Optimisation and Comparison Framework for Monocular Camera-Based Face Tracking

Stefan Marks
Department of Computer Science
The University of Auckland
New Zealand
Email: smar189@aucklanduni.ac.nz

John Windsor
Department of Surgery
The University of Auckland
New Zealand
Email: j.windsor@auckland.ac.nz

Burkhard Wünsche
Department of Computer Science
The University of Auckland
New Zealand
Email: burkhard@cs.auckland.ac.nz

***Abstract*—Tracking the position and orientation of the human face with respect to a camera has valuable applications in human computer interaction (HCI). Examples are navigating through a virtual environment, controlling objects using head gestures, and enabling avatars in a virtual environment to reflect the user's behaviour.**

Tracking performance can be heavily influenced by environmental parameters. Developers and users of face tracking plug-ins without computer vision experience need guidelines how to optimise face tracking performance in real world set-ups and they need measures how environmental parameters influence the results.

In this paper we develop a qualitative framework for determining ideal working conditions of face tracking algorithms. We apply our framework to a commercially available face tracking solution and present the results of this analysis.

I. Introduction

Detecting and tracking the human face with a camera has a variety of applications such as Human Computer Interaction (HCI). We have developed tools for mapping user behaviour onto avatars in virtual environments and we use head tracking for view point control in tasks where both hands are occupied (e.g., surgery simulations) [1]. Most consumer-level applications track a human face with a single, monocular camera. This inexpensive and easy setup is available to most users and is already capable of delivering the necessary data for interaction with a computer program by means of head movement.

In recent years, technology and algorithms have been steadily advanced and there is a wide selection of tracking methods and implementations available (e.g., [2], [3], or see [4] for an overview). Yet, to the best of our knowledge, there is no study dedicated to setting up a framework that allows for the comparison of different face tracking methods.

We were interested in the ideal conditions in terms of lighting, camera position, etc., that are necessary to achieve good tracking results. For this purpose, we have developed a framework for measuring the influence of changes in those conditions on the quality of a face tracking algorithm.

The results are useful to both developers and users of face tracking algorithms in order to optimise performance or estimate errors where environmental parameters can not be changed. In order to demonstrate the framework, we measure the quality of faceAPI [5], a commercial face tracking system that is freely available for non-commercial purposes. Our framework is also useful for comparing different tracking algorithms.

II. Methodology

A. Definition of Quality

An optimal head tracking system should be capable of measuring the position and orientation of the tracked head precisely and without too much jitter. Ideally, it should also be insensitive to the position of the head in the camera image, changes in lighting, changes in the facial geometry, occluded features, etc.

Head tracking quality can be evaluated in different ways. Many researchers are interested in the absolute error of a tracked position to a ground truth. This can be achieved by using predefined motion paths (e.g., linear motors), a robot arm [2] or a secondary position measurement system.

We found that for most consumer level applications absolute precision is of secondary importance because ordinary users either do not want to or can not calibrate a system appropriately. We found that instead stable results and relative movements are more important. Instabilities such as jitter create very disturbing visual artefacts when rendering the result or using them for navigation. Relative movements are important so that motions and head positions are consistent when repeating them.

We hence define "tracking quality" as the standard deviation of measurements of the stationary head over a certain period of time.

For all experiments, we log the tracked position and rotation of the head for each video frame together with the time since start of the tracking.

Table I: Structure of a logfile used for recording the head tracking experiments.

frame	time.abs	time.rel	state	pos.x	pos.y	pos.z	rot.x	rot.y	rot.z
0	0.000	0.000	started						
51	2.064	2.064	fixed						
51	2.064	0.000	tracking	-26.090	8.118	683.417	-8.011	4.971	3.382
...	...	...	...	...	...	...	...	...	...
298	12.072	0.041	tracking	-26.356	8.036	683.508	-7.315	4.340	3.285
299	12.113	0.041	stopped						

978-1-4244-4698-8/09/$25.00 ©2009 IEEE

To evaluate measurement series, we calculate the 10% trimmed average $pos.avg.A$ and $rot.avg.A$ of all $pos.A_n$ and $rot.A_n$ of a measurement (with $A \in [X, Y, Z]$ and $n \in [1, 2, \ldots, N]$).

After that, we determine the relative positions and rotations $pos.rel.A_n$ and $rot.rel.A_n$.

$$\begin{aligned} pos.rel.A_n &= pos.A_n - pos.avg.A \\ rot.rel.A_n &= rot.A_n - rot.avg.A \end{aligned} \tag{1}$$

Finally, we calculate the relative positions $pos.rel_n$ and rotations $rot.rel_n$.

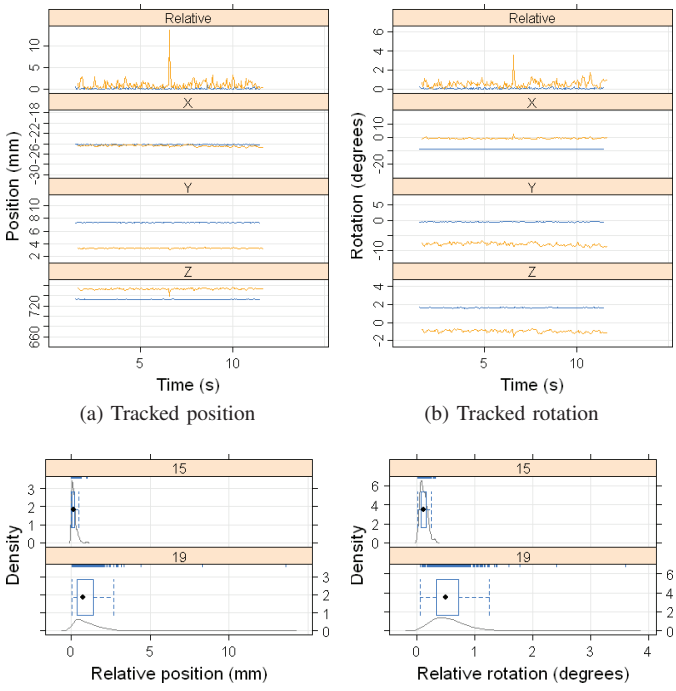
$$\begin{aligned} pos.rel_n &= \sqrt{pos.rel.X_n^2 + pos.rel.Y_n^2 + pos.rel.Z_n^2} \\ rot.rel_n &= \sqrt{rot.rel.X_n^2 + rot.rel.Y_n^2 + rot.rel.Z_n^2} \end{aligned} \tag{2}$$

We assume small angles ($< 10°$) for the relative rotation, so we use the same formula for positional and rotational components. For larger angles, the total angular deviation should be calculated differently.

The final values for $pos.rel_n$ and $rot.rel_n$ are then used for calculating standard deviation, variance, interquartile ranges, etc. We created R-modules [6] for automatically loading, evaluating and plotting these results.

Two sample measurements with a high and low quality together with the corresponding density plots are depicted in Figure 1. The upper row visualises the measured data over time. The lower row shows density plots of the relative measured position and rotation, overlaid with a box-and-whisker plot and a "ceiling rug" of the individual measured data points.

In general, boxplots demonstrating a narrow interquartile distance (the horizontal size of the blue box) serve as an indicator for high quality tracking results, where larger interquartile distances are an indicator for more jitter and thus a lower quality of the tracking results.



(a) Tracked position (b) Tracked rotation

(c) Density plot of the tracked relative position (d) Density plot of the tracked relative rotation

Figure 1: *Upper row*: Sample data for two measurements. Measurement 15 (blue) is of good quality, whereas Measurement 19 (orange) demonstrates strong jitter and bad quality. *Lower row*: Density plots are used to visualise the amount of jitter.

B. Hypotheses

We set up the following list of hypotheses to be validated by our experiments.

1) The position of the light source influences the quality of tracking. Frontal lighting is better than light coming from the sides.
2) The intensity of the light source influences the quality of tracking. More light results in better quality as it increases contrast, reduces the necessary camera exposure time and decreases noise and blur. Too much light decreases the tracking quality as image details are lost in saturated pixels.
3) The position and rotation of the head within the camera view volume influences the quality of tracking. A closer distance to the camera (Z-axis) enables a higher quality of tracking as the head appears larger and can be tracked more accurately. The lower limit is the Z-position when the camera image of the head gets too big and parts of the facial area lie outside of the image. In contrast, the deviation from the central axis in X and Y direction should not influence the quality of the tracking as it only changes the position of the head in the 2D image, but creates no distortion or change of size.
4) The rotation of the head within the camera view volume influences the quality of tracking. Less rotation around the X-axis (pitch) and Y-axis (yaw) improves the quality of the tracking. Rotation around theses axes results in non-planar distortion of facial features in the 2D image and requires more work to keep track of these features. In contrast, rotation around the Z-axis does not influence the tracking quality. The facial features are only rotated but not distorted or occluded. Tracking should not be negatively influenced by this rotation.
5) Occlusion of facial features by hair, glasses, beard, headset, etc. influences the quality of the tracking. The quality of tracking is higher when less facial features are occluded. The tracking quality is influenced most when key features like eyes or mouth are occluded.

This list is by far not exhaustive, and we discuss further items that could be included in Section IV.

C. Setup and Equipment

The general setup consists of the following items (see Figure 2):

1) Camera: For the experiments, we use a Logitech QuickCam Express Go webcam, mounted on a tripod. The camera has a resolution of $640\,\text{px} \times 480\,\text{px}$ and a Field of Vision

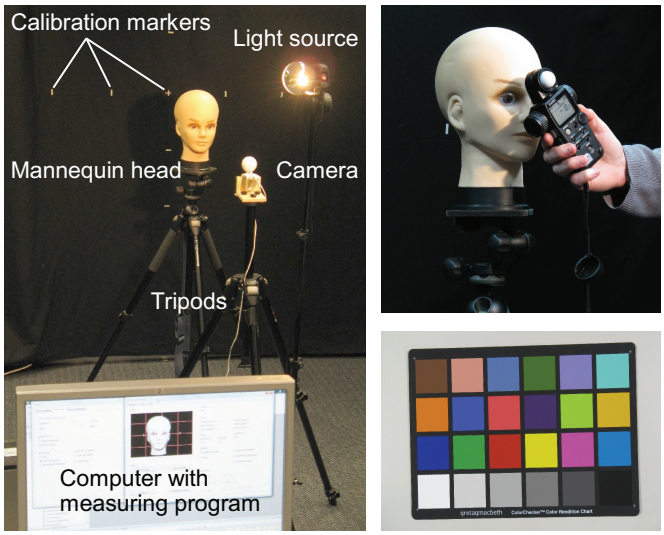


Figure 2: Photos of the experimental setup in the Computer Vision Laboratory. *Left*: General setup with the light source, the head and the controlling computer. *Top right*: The used light meter and the position during measuring the amount of incident light. *Bottom right*: The used colour calibration chart.

(FoV) of 32.4° (acquired during the calibration step described in Section II-D2).

2) Mannequin Head: We use a life-sized mannequin head for our experiments The head is mounted on a tripod and can be rotated manually along the X, Y, and Z-Axis.

3) Light Source: A single 150 W light bulb is used as a light source. By mounting it on a tripod, were are able to vary its position and direction. More advanced lighting systems would also allow to control the brightness directly and attach a diffuser for creating diffuse light.

To avoid the influence of other light sources, the walls and windows of the room are fitted with black, light absorbing cloth.

4) Light Meter: To ensure equal lighting conditions for the measurements, we utilise a Sekonic L-758D light meter. Though this device is not capable of displaying an absolute value of incident light, we use a special "exposure value" mode, which displayed a relative reading that guaranteed equal lighting for all measurements (see Figure 2, top right).

5) Colour Calibration Chart: To ensure equal light colour conditions for the measurements, we use a Gretag Macbeth™ColorChecker™Color Rendition Chart to adjust the camera colour controls (see Figure 2, bottom right).

6) Coordinate System: For physical measures, markers are attached to the wall opposite of the camera and on the floor. They define a basic coordinate system for the physical measurements and are used during the calibration phase (see Section II-D).

7) Software: A computer program controls the camera and records the measured positions of the head. The program also assists in correctly aligning objects in the field of vision of the camera and to detect overexposure.

D. Calibration Phase

The calibration phase consists of three steps:

1) Geometric Setup: The camera is mounted on the tripod, using an L-shaped support plate. With the help of an alignment grid in the measuring program, the camera can then be adjusted so that it points directly at the centre of the calibration pattern and is not tilted in any way.

2) Calculation of the Camera FoV: The calibration pattern consists of a centre cross and small lines in 25 cm intervals (see Figure 2). With $w_{C,R}$ being the physical horizontal size of the calibration pattern, $w_{C,C}$ being the width of the calibration pattern in the camera view, $w_{S,C}$ being the horizontal resolution of the camera, and l being the physical distance of the camera from the calibration pattern, the following formula can be used to calculate the FoV of the camera:

$$FoV = 2 \cdot arctan\left(\frac{w_{S,C} \cdot w_{C,R}}{w_{C,C} \cdot 2 \cdot l}\right) \quad (3)$$

For the Logitech camera, with $w_{C,R} = 1\,\mathrm{m}$, $w_{C,C} = 550\,\mathrm{px}$, $w_{S,C} = 640\,\mathrm{px}$, and $l = 2\,\mathrm{m}$, we obtained a FoV of 32.4°

3) Colour Calibration: The colour calibration chart is placed in front of the camera. Using the white balance controls of the camera, the picture is adjusted as so that the colours show no tint.

4) Lighting Calibration: Our measuring program provides an indicator for saturated pixels. Saturated pixels have at least one of their Red/Green/Blue values at maximum, and may degrade the performance of a facial tracking algorithm as colour information is lost.

After the mannequin head has been set up in front of the camera, the lighting has to be adjusted so that a bare minimum of saturated pixels is visible. When this is done, the lighting meter is used to measure the relative exposure value (EV) directly in front of the mannequin head.

In our setup, we measured an exposure value of 10.0 EV, and kept this value constant for all experiments except for measuring the influence of lighting intensity (see Section III-B).

III. Results

This chapter summarises the main aspects of our evaluation framework and the corresponding results for the faceAPI. More detailed results can be found at [7].

A. Position of the Light Source

For this experiment, we moved the light source around the head to predefined positions. Assuming horizontal symmetry, we only covered positive yaw values of the light source. Also, as tracking failed for all pitch angles below 10°, we reduced the amount of tested positions below 0° pitch. Room space constraints also limited the light source position to a maximum pitch angle of 40° and a maximum yaw angle of 30°.

For every position, we initiated tracking 5 times, stopping either after 10 s of unsuccessful tracking or 10 s after successful tracking.

For the evaluation (see Figure 4), we calculated the relative tracking positions and rotation as described in Section II-A

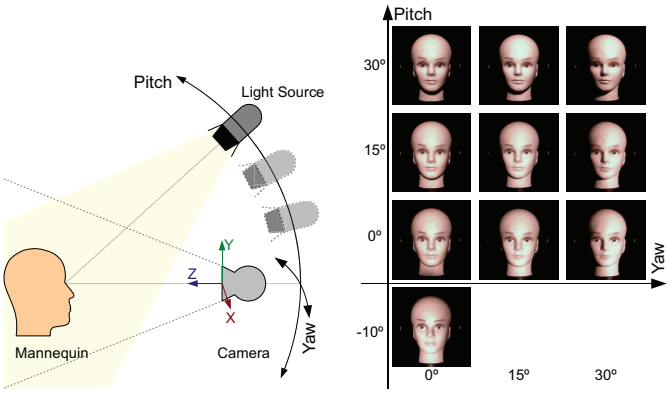
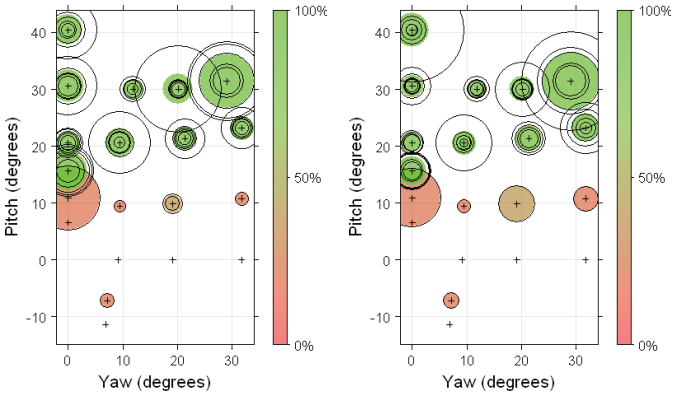


Figure 3: *Left*: Schematic view of the experimental setup. *Right*: Camera view of the mannequin head with varying lighting positions.



(a) Standard deviation of the tracked relative position

(b) Standard deviation of the tracked relative rotation

Figure 4: Evaluation of measurements with varying lighting positions.

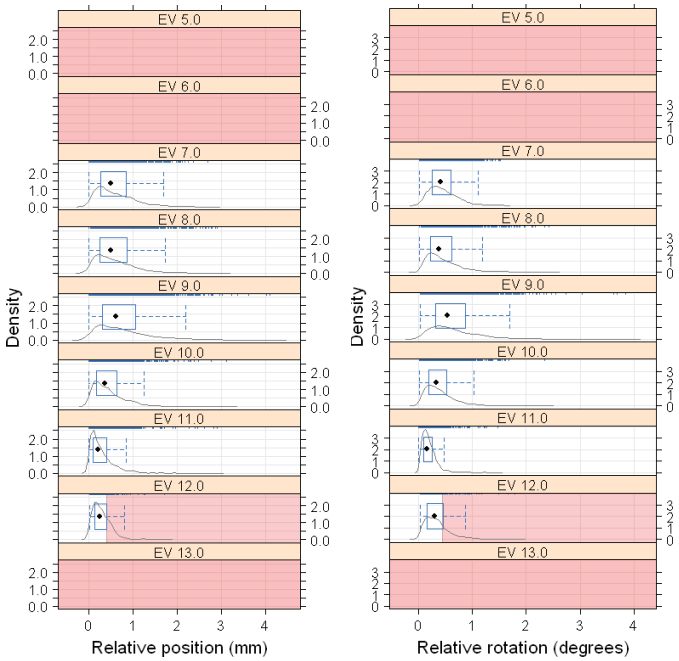
and plotted the standard deviation of these values at the corresponding pitch/yaw position.

The black circles indicate the individual standard deviation for each single tracking process, the size of the coloured circles indicates the average of the standard deviation for all tracking processes at that light source position. The colour of the circle indicates the percentage of successful tracking. Some of the tracking results in the regions with $pitch \leq 10°$ demonstrate a low standard deviation, but tracking only succeeded in one out of five attempts. Thus, results with colours other than bright green should be disregarded.

Our results confirm hypothesis 1: The optimal position for the light source lies within a cone of $20° \leq pitch \leq 30°$ and $-20° \leq yaw \leq 20°$. Jitter increases for $yaw > 20°$ and for $pitch > 30°$. In addition, for all $pitch < 20°$, tracking does not perform reliably.

B. Intensity of the Light Source

Determining the ideal lighting intensity was achieved by varying the distance of the light source to the mannequin head. We varied the measured lighting intensity i at the head in full steps from $5\,\mathrm{EV}$ to $13\,\mathrm{EV}$. For each lighting intensity, we



(a) Density plot of the tracked relative position

(b) Density plot of the tracked relative rotation

Figure 5: Evaluation of the influence of lighting intensity.

initiated tracking 5 times and recorded the tracking results for $10\,\mathrm{s}$ or stopped after $10\,\mathrm{s}$ of unsuccessful tracking (see Figure 5).

For $i \leq 6\,\mathrm{EV}$ and $i \geq 12\,\mathrm{EV}$, tracking did not succeed at all or only partially (represented by the red bars in the density plots). The tracking quality is best for $i = 11\,\mathrm{EV}$, though this value already creates a large amount of saturated pixels.

We would recommend $i = 10\,\mathrm{EV}$ as the ideal lighting intensity, because this value produces only very few saturated pixels and thus guarantees more headroom in case the user moves away or towards the light source.

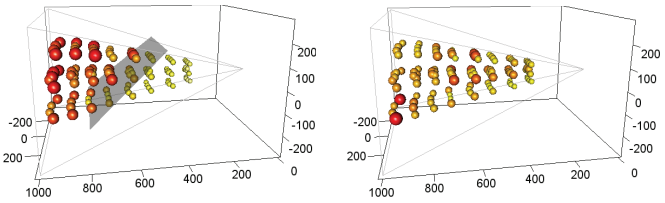
Our results confirm hypothesis 2: For optimum tracking results, lighting has to be chosen as bright as possible without causing saturated pixels.

C. Influence of the position in the view volume

To verify hypothesis 3, we positioned the head in regular intervals in the view volume of the camera. Starting at $Z = 30\,\mathrm{cm}$ distance, we used an alignment grid in the measuring program (visible on the computer screen in Figure 2) to move the head into 6 offset positions on the X- and Y-axis. The centre of the eyes was placed on the crossing points of the grid. Similar to experiment III-A, we assumed horizontal symmetry and only covered the positive X-offsets.

The six X- and Y-offset positions were repeated while moving the head along the Z-axis away from the camera in $10\,\mathrm{cm}$ steps, covering a total range of $30\,\mathrm{cm} \geq Z \geq 100\,\mathrm{cm}$.

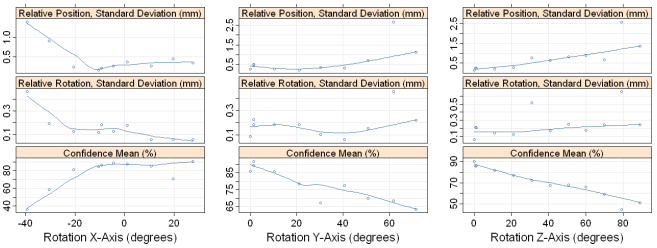
The tracking was initiated at the beginning and continued for the duration of the experiment. For each position, we recorded $10\,\mathrm{s}$ of data.



(a) Standard deviation of the tracked relative position

(b) Standard deviation of the tracked relative rotation

Figure 6: Evaluation of the measurements with different head positions. The colour and diameter of the spheres indicate the quality of the tracking results – small/yellow:good, large/red:bad.



(a) Pitch (X-Axis) (b) Yaw (Y-Axis) (c) Roll (Z-Axis)

Figure 7: Evaluation of the measurements with different head rotations.

The position on the X- and Y-axis seems not to be of significant influence on the quality of the tracking, in contrast to the position on the Z-axis. The grey plane in Figure 6a marks the 50% boundary between good and bad quality results. Good tracking quality is achieved for head positions $80\,\mathrm{cm} \geq Z \geq 30\,\mathrm{cm}$ with a tendency to better results in the lower half of the viewing cone.

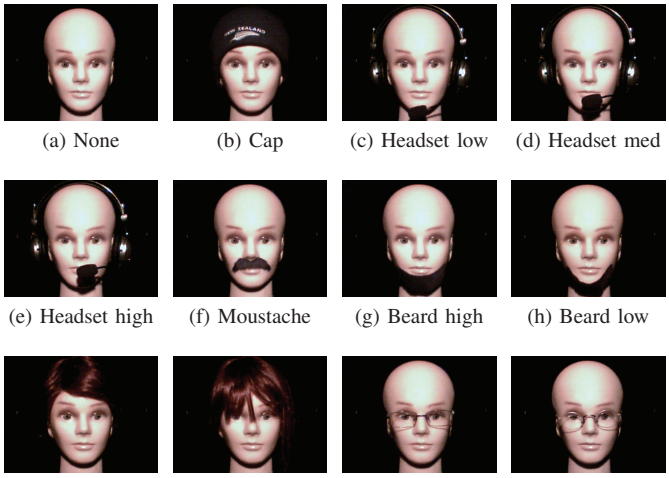
The reason for the tracking results becoming worse at the far end of the working space is that the projected size of the head gets smaller and thus, less pixels can be used for the tracking process, resulting in increased jitter for the position result.

Our results confirm hypothesis 3: The quality of tracking depends mostly on the position on the Z-axis. The most accurate results are achieved with a distance to the camera of $80\,\mathrm{cm} \geq Z \geq 30\,\mathrm{cm}$.

D. Influence of the rotation

To verify hypothesis 4, we tilted the head in $10°$ steps independently along all 3 axes. The tripod head we used for mounting the mannequin head allowed for a rotation range of $-30° \leq pitch \leq 180°$ for the X-axis, $360°$ for the Y-axis, and $0° \leq roll \leq 180°$ for the Z-axis. During the rotation, we kept the head at a constant distance of $60\,\mathrm{cm}$ to the camera and always fully visible inside the camera frustum. The range of the angles along the three axes was limited by the range of the tripod head or when tracking was lost.

We started the tracking on the initially unrotated head, and continued during the rotation along one axis. Each orientation was measured for $10\,\mathrm{s}$.



(a) None (b) Cap (c) Headset low (d) Headset med

(e) Headset high (f) Moustache (g) Beard high (h) Beard low

(i) Hair, Style 1 (j) Hair, Style 2 (k) Glasses 1 (l) Glasses 2

Figure 8: Camera view of some of the different setups for occluded facial features.

In general, rotation along any axis away from the neutral position influences the tracking results negatively. The increase in jitter is most apparent for negative pitch angles (X-axis), when the head points downwards (see Figure 7).

For all other directions, the influence of the angle on the amount of jitter is more or less linear. For the yaw angle (rotation around the Y-axis) we only looked at positive angles, assuming horizontal symmetry again as in experiment III-A.

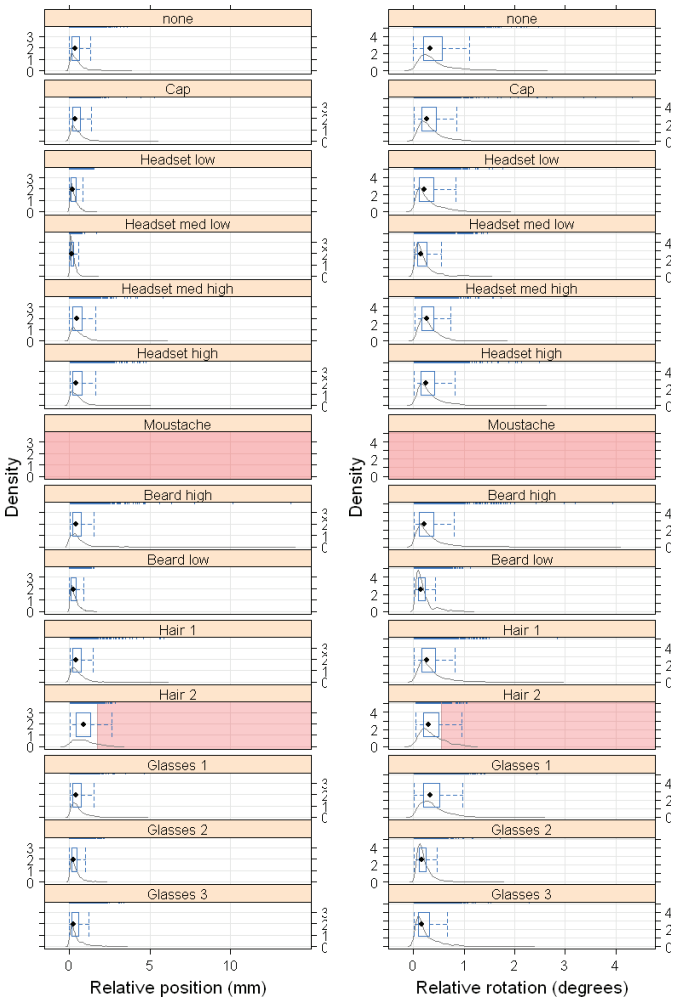
Our results confirm parts of hypothesis 4: Rotation around the X- and Y-axes negatively influences the quality of the tracking. Nevertheless, our findings of negative influence of Z-axis rotation are unexpected. However, the impact on our application is not that significant, as the head roll angle is anatomically limited to $54°$ with a comfortable maximum value of $20°$ [8], thus being only about half the angular range we tested.

E. Influence of occluded facial features

Tracking algorithms should be very robust against partial occlusion of facial features. We tested the influence of headgear, headsets, beards, hair, and glasses on the tracking quality (see Figure 8). For each occlusion element, we initialised tracking 5 times and recorded the tracking results for $10\,\mathrm{s}$.

Figure 9 demonstrates that tracking failed completely for the moustache and in 4 out of 5 tracking attempts for the hairstyle partially occluding the eyes. This reflects the fact that the eyes and the part between the nose and the mouth are playing a crucial part for recognising a face in the video image.

All other occluded features resulted in 100% tracking, but as we expected, with different quality. Objects occluding the mouth, like a high headset microphone or a chin beard covering the area from the lower lip down to the chin resulted in tracking data severely deviating from the expected position. Interestingly, a beanie even improved the quality of tracking compared to the bald head, possibly because it reduces the amount of skin coloured pixels that are candidates for the



(a) Density plot of the tracked relative position

(b) Density plot of the tracked relative rotation

Figure 9: Evaluation of the measurements with different occluded facial features.

face region. Glasses with thick frames (Glasses 1) had more negative impact on the results than glasses with no or thin frames.

The headset had no influence as long as the microphone was located below the lower lip. With the microphone covering the lip, tracking quality deteriorated.

Generally, occluded facial features lead to a lower quality of tracking results, which supports hypothesis 5. As long as the most important facial features, like eyes, nose, and mouth, are clearly visible, tracking is successful. Increasing occlusion of these key features leads to lower tracking quality.

IV. Conclusion and Future Work

We have described a series of experiments to prove our hypotheses about the influence of several factors on the quality of head/face tracking, in our specific case for the faceAPI. The majority of these hypotheses has been proved with a minor number of surprises, like the influence of the head roll angle (see Section III-D).

The general recommendation for good tracking results are as follows:

Lighting: The light source should shine onto the face from within a cone of $20°$ and $30°$ pitch and $-20°$ and $20°$ yaw.

Workspace: The ideal workspace lies between $30\,\mathrm{cm}$ and $80\,\mathrm{cm}$ distance of the camera with a tendency to the lower part of the camera view frustum. This, of course, depends heavily on the focal length of the lens used in the specific camera model.

A good position for the camera is the upper rim of the monitor or the laptop lid. The camera should be centred horizontally on the rim and should not be tilted or rolled at any angle.

User: The user should take care that the face is not occluded by items like the microphone of the headset or long hair. Hair can be tied back, and a headset can be adjusted so not to occlude the lips.

Depending on thickness and colour, beards might cause problems when they cover parts of the lips and/or the space between the nose and the upper lip.

Glasses and headgear do not reduce the quality of the tracking results as long as eyes and eyebrows are clearly visible.

The list of hypotheses in Section II-B is not exhaustive and will be extended in the future. A planned extension is the influence of camera models. With different lens parameters and/or different sensitivities, different camera models would gravely influence some of the presented results. But, when extending the framework, one has to weigh the benefits of gaining additional information against the additional workload that this certain aspect might bear. By adding another camera, it would be necessary to repeat all experiments. In contrast to this "multiplying" influence, other framework additions would result in only one more experiment.

References

[1] S. Marks, J. Windsor, and B. Wünsche, "Enhancing Virtual-Environment-Based Teamwork Training with Non-Verbal Communication," in *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2009*. Honolulu, HI, USA: AACE, Jun. 2009, pp. 4133–4144. [Online]. Available: http://www.editlib.org/p/32078

[2] Y. Matsumoto, N. Sasao, T. Suenaga, and T. Ogasawara, "3D Model-based 6-DOF Head Tracking by a Single Camera for Human-Robot Interaction," *IEEE International Conference on Robotics and Automation*, 2009.

[3] R. Gross, I. Matthews, and S. Baker, "Active Appearance Models with Occlusion," *Image and Vision Computing*, vol. 24, no. 1, pp. 593–604, 2006.

[4] J. J. Wang and S. Singh, "Video analysis of human dynamics–a survey," *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, 2003.

[5] Seeing Machines. (2009) faceAPI. [Online]. Available: http://www.seeingmachines.com/faceAPI.html

[6] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[7] S. Marks, J. Windsor, and B. Wünsche, "Evaluation Framework for Face Tracking Algorithms," The University of Auckland, Tech. Rep., Sep. 2009. [Online]. Available: http://www.cs.auckland.ac.nz/~stefan/documents/StefanMarks%20-%20FaceTrackingEvaluationFramework.pdf

[8] A. R. Tilley, *The Measure of Man and Woman*, 2nd ed. John Wiley & Sons, 2002.