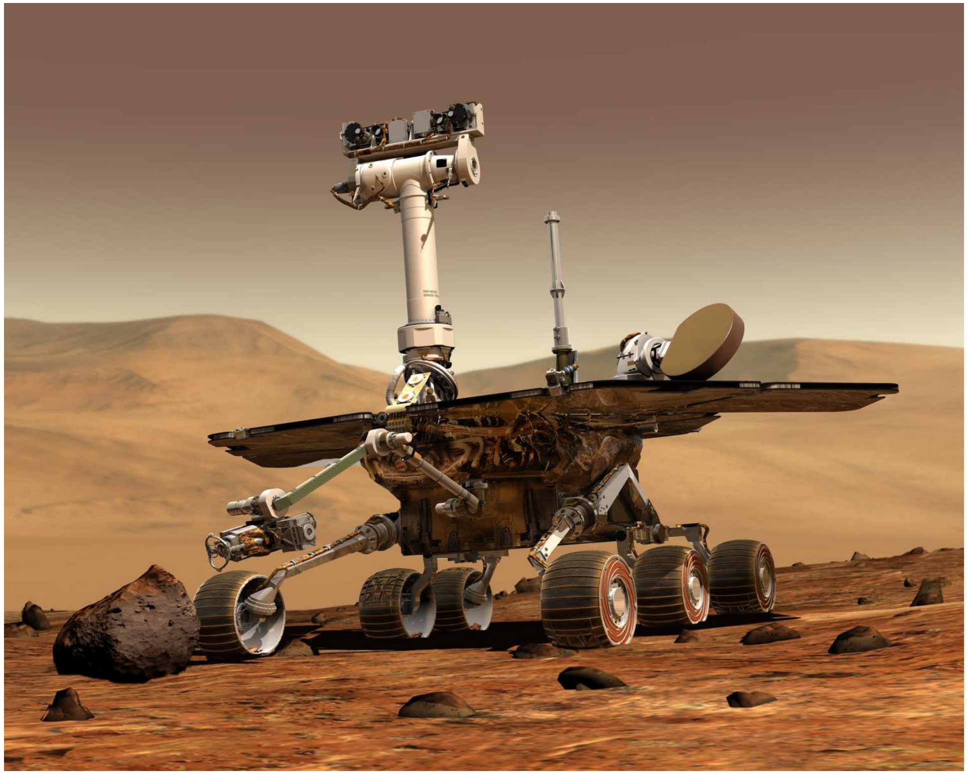


Figure 9: NASA's Mars rovers use stereo vision. [Society, 2008]

Quality Control

- Ensure parts have the right dimensions

Figure 10: Assessing the quality of tires. [Weise et al., 2007]

Face Pose Estimation

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

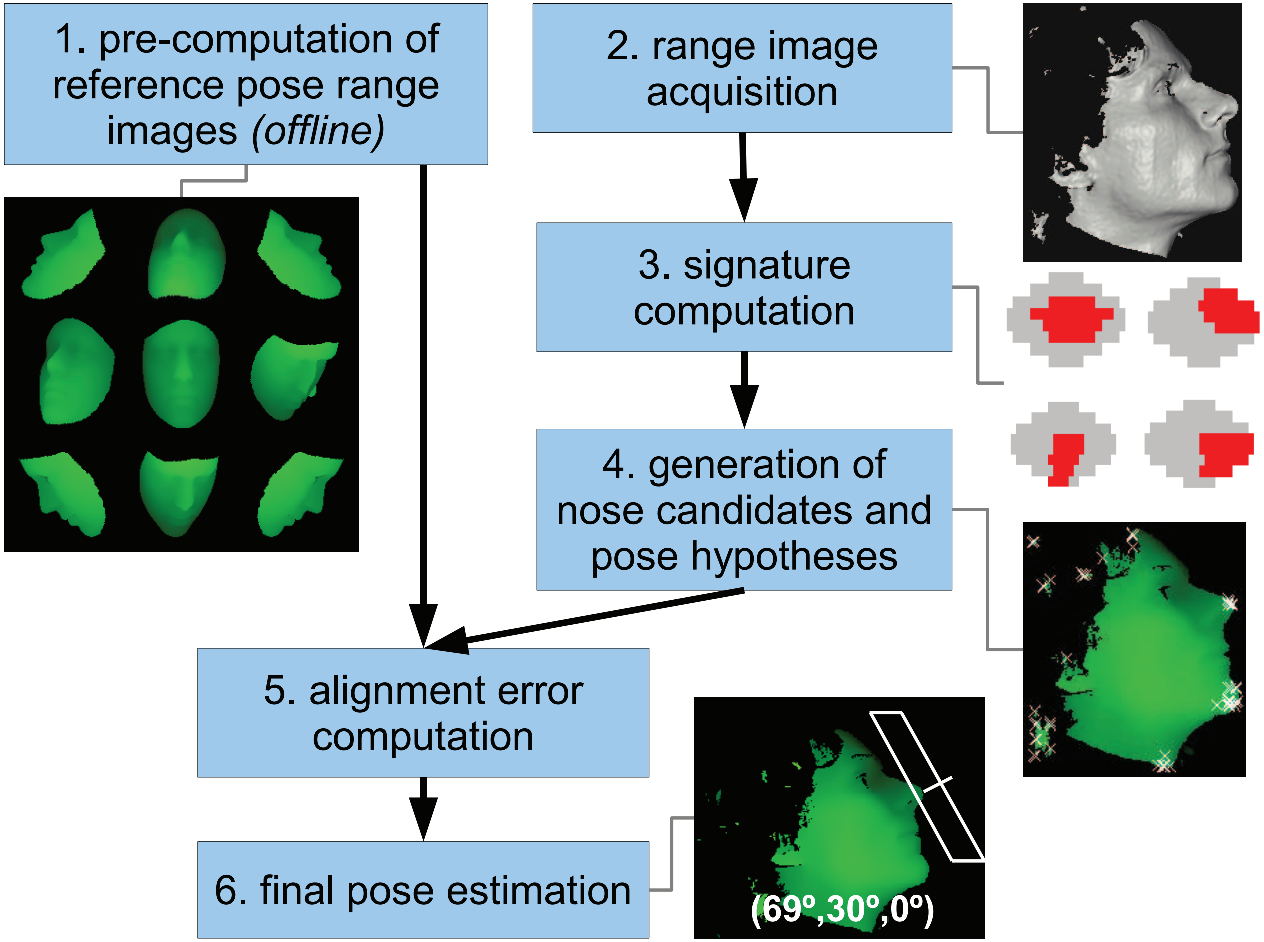
Related Work

- Pose Estimation from 2D Images
  - Analyze the entire image
  - Require exact localization of faces, or require pose-dependent features
  - Sensitive to illumination, shadows, lack of features, and occlusions
  - Some systems use stereo or multi-view images
- Pose Estimation from Range Images
  - Require range images
  - Previous work had different limitations, e.g. non real-time, require initialization or only works on single faces.

Main Article

- "Real-Time Face Pose Estimation from Single Range Images" By Michael D. Breitenstein, Daniel Kuettel, Thibaut Weise, Luc van Gool, Hanspeter Pfister [Breitenstein et al., 2008]
  - Robust to:
    - Large pose variation
    - Facial variations (expressions, emotions)
    - Occlusion (glasses, hair, gestures)
    - Multiple faces in the field of view
  - No manual initialization or interaction
  - Works independently on each frame
  - Real-time
  - Parallel computations on Graphics Processing Unit

Overview of the Algorithm

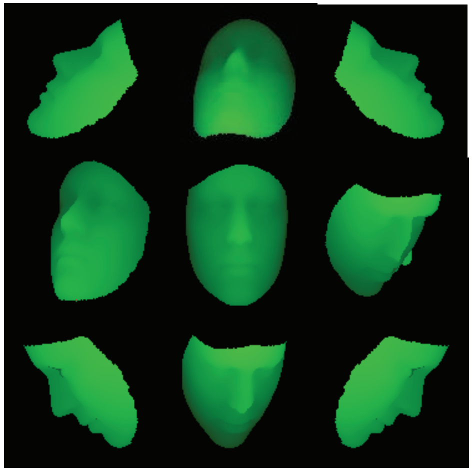


[Breitenstein et al., 2008]

1. Precomputation of Reference Pose Range Images

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

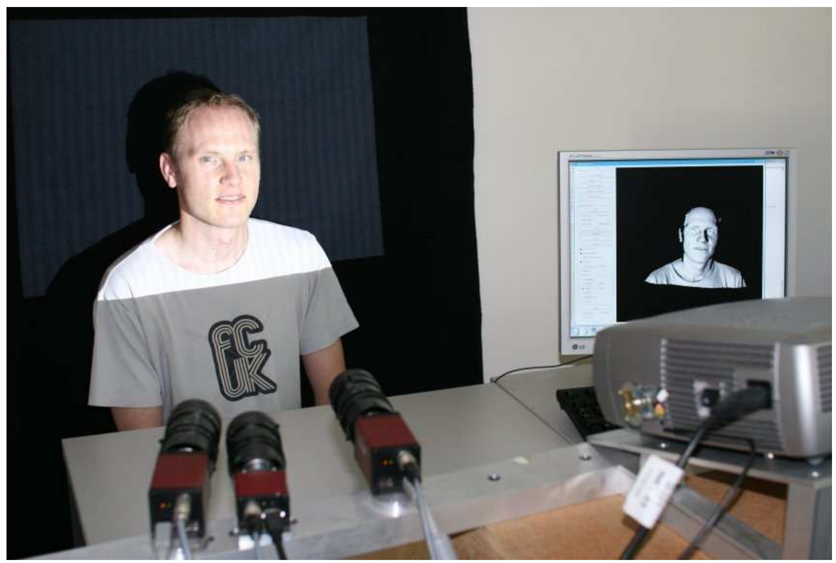
(a) Average face model



(b) Reference pose range images

2. Range Image Acquisition

- The range images are acquired from a structured light scanner [Weise et al., 2007]
- Calculate the depth from the deformation of the light patterns

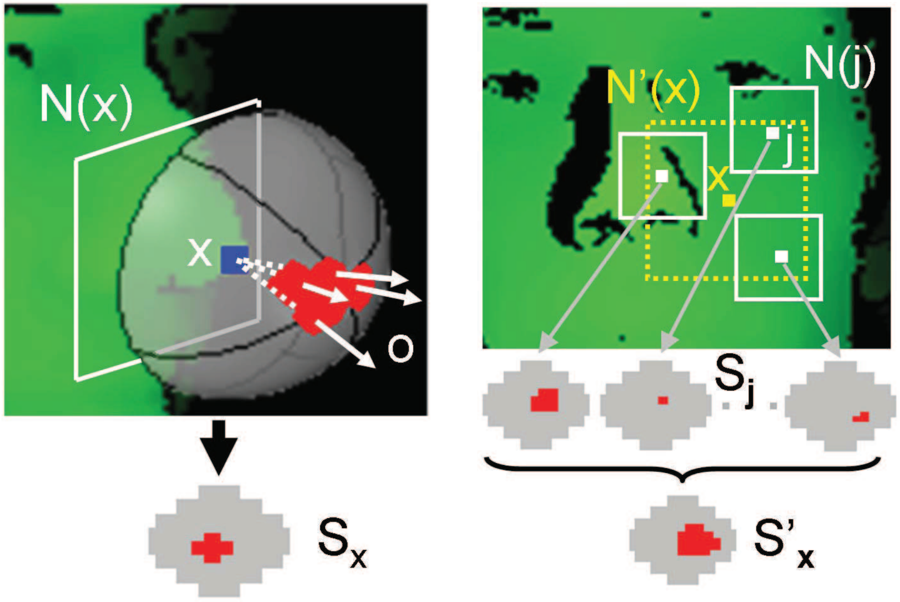


(c) The range scanner using structured light

(d) Input range image

3. 3D Shape Signature Computation (1)

- 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

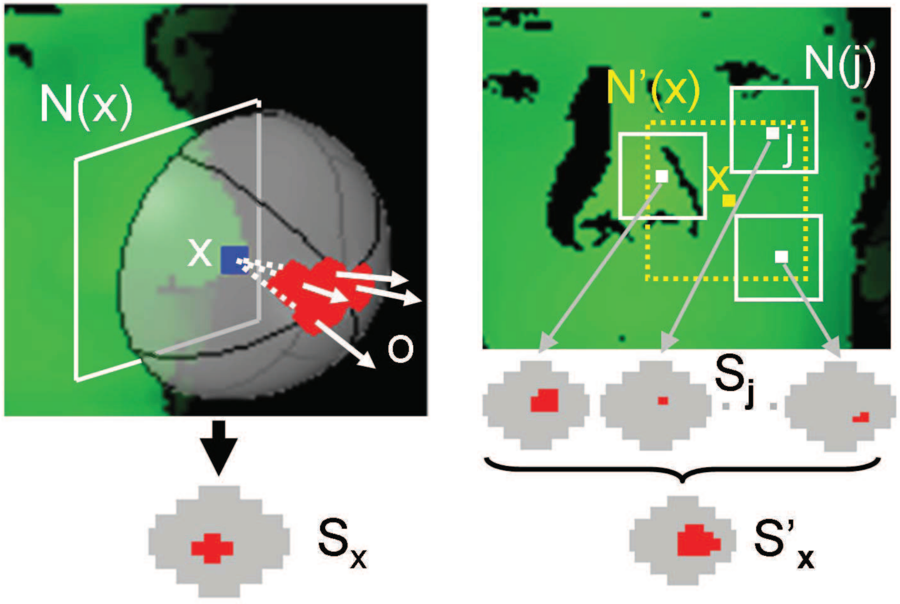


single signature

aggregated signature

3. 3D Shape Signature Computation (2)

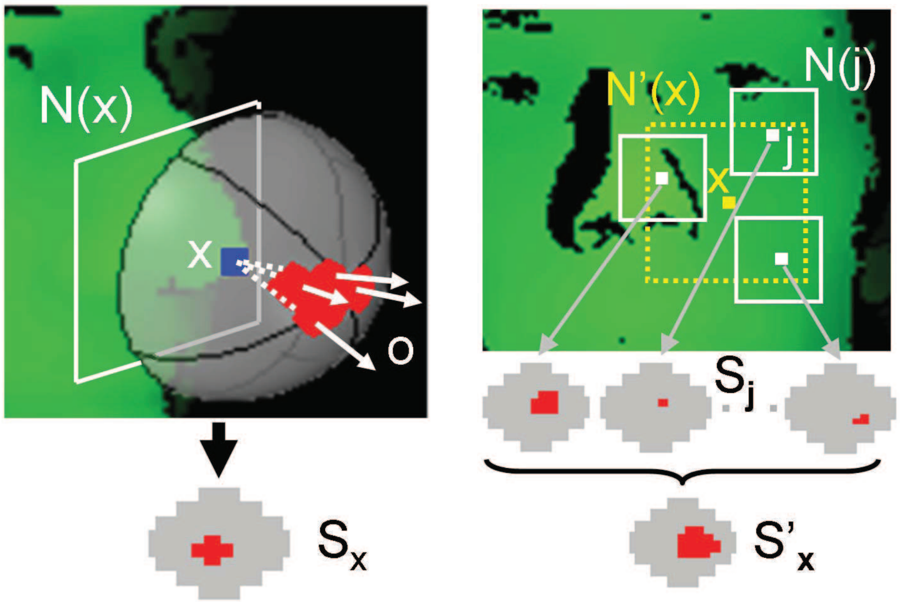
- Find nose tip in range image for initial alignment of input and reference range images
- Single signature (matrix) for each pixel x:
  - Each cell corresponds to one orientation o
  - Cell marked if x is a local directional maximum for N ($=$ max. along o compared to pixels in neighborhood N (32×32))
  - Computed for 56 orientations



single signature   aggregated signature

3. 3D Shape Signature Computation (3)

- Find nose tip in range image for initial alignment of input and reference range images
- Because signatures are sparse, merge signatures in neighborhood N'
  - Cell marked if a pixel in N' (32×32) is a local directional max. for N

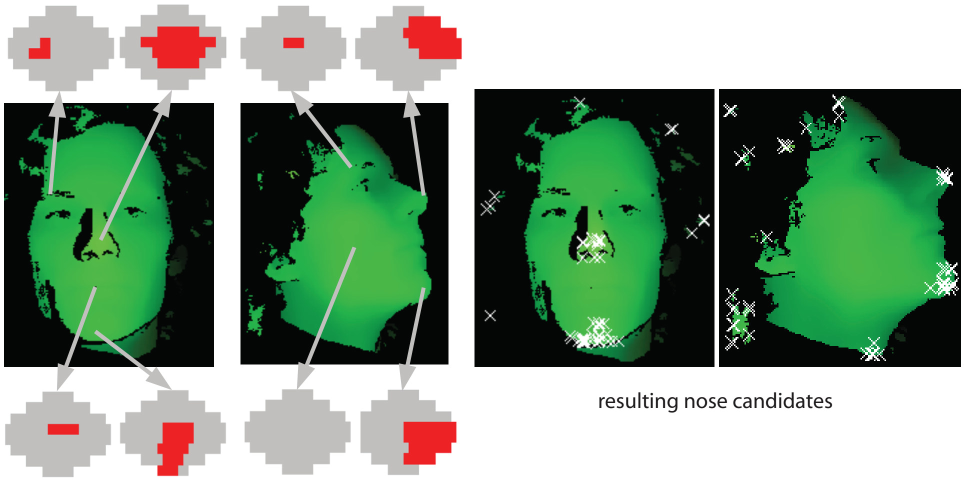


single signature

aggregated signature

4. Generation of Nose Candidates

- 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:
  - More than 5 cells have to be marked
  - Pixel is representative for area (Single signature contains mean orientation of area)
- 3D positions and mean orientations of nose candidate pixels form a set of head pose hypotheses



resulting nose candidates

typical signatures for different facial regions

5. Alignment Error Computation

- Evaluate alignment of two range images $M_{\mathbf{o}}$, $I_{\mathbf{x}}$
  - Nose and chin positions annotated in pose reference image $M_{\mathbf{o}}$
  - Input image $I_{\mathbf{x}}$ translated to nose candidate position x

Total error $=$ depth difference error $+$ coverage error $+$ constant

$$e(M_{\mathbf{o}}, I_{\mathbf{x}}) = e_d(M_{\mathbf{o}}, I_{\mathbf{x}}) + \lambda \cdot e_c(M_{\mathbf{o}}, I_{\mathbf{x}}) + C \quad (1)$$

- Term e_d computes differences between $M_{\mathbf{o}}$ and $I_{\mathbf{x}}$ for all foreground pixels (i.e., pixels where a depth was captured).
- Term e_d is the ratio of foreground pixels in $M_{\mathbf{o}}$ without correspondence in $I_{\mathbf{x}}$.
- Constant is added for additional robustness.

5. Alignment Error Computation

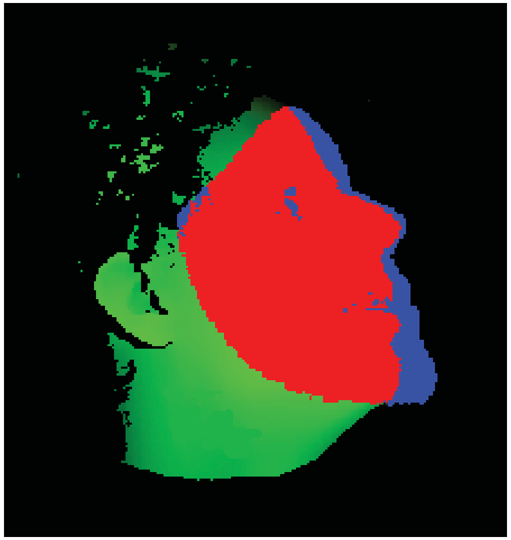


Figure 11: An alignment of one reference pose range image and the input image.

- Red area is the intersection of all foreground pixels between $M_{\mathbf{o}}$ and $I_{\mathbf{x}}$
- Blue area is all foreground pixels of the reference image that have no correspondence in the input image

6. Final Pose Estimation

- Roughly the five best post estimations are found, then fine sampled around the five estimates (125 pose hypothesis)
- The pose with the smallest error is the final pose estimation
- Results:
  - 97.8 % succes rate for error less than 15 degrees
  - 55.8 fps for actual algorithm, and 15 fps including image acquisition
  - Necessary resolution for reference models is only $32\times$ 32 pixels
  - Robust to variations (occlusion, facial expressions)
  - Works for a very large pose range

Video

Our Contribution

- Definition of Range Images
- Range Image Acquisition
- Applications of Range Images
- Face Pose Estimation using Structured Light
- Our Contribution: Face Pose Estimation using Stereo Vision

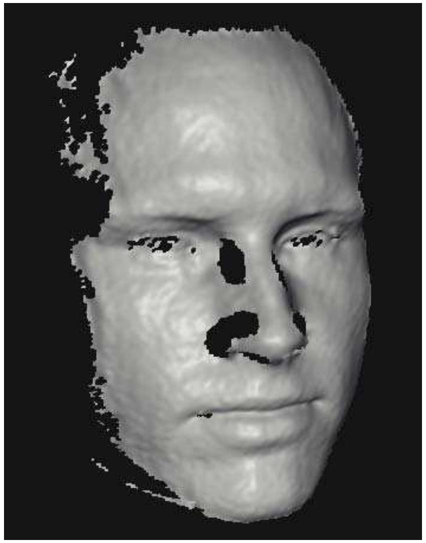
Stereo Camera

- Acquire range images using stereo camera
- No light patterns in the face
- Poor quality compared to structured light
- Our task is to get as good result as possible using the input images from stereo camera

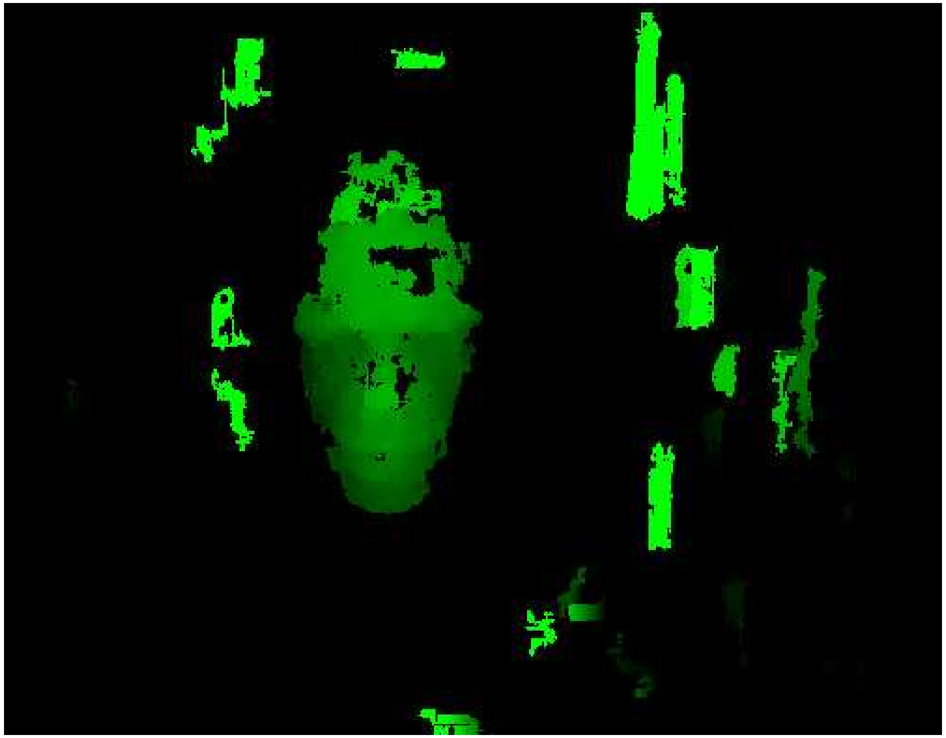
Figure 12: The BumbleBee XB3 stereo camera.

Input Data

Comparison of the input data when using structured and stereo camera.



(a) Range image from structured light



(b) Range image from stereo camera

Methods to Improve Input Data

- Different methods applied to get better data
  - Interpolation (fill holes in the face)
  - Skin segmentation (remove noise, and non-face pixels)
  - Filters (Gaussian, median)
  - Use face detection to locate region of interest (ROI) (improve frame rate, and remove surrounding noise)

Video

Post Processing

- Different methods applied for post processing to get better pose estimate
  - Face detection (nose always in the center of the face detection area)
  - Tracking of facial features (KLT)
  - Difference between center of mass of the face and the center of ROI
  - Temporal estimation (pose change from the previous pose)

Screenshot

Questions?

Questions?

References

[About.com, 2008] About.com (2008). Exercises on the ball. [Online; accessed 15-October-2008].

[Ballinger, 2007] Ballinger, B. (2007). a christmas bow. [Online; accessed 15-October-2008].

[Boucher, 2005] Boucher, T. (2005). How augmented reality will *really* work. [Online; accessed 15-October-2008].

[Breitenstein et al., 2008] Breitenstein, M. D., Kuettel, D., Weise, T., Gool, L. V., and Pfister, H. (2008). Real-time face pose estimation from single range images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

[Carsten Høilund and Jeppe Jensen and Simon J. K. Pedersen, 2008] Carsten Høilund and Jeppe Jensen and Simon J. K. Pedersen (2008). Occlusion Handling in an Augmented Reality System using Stereo Vision and Background Subtraction.

[Gvili et al.,] Gvili, R., Kaplan, A., Ofek, E., and Yahav, G. Depth keying. Technical report, 3DV Systems Ltd.

[Heresy, 2007] Heresy, T. B. (2007). 2d becomes 3d. [Online; accessed 15-October-2008].

[Society, 2008] Society, T. P. (2008). Planetary news: Mars (2008). [Online; accessed 15-October-2008].

[Weise et al., 2007] Weise, T., Leibe, B., and Gool, L. V. (2007). Fast 3d scanning with automatic motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*.

[Wikipedia, 2008a] Wikipedia (2008a). Stereoscopy — wikipedia, the free encyclopedia. [Online; accessed 15-October-2008].