Improving Head Pose Estimation in Radiation Therapy Through Photogrammetric Techniques for Machine Learning Applications

Cyrill Milkau [1], Sebastian Preußel [2], Sarah Guy [3], Danilo Schneider [1]

[1] Faculty of Spatial Information, HTW Dresden – University of Applied Sciences, Germany – cyrill.milkau@htw-dresden.de, danilo.schneider@htw-dresden.de
[2] Institute of Photogrammetry and Remote Sensing, Dresden University of Technology, Germany–sebastian.preussel1@tu-dresden.de
[3] Department of Radiotherapy and Radiation Oncology, Dresden University of Technology, Germany – sarah.guy@ukdd.de

Keywords: pose estimation, keypoint detection, 3d-object observation, uncertainty estimation, bundle adjustment, marker-based tracking

Abstract

Reliable head pose estimation is critical for human activity monitoring, yet machine-learning (ML) approaches typically provide no explicit measures of uncertainty or reliability. We propose a hybrid strategy that embeds ML-based pose estimation within a rigorous photogrammetric framework. A calibrated multi-camera system observes both static reference markers and dynamic markers attached to the human head. Through bundle adjustment, 3D marker positions together with uncertainty, redundancy, and correlation measures are estimated. By modelling the head-mounted markers as an inner sub-network, stability and reliability indicators are derived and reprojected into image space, directly informing markerless ML-based keypoint detection. We validate the approach on synthetic data and a real test measurement in the context of radiotherapy patient monitoring. Results show that the framework reliably separates rigid head motion from local facial deformation, achieving a signal-to-noise ratio above 6 for face deformation detection, and that studentized per-point stress indicators effectively partition the marker network into deformation-active and noise-dominated subsets, providing a principled basis for spatially adaptive keypoint weighting

1. Introduction

Facial landmarks are commonly extracted from 2D heatmaps and used in head motion estimation without explicit consideration of their geometric configuration. In practice, these points are often deployed in uncontrolled, "in-the-wild" settings, where their spatial arrangement and statistical properties—such as spatial consistency, positional stability, and distribution—are not optimized for measurement sensitivity. This work presents a photogrammetric framework to enhance the robustness of such systems by leveraging the geometric and statistical characteristics of 2D landmark estimates to inform the design of the measurement setup. By aligning marker placement and camera configuration with the inherent reliability and spatial structure of the detected points, the system becomes more sensitive to meaningful head and facial deformations. The approach is validated on post-processed synthetic and real data. Limitations for real-time implementation are discussed. The results suggest that geometric and statistical properties of 2D heatmap outputs—such as spatial consistency and positional stability—can guide the design of a more robust and deterministic measurement setup.

In (Hempel et al., 2022) predictions of rotation matrices are estimated based on a ML model applied on image crops (Chun et al., 2024) uses bidirectional regression that combines 3d facial landmark reconstruction with head pose assuming a geometric transformation model, though assuming pinhole model yet. Other works focus on morphable models like FLAME (Feng et al., 2020) mostly with close-range or wide-angle conditions without proper extended distortion correction. In their work, (Li et al., 2024) addressed fisheye distortion for head pose estimation by learning a location-dependent correction within the network rather than calibrating the lens explicitly. Our approach differs fundamentally from the deep learning-dominated head pose estimation literature by grounding head and face motion analysis in a rigorous multi-camera photogrammetric framework also benefiting from the fact that limiting the approach to operating inside a controlled environment is highly relevant for the targeted field of application. The main motivation behind this work is to improve the patient positioning during radiotherapy treatment by gating head and face deformation throughout the radiation session. This could replace the current demand for immobilization masks which can cause anxiety and stress for the patient while enforcing their head in a fixed position.

2. Proposed strategy

The core idea of this approach is to describe the motion of a rigid object as a function of derived parameters. In practice, this target object is approximated by a set $\boldsymbol{S}$ of 3D points, along with several weak constraints (e.g., connecting edges) or strong constraints (e.g., fixed scale). The motion is described by grouping the 3D points into clusters and evaluating their change over specific time periods (epochs). A Helmert transformation is then applied to the same set at two individual epochs, with the objective of decomposing the determined movement between them into a rigid component and a non-rigid residual. First, the centroids $\overline{\boldsymbol{P}}^{ref}, \overline{\boldsymbol{P}}^{cur}$ (i.e., of same points at different epochs) are used to calculate the centred coordinates $\boldsymbol{p_i}, \boldsymbol{q_i}$. Then, a scale factor s_{pq} (optionally), a rotation $\boldsymbol{R}_{pq}$ and translation $\boldsymbol{t}_{pq}$ are determined using singular vector decomposition (SVD) on the cluster, shift centroid and then determine residuals $\boldsymbol{r}_i$ between the transformed $\boldsymbol{P}_i^{cur}$ and the initial $\boldsymbol{P}_i^{ref}$. These represent each point’s displacement component that is not covered by any rigid-body motion. The total deformation energy $E^{\hat{c}}$ stored in the current epoch (over the current set $\boldsymbol{S}^{cur}$ with respect to the reference set $\boldsymbol{S}^{ref}$) is a scalar summary of how much the point configuration deviates from rigidity. Based on these metrics, two absolute

motion indicators are introduced for each cluster $\hat{c}$: Face motion indicator $F^{\hat{c}}$ and head motion indicator $H^{\hat{c}}$ as well as a combined log-ratio indicator $\lambda^{\hat{c}} = \ln(F/H)$. If face motion dominates, then $\lambda^{\hat{c}} > 0$, and $\lambda^{\hat{c}} < 0$ vice versa. For point-base evaluation a deformation intensity $s^{\hat{c}}$ is defined in which vertices (3D points of $\boldsymbol{S}$) and edges (connections between them) are considered. In between steps are necessary which will explained in the following sections. The full workflow is shown in Figure 1. The main objective of this work is showing the general concept of the motion indicators and deformation intensity without the overhead of real-time evaluation, therefore everything is evaluated in post-processing.

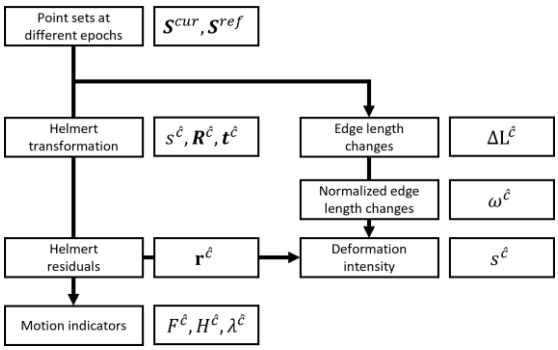

Figure 1. Workflow

2.1 System overview

The hardware setup consists of four 10-GigE cameras connected to a frame grabber video card inside a workstation. The cameras are mounted on the wall and ceiling to ensure stability (see Figure 2). Two separate C++ applications, backed by a static linked library and connected via TCP/IP network to exchange intermediate results, form the core software structure. The system builds upon the design presented in (Milkau et al., 2026).

A local cartesian coordinate system X^w was used as the reference frame system. It is realized as a right-handed system with an arbitrary root point and arbitrary orientation towards a fixed reference target point. A test field of 98 evenly distributed rotational invariant marker control points is built into the scene. All control points coordinates are determined within X^w using a Leica MS60 total-station. The cameras exterior orientation parameters (EOP) are determined in X^w by spatial resection using direct linear transformation (DLT) on the control markers. To determine the interior orientation parameters (IOP) for each camera, a-priori calibration is performed. To achieve this, a calibration cube marked with control points is placed inside the measurement space and captured by the cameras for calibration. With IOP and EOP j individual camera and image coordinate systems X^{c_j}, X^{I_j} are defined. Observations in X^{I_j} can be transformed to world coordinates in X^w. The general data flow from acquisition of 2D data to estimation of 3D data is shown in Figure 3. The IOP and EOP are assumed constant over the short period of acquisition time. While control markers are associated to control points and therefore assumed to remain fixed in space, observation markers are linked to observation points which are the main objective in the detection pipeline. For convenience, certain marker ids are preserved for the control points, all others are treated as observation points. The main goal of this setup is to be able to track observation points position over time.

Figure 2. top: Camera setup; bottom: Test field

2.2 Position estimation

To obtain 3D observation point coordinates, a simple bundle adjustment was implemented. The pose of each camera is transmitted once to the evaluation software prior to the acquisition procedure. These parameters are used to estimate each incoming 2D observations point initial position using DLT - warm-up phase (I). After some iterations, the bundle adjustment (II) solves for each point individually (no parametric constrains between 3D points – free network adjustment). The 3D observation points are determined directly in X^w. Projecting any 3D point $\boldsymbol{P}$ from X^w in X^{I_j} with the extended collinearity equations, IOP $\boldsymbol{i}_j = (c_k, x_0, y_0, A_1, A_2, A_3, B_1, B_2, C_1, C_2)$ (plus distortion model) and EOP $\boldsymbol{e_j}(\boldsymbol{R}_j, \boldsymbol{t}_j)$ is done by

$$\boldsymbol{p}_j = \boldsymbol{R}_j(\boldsymbol{P} - \boldsymbol{t}_j) \tag{1}$$

In (I) the linear system of the projection is solved by finding a non-trivial solution $\boldsymbol{P_j} = X_{h4}^{-1}\,[X_{h1} \quad X_{h2} \quad X_{h3}]$. Based on this initial position the following 2D marker observations are incorporated into (II). The structure is the common approach of the parameter vector $\boldsymbol{x_{M\times 1}}$ (M 3d observation points $\boldsymbol{P_{i=0:M-1}}$ in X^w), the observation vector $\boldsymbol{l_{2N\times 1}}$ (N image coordinates of 2d observation markers), the functional model $f(\boldsymbol{x}) = \pi(\boldsymbol{P}, \boldsymbol{e}, \boldsymbol{i})$ and design matrix $\boldsymbol{A} = \partial_{\boldsymbol{x}} f$ – all solved by Gauss-Newton solver with optional Levenberg-Marquardt damping (Luhmann et al., 2023). The general uncertainty estimation is done straightforward. With a-posteriori variance of unit weight σ_0^2 the point covariance and cross-covariance are estimated via $\boldsymbol{\Sigma}_{\mathrm{xx}} = \sigma_0^2 \boldsymbol{N}_r^{-1}$ with regularized normal matrix. And following standard deviation $\sigma_{\mathrm{x_k}} = \sqrt{(\boldsymbol{\Sigma}_{\mathrm{xx}})_{kk}}$ and correlation $\rho_{\mathrm{kl}} = (\boldsymbol{\Sigma}_{\mathrm{xx}})_{kl}/(\sigma_{\mathrm{x_k}} \cdot \sigma_{\mathrm{x_l}})$. Additionally, the IOP uncertainty contribution to parameter covariance $\boldsymbol{\Sigma}_{\mathrm{x}}^{\boldsymbol{i}}$ is propagated through the normal equations, so the total parameter covariance adds to $\boldsymbol{\Sigma}_{\mathrm{x}}^{Tot} = \boldsymbol{\Sigma}_{\mathrm{xx}} + \boldsymbol{\Sigma}_{\mathrm{x}}^{\boldsymbol{i}}$ (see Appendix B for further details). The sub-pixel ellipse centre’s covariance σ_{uv}^2 is derived from the fitted ellipse semi-axes. The bundle solves only for the 3D observation point coordinates (Σ_{xx}) yet the IOP and EOP influence is modelled independently by Σ_I and $\Sigma_E$and aggregated yielding a final Σ_x.The per-point covariance matrix Σ_{P_i} (diagonal block of Σ_x) contains the per-point covariance $\sigma_{P_i}^2$ which contributes in the edge length variance σ_{e}^2.

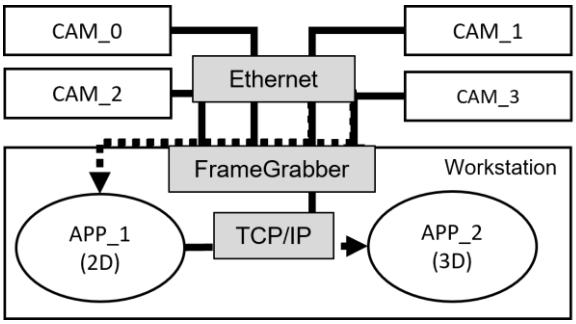

Figure 3. Data flow - camera buffer is fed into the target application via Ethernet; 2d observations are then passed internally to an evaluation application for 3d estimation.

3. Synthetic data

A proof-of-concept for the presented approach is the use of synthetic prior to real data. For that a face net like in Figure *4* is defined as set of 3D points with a perfect symmetric distribution around the axes $\overline{\boldsymbol{P}_1\boldsymbol{P}_{27}}$ and $\overline{\boldsymbol{P}_2\boldsymbol{P}_{28}}$. The centroid of this net is then placed in X^w at an arbitrary position and orientation within the covered measurement space. By projecting the set into X^{I_j} along with some normal-distributed noise, the synthetic 2D points can be used to re-project into X^w following (x) and the normal adjustment pipeline. For zero-noise the 3D position is identical to the initial 3D set. All experiments are post-processed results from pre-captured image buffers. Certain transformation chains were then applied to the synthetic set, to mimic three real world kinematic motion scenarios: static head with static face (SHSF), static head with dynamic face (SHDF) and dynamic head with static face (DHSF). These included rotation and translation of points to simulate movement (see Table 1).

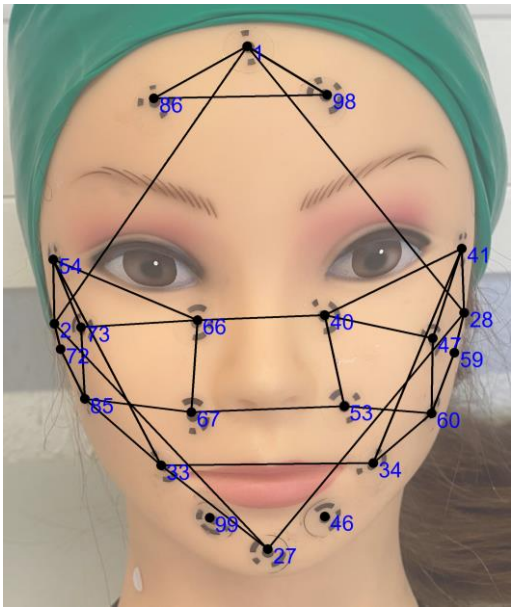

Figure 4. Twenty-two Observation markers of the face net

The evaluation steps are clearly separated and build up on each other:

(1) Epoch-wise Acquisition of image buffer epoch-wise (epoch is one simultaneously capture of all cameras)
(2) Epoch-wise Marker detection in each image
(3) Epoch-wise bundle adjustment based on four camera views
(4) Epoch-aggregated analysis on the adjustment results for each scenario to retrieve motion indicators and deformation intensity

Scenario	Motion sequence
SHSF	No transformation
SHDF	Shift $\pm$10mm to centroid
DHSF	Rot. +45° - Rot. -90° - Rot. +45°

Table 1. Analysed Scenarios of synthetic motion

Each individual bundle adjustment result is per se independent from the previous ones (no inter-connections). To evaluate net deformations, either the scenarios first n epochs or an epoch-aggregated SHSF scenario serves as reference (both are modeled without any motion signal). Then, each single epoch is transformed against the reference epoch (see section 2.2). Scale is fixed as this would correlate with the to be observed deformation. Upon real deviation this will yield strain on the vertices and their connecting edges, which is modeled by the Helmert residual:

$$\boldsymbol{r}_i^c = \boldsymbol{P}_i^c - \left(\boldsymbol{R}^c \boldsymbol{P}_i^{ref} + \boldsymbol{t}_i^c\right) \tag{2}$$

The total deformation energy “stored” in that transformation divided by all points yields the face motion indicator

$$F^c = \sqrt{\sum_{i=1}^{n} \left\|\boldsymbol{r}_i^c\right\|^2 / n} \tag{3}$$

With the transformation’s axis rotation angle θ and a rotation-to-length scaling factor κ

$$H^c = \|\boldsymbol{t}_i^c\| + \kappa \cdot \theta \tag{4}$$

is the head motion indicator – λ^c follow from section 2. The sets edge length changes ΔL as well as the Helmert residuals are normalized by their standard deviation (Niemeier, 2008) yielding the final deformation intensity

$$s_i^c = \frac{r_i^c}{\sigma_r} + \frac{\Delta \mathrm{L_i^c}}{\sigma_{\Delta L}} = \frac{r_i^c}{\sigma_r} + \omega \tag{5}$$

Figure 5 shows the result for the synthetic dataset. Here the aggregated SHSF with small random noise is taken as reference scenario. The upper plots show the expected behavior of the motion sequence in Table 1 for SHDF (top left: face motion triggers and stays constant until small increase at the end) and DHSF (top right: sinusoidal movement of head w.r.t reference). The bottom shows the combined indicator which follows the signal for DHSF but acts noisy on the SHDF. This could be due to the fact that the face deformation is too much compensated by a fairly good head transformation, that the two indicators are indeed correlated (see top right). For real data this high correlation is not expected and the log-ratio indicator should follow the cleaner signal.

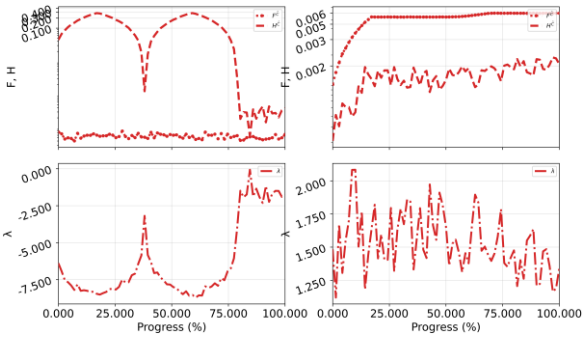

Figure 5. Motion indicators (synth) for DHSF (left) and DHSF (right); top: H (dotted) and F (dashed) with log-scale; bottom: λ

4. Real data

In contrast to the synthetic motion sequence, the real person’s movements like head rotation or facial deformation are based on more complex motion chains and modelling them is not trivial. The system is therefore designed to abstract the head-face region with the markers and their individual relations (face net) and observes them with the cameras. As of now the system operates in software-synchronization only, as there is currently no external hardware trigger available. Thus, the operation principle is: camera stream runs asynchronously and occasionally the most recent frame buffers are retrieved using synchronous threading inside the application. The temporal offset between each camera’s buffer is assumed to be constant and mainly determined by the frame rate (i.e., effective buffer size) and Ethernet speed. The acquisition rate $1/\Delta t_{acq}$ is therefore set empirically to ensure that each grab delivers the next available frame. When it is too high, frames may get processed multiple times, when it is too low, frames may get skipped. The camera setup intends to cover the full head-face region and ideally achieve redundant observations for as many points as possible.

Scenario	Motion sequence
SHSF	Person shows no head or face motion
SHDF	Person grimaces and relaxes periodically
DHSF	Person rotates head sagittal lateral

Table 2. Analysed scenarios of real motion

The observation markers are made of 22 self-adhesive medical face patches on which the marker code is printed. They are attached onto the target persons face following general symmetry and distribution to ensure overlapping visibility, creating a face net like Figure 4. The to be observed test person is then placed inside the test field (see Figure 2 bottom) and observed via the camera setup. Each cameras buffer is read via the API and fed into the target application. After the buffer is received (A), it is evaluated inside a pre-defined region of interest (ROI) (B) on which a custom marker detection algorithm is applied to identify possible markers (C). The markers 2d pixel coordinates are the main output (D) and transferred internally to another evaluation application for further computation. Figure 6 shows the individual steps A-D. Step (A) is performed in real-time, while (B)-(D) are in post-processing. The same scenarios from Table 2 were performed – this time by a real test person.

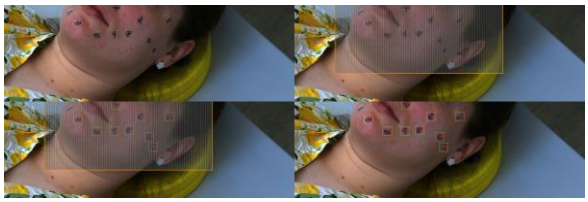

Figure 6. Image buffer (A; top left) is ROI-overlayed (B; top right) for marker detection (C, bottom left) and tracking (D, bottom right)

The pixel coordinates of m elliptical observation markers in camera j are defined as a set of $\{(u_{jn}, v_{jn})\}$. They originate from fitting a robust ellipse into a connected contour with the general conic equation that needs to be minimized. Given that elliptic observation markers move in space, the pixel coordinates change in each frame. To account for that, a meaningful region of interest (ROI) is defined and a nonlinear-velocity Kalman filter applied. The ROI limits the domain of the Kalman filter for all possible markers while the latter further specifies the per-marker location search space. For general least-squares fit of an initial ellipse the library OpenCV (Bradski, 2000)is used together with some image enhancing techniques such as geometry filter for rejecting outliers before refinement. The detection is based on the results of (Hardner et al., 2024) but extended by CUDA-support for GPU acceleration and the Kalman filter described in (Milkau et al., 2026) (reducing the search space for the initial ellipse to a minimal extend). By updating the center and dimension of the ROI it accounts for the expected movement of markers during the tracking (see Figure 7).

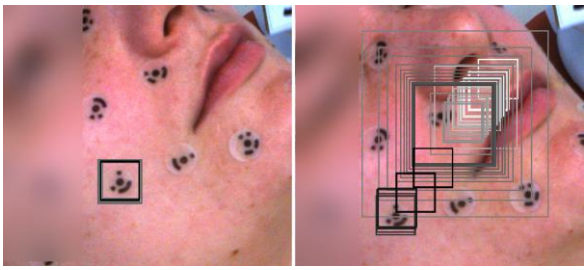

Figure 7. ROI for ellipse detection gets updated from start (black) to finish (white); left: static movement of observation marker (i.e., no effective ROI update), right: dynamic movement of marker (i.e., ROI shift and rescaled)

Detected observation marker positions are aggregated from all cameras, and solved for their positions using the bundle adjustment. The resulting 3D face net is shown in Figure 8. Due to improper detections as well as little overlapping areas, the face net is usually never complete which is solved by averaging adjustment results temporally (aggregating resolutions from identical frames due to higher publish-rate from the application than frame-rate from the actual cameras) and spatially (cluster m consecutive results to avoid fluctuation). The motion indicators for the real SHDF and DHSF compared against the SHSF scenarios are shown in Figure 9. The DHDF signal clearly reproduces the movement (background color indicates direction change), while some delay is present, which may be induced by the applied spatial smoothing of consecutive clusters.

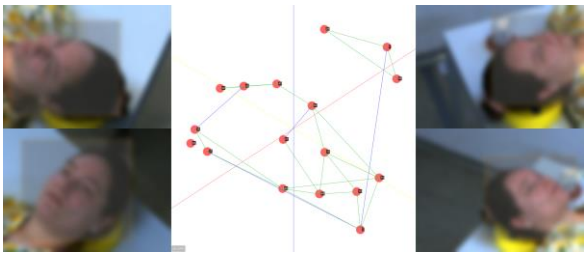

Figure 8. Estimated 3d observation points based on 2d image measurement and bundle adjustment

The SHDF plot shows remarkably well how the face deformation indicator reacts to small periods of face tension (green background indicates increased tension, red background decreased tension, white is hold). The head motion indicator slightly follows the tension curve, but the log-scaled indicator reacts much better to the actual signal. The latter can be expressed via some Signal-to-noise ratio that is defined as

$$SNR = \frac{|\bar{F} - \bar{H}|}{\sqrt{\sigma_F^2 + \sigma_H^2}} \tag{6}$$

from the within-run mean standard deviations of the F and H time series, computed across all epochs of the scenario being analyzed. It indicates the mean gap between F and H relative to the combined epoch-to-epoch scatter of both indicators – in this case 1.16 for SHSF, 1.64 for DHSF and 6.89 for SHDF. It follows that the face deformation produces a much cleaner separation between the F and H than head rotation does, which could help distinguishing dominant signals in mixed motion scenarios in the future.

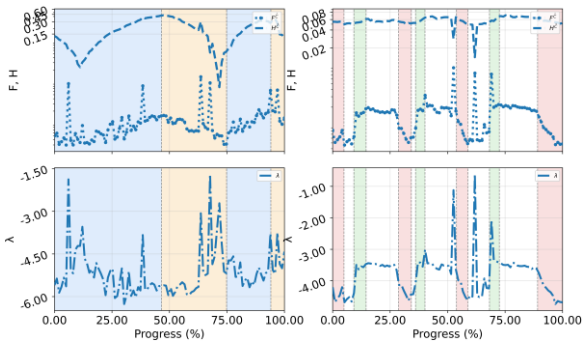

Figure 9. Motion indicators (real) for DHSF (left) and DHSF (right); top: H (dotted) and F (dashed) with log-scale; bottom: λ with linear scale

Table *3* allows to analyze the impact of individual points to the overall results. Both scenarios are listed in which the main contributing points are bold. They carry up to 25% of the overall deformation intensity alone which marks them as highly suitable for the individual detection. Some points were excluded in the analysis due to their low appearance rate throughout the experiment (e.g., point 72). Some were also detected as instable after the run (e.g., point 2). Remarkably is the systematic effect of points from camera 1 and 3 with standard deviations of several decimeters. They are damped down during the analysis, though their initial estimation hints a deeper problem with the camera setup that needs to be investigated further. In Figure 10 and Figure 11, the individual points are listed by their relative contribution to the deformation energy, and their belonging to a subset that defines together 68%, 95% and 99% of the full deformation energy and is therefore a representative reduction at that required level. In Figure 12 the face net is visualized exposing the clear beneficial left part of the face to contribute most of the deformation intensity – the side where camera 0 and 2 are placed. Also, one can see that the edge strain – the normalized edge length change – slightly higher is in the area of cheekbones.

ID	σ_P	c_j	$s_{i,SHFD}$		$s_{i,DHSF}$	
	$[mm]$		$[-]$	[%]	$[-]$	[%]
33	1.15	0;2	**3.854**	25	**2.944**	17
86	1.30	0;2	**3.761**	24	1.095	6
73	0.85	0;2	**1.419**	9	**1.632**	9
99	1.35	0;2	**1.362**	8	**1.467**	8
85	0.90	0;2	**1.182**	8	0.972	6
2	1.07	0;2	0.934	6	**5.211**	30
54	1.02	0;2	0.776	5	**1.489**	9
66	1.35	0;2	0.683	4	0.996	6
27	7.18	>	0.578	4	0.414	2
72	0.98	>	outlier	(-)	0.618	4
67	10.63	>	0.173	1	0.277	2
46	11.84	>	0.219	1	0.103	1
34	14.78	>	0.106	*1*	0.083	*1*
1	46.35		0.129	*1*	0.069	*1*
53	23.37	1;3	0.092	*1*	0.057	*1*
98	131.72	1;3	0.046	*1*	0.018	*1*
47	28.09	1;3	0.040	*1*	0.043	*1*
59	40.91	1;3	0.022	*1*	0.035	*1*
40	71.95	1;3	0.021	*1*	0.018	*1*
41	52.37	1;3	0.018	*1*	0.033	*1*

Table 3. Marker statistics position standard deviation, camera origin, absolute deformation intensity and relative contribution to it per scenario (SHDF/DHSF)

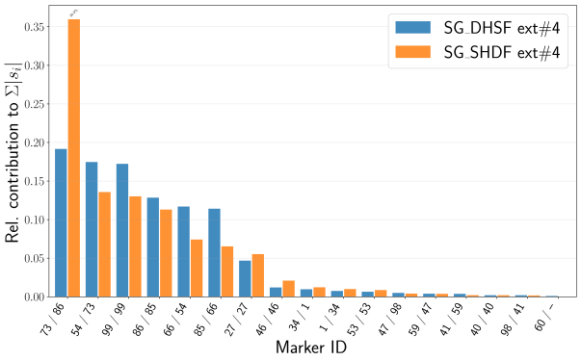

Figure 10. Relative contribution of each marker to the overall deformation intensity

By projecting the accumulated deformation intensity from X^w to X^I, the image regions affected by face deformation can be visualized and furthermore used as additional input heatmap layer for any head pose estimation using a machine learning model. For that, the heat at pixel $(\tilde{u}, \tilde{v})$ based on the vertex i 2d observation point (u_i, v_i) is

$$\mathcal{H}(\tilde{u}, \tilde{v}) = \sum_i s_i \cdot h(r, c, u_i, v_i, \sigma_F) \tag{7}$$

The function h is designed as a strongly monotonically decreasing function and depend additionally on a const value σ_F that controls the spatial extent of each points influence. A Normalized Deformation Indicator (NDI) is then defined as

$$\mathrm{N} = \frac{\mathcal{H}^{DHSF} - \mathcal{H}^{SHDF}}{\mathcal{H}^{DHSF} + \mathcal{H}^{SHDF}} \tag{8}$$

This finally allows to deduce a heatmap (Figure 13) for the camera for that specific scenario (a kind of calibration state) which enables a geometric and statistically reliable weighting scheme like ML models usually lack.

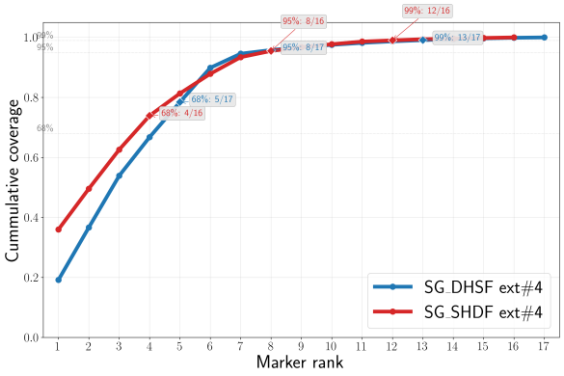

Figure 11. Aggregated coverage of the overall deformation energy per scenario

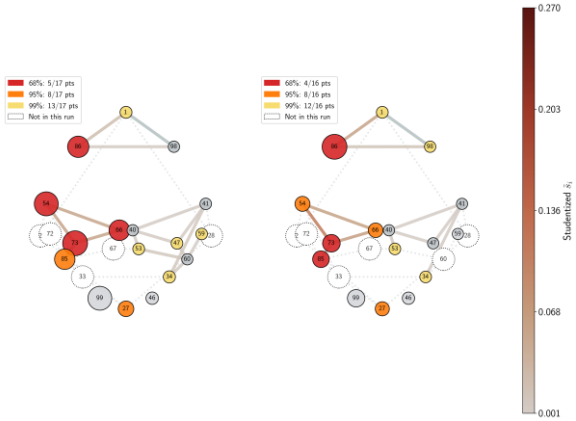

Figure 12. Face net visualization of marker contribution (size) to overall deformation with percentile-subset (color legend) and edge strain (color bar); left: SHDF; right: DHSF

5. Conclusion

Reaching SNR of 6.89 for face deformation is a promising sign, that the approach is fundamentally working. The temporal and spatial clustering is necessary though to account for data exchange limitation. The heatmap highlights inner-face deformation down to individual and point-subset contribution at statistical significance. The inconsistent detection propagates through the whole chain and causes the signal to weaken. Also, facial deformations and rigid-body motion may find a compromise which causes invisible artifacts in the adjustment resulting in bad separation ratios. Still a system was presented that correctly analyses each epoch deformation as head or face dominated using the motion indicators. At point level it identifies

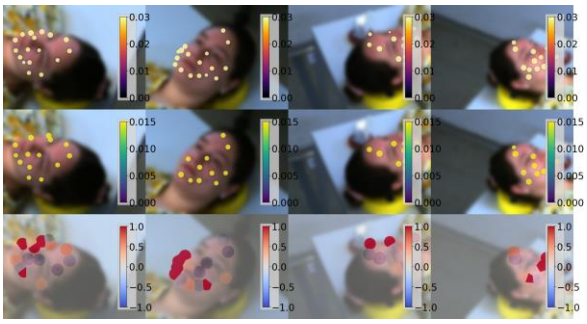

Figure 13. Heatmaps $\mathcal{H}$ from reprojected deformation indicator of DHSF (top row) and SHDF (middle row); Calculated NDI from $\mathcal{H}^{DHSF}$ and $\mathcal{H}^{SHDF}$ (bottom row)

specific markers contribution to the total deformation energy. At pixel level the proposed NDI projects deformation classification to the image producing a spatial heatmap that enables downstream ML model to account for face deformation without explicitly training them for it, by providing weighting possibilities. Mixed motion sequences are not yet covered, as the log-scale discrimination is only marginal. Also, a best reference motion is yet to be developed based on the first results of favorable point subsets. In the future, the real time application as well as a connection to head estimation models is to be investigated.

Acknowledgements

This work is co-funded by the European Union (EFRE/JTF Technology Funding Line 2021-2027; No. 100699425) and co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

Co-funded by the European Union

This project is co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

References

Bradski, G., 2000. The OpenCV Library. Dr Dobbs J. Softw.Tools.

Chun, S., Kim, B., Chang, H.J., Chang, J.Y., 2024. Bidirectional Regression for Monocular 6DoF Head Pose Estimation and Reference System Alignment. https://doi.org/10.48550/ARXIV.2407.14136

Feng, Y., Feng, H., Black, M.J., Bolkart, T., 2020. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. https://doi.org/10.48550/ARXIV.2012.04012

Hardner, M., Liebold, F., Wagner, F., Maas, H.-G., 2024. Investigations into the Geometric Calibration and Systematic Effects of a Micro-CT System. Sensors 24, 5139. https://doi.org/10.3390/s24165139

Hempel, T., Abdelrahman, A.A., Al-Hamadi, A., 2022. 6d Rotation Representation For Unconstrained Head Pose Estimation, in: 2022 IEEE International Conference on Image Processing (ICIP). Presented at the 2022 IEEE International Conference on Image Processing (ICIP), IEEE, Bordeaux,

France, pp. 2496–2500. https://doi.org/10.1109/ICIP46576.2022.9897219

Li, B., Zhang, D., Huang, C., Xian, Y., Li, M., Lee, D.-J., 2024. Location-guided Head Pose Estimation for Fisheye Image. https://doi.org/10.48550/ARXIV.2402.18320

Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2023. Close-range photogrammetry and 3d imaging, 4th ed. ed, De Gruyter Stem. De Gruyter, Boston.

Milkau, C., Preußel, S., Maas, H.-G., Schneider, D., 2026. Bildgestützte Lokalisierung von Zielkörpern in der Strahlentherapie. Beitr. Oldenburger 3D-Tage.

Niemeier, W., 2008. Ausgleichungsrechnung: Statistische Auswertemethoden, 2., überarbeitete und erweiterte Auflage. ed. Walter de Gruyter, Berlin.

Appendix

A. Position estimation (extended)

$$u_j = x_0 + \left(\Delta x'_{rad} + \Delta x'_{tan} + \Delta x'_{aff} + \ c_k \cdot p_x/p_z\right) \qquad (x)$$

$$v_j = y_0 - (\Delta y'_{rad} + \Delta y'_{tan} - \ c_k \cdot p_x/p_z) \qquad (x)$$

For each point $\boldsymbol{p}_i(X_i, Y_i, Z_i)$ the DLT is performed like:

$$\begin{pmatrix} X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0 & -u_iX_i & -u_iY_i & -u_iZ_i \\ 0 & 0 & 0 & 0 & X_i & Y_i & Z_i & 1 & -v_iX_i & -v_iY_i & -v_iZ_i \end{pmatrix} \begin{pmatrix} L_1 \\ \dots \\ L_{11} \end{pmatrix} = \begin{pmatrix} u_i \\ v_i \end{pmatrix}$$

i.e., a $2n \times 11$ linear system $\boldsymbol{AL} = \boldsymbol{l}$, solved via normal equations and LM regularization. From $\boldsymbol{L}$ the IOP are extracted