Motion Correction for Separate Mandibular and Cranial Movements in Cone Beam CT Reconstructions

Lukas Birklein[a)] and Stefan Niebler and Elmar Schömer

Institute of Computer Science, Johannes Gutenberg University, 55099 Mainz, Germany

Robert Brylka and Ulrich Schwanecke

Computer Vision & Mixed Reality Group, RheinMain University of Applied Sciences,

65195 Wiesbaden, Germany

Ralf Schulze

Devision of Oral Diagnostic Sciences, Dept. of Oral Surgery and Stomatology, University of Bern, 3011 Bern, Switzerland

Version typeset March 13, 2023

[a)]Author to whom correspondence should be addressed. email: lukas.birklein@uni-mainz.de

Abstract

Background: Patient motions are a repeatedly reported phenomenon in oral and maxillofacial cone beam CT scans, leading to reconstructions of limited usability. In certain cases, independent movements of the mandible induce unpredictable motion patterns. Previous motion correction methods are not able to handle such complex cases of patient movements.

Purpose: Our goal was to design a combined motion estimation and motion correction approach for separate cranial and mandibular motions, solely based on the 2D projection images from a single scan.

Methods: Our iterative three-step motion correction algorithm models the two articulated motions as independent rigid motions. First of all, we segment cranium and mandible in the projection images using a deep neural network. Next, we compute a 3D reconstruction with the poses of the object's trajectories fixed. Third, we improve all poses by minimizing the projection error while keeping the reconstruction fixed. Step two and three are repeated alternately.

Results: We find that our marker-free approach delivers reconstructions of up to 85% higher quality, with respect to the projection error, and can improve on already existing techniques, which model only a single rigid motion. We show results of both synthetic and real data created in different scenarios. The reconstruction of motion parameters in a real environment was evaluated on acquisitions of a skull mounted on a hexapod, creating

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.16347

a realistic, easily reproducible motion profile.

Conclusions: The proposed algorithm consistently enhances the visual quality of motion impaired CBCT scans, thus eliminating the need for a re-scan in certain cases, considerably lowering radiation dosage for the patient. It can flexibly be used with differently sized regions of interest and is even applicable to local tomography.

I. Introduction

Cone Beam Computed Tomography (CBCT) is an established three-dimensional (3D) radiographic imaging technique. Introduced in dental imaging in the late 1990's[1,2], the technique is now also widely used in other medical disciplines[3,4,5,6]. Owing to the implementation of flat-panel detectors as image receptors and to its technical design, an inherent shortcoming of CBCTs for maxillofacial application lies in long rotation times of 10 to 40 s, during which several hundreds of projection radiographs used for 3D reconstruction are acquired[7]. The backprojection process for 3D reconstruction relies on a priori knowledge of the imaging geometry for each projection radiograph up to voxel size accuracy[8]. If a patient moves during the acquisition time, errors in the reconstruction chain necessarily occur. Due to the long scan times, patient motion is an issue in CBCT[7,9]. The most common reconstruction technique relies on the Feldkamp algorithm[10] which essentially is a 3D adaptation of the classical fan-beam filtered backprojection[11]. However, iterative reconstruction techniques are known to be more flexible and can incorporate statistics, physical models and a priori knowledge[12]. This can be used to create more exact results, which is especially important when using a projection based motion correction method. In this context it is noteworthy that CBCTs do not produce standardized gray values in the sense of e.g. Hounsfield units (HU), i.e. they cannot be compared between machines. The scale of the reconstructed gray values even differs within volumes and also between different exposure settings in one machine.[13,14]

Periodic motions, such as those caused by breathing or heart-beat, has been a topic of lively research over the last 20 years[15,16,17,18,19]. Typically, such approaches use models, for instance surrogate motion models, to estimate motion that is otherwise not directly accessible[19,20], instead of real yet inaccessible organ motion, e.g. lung deformation estimated from the thoracoabdominal surface[19] or by measuring air flow changes by spirometry[21,22]. Inherent errors from such surrogates lie in potential miscorrelation with the internal organ motion they shall represent[23]. While such techniques may be capable to address periodic motion, patient movement characteristics in maxillofacial CBCT can not easily be described by such models, since they include multiplanar movements, head rotation or swallowing[24]. Cases are also observed in which only the mandible is moved against the remaining (resting) skull[24]. Techniques, which are applicable in scenarios more similar to ours, include methods of autofocus[25,26,27,28], consistency conditions[29,30,31] and learning based approaches[32,33,34]. However, methods using consistency conditions are currently not able to correct separate cranial and mandibular motions, which is the goal of our work. Approaches based on an autofocus metric are typically able to compensate non-rigid transformations, but usually incorporate temporal regularization of some kind to reduce the motion parameter space, hindering their application in cases of inconsistent or sudden motions of the patient. Motion correction methods based on deep learning are another lively field of research and will play an important role in the future. Berger et al.[35] presented an approach which is able to correct multiple rigid motions in one field of view (FOV). This method relies on a prior motion-free reconstruction of the same region, performing a 3D/3D registration, followed by a bone-wise 2D/3D registration. Another noteworthy work in that field was presented by Flach et al.[36], performing deformable 3D/2D registration by applying a regularized deformation field to the reconstructed volume. In their method, the authors also make use of an artifact-free reconstruction by splitting the scan into a "prior" and "intervention" phase, of which only the second one was motion impaired. Unfortunately, such an artifact-free reconstruction is usually not available in clinical applications.

Based on the work of Niebler[37], in this paper we introduce a marker-free, projection-based iterative framework for correcting movements of the facial skull or of the mandible, which may move relative to the cranium in certain cases. Without any priori knowledge, we model the motion as two separate rigid motions of these two components.

The paper is structured as follows: After an explanation of the CBCT reconstruction model the method is described, followed by the segmentation process of the mandible from 2D projection images. The next section highlights the implementation of the algorithm. Results from synthetic experiments and real world data are then described.

Subsequently, the method is discussed and conclusions are drawn. Lastly future work directions are presented.

II. Materials and Methods

A typical CBCT setup consists of an X-ray source S and a panel detector D rotating around an object $\tilde{x}$, generating $n \in \mathbb{N}$ individual X-ray images $b^{(i)}, i \in \{1, ..., n\}$, each of dimension $w \times h$. This results in a total number of $n \times w \times h$ pixels $b_c^{(i)}$, each of which, in simplified terms, can be associated with one X-ray $r_c^{(i)}$ running from S through $\tilde{x}$ into $b_c^{(i)}$ in a straight line along the path $\gamma_c^{(i)}$. On this path, the intensity of $r_c^{(i)}$ is weakened by the attenuation coefficient of the matter which it is currently passing through.

Since in practice it is only possible to reconstruct a discrete approximation of $\tilde{x}$, we define a grid of $m_x \times m_y \times m_z$ voxels v each with a given pitch (size) over the region of interest (ROI) Ω, where $m_x, m_y, m_z \in \mathbb{N}$ denote the number of voxels in x, y, z-dimension, respectively. Each voxel is then assigned the attenuation value of $\tilde{x}$ at its spatial position. The following notation will use the continuous and discrete versions of $\tilde{x}$ interchangeably.

Using this discrete version of $\tilde{x}$, the reconstruction problem boils down to solving a linear system of equations of the form

$$A\tilde{x} = b. \tag{1}$$

The system matrix A has a block diagonal shape, as projection images are independent from one another, and can be fully determined based on the specifications of the CBCT-machine. Since A is usually not a square matrix, Eq. (1) cannot be solved directly. We can use the corresponding (damped) least squares problem

$$\arg\min_{\tilde{x}} \|A\tilde{x} - b\|_2^2 + \lambda R(\tilde{x}), \tag{2}$$

with a suitable regularization term R and parameter $\lambda \in \mathbb{R}$, to minimize the residual error of the projection by using the conjugate gradient method for least squares ($CGLS$) on the normal equation.

II.A. Method Overview

By assuming a static object $\tilde{x}$, standard reconstruction algorithms cannot account for patient motion, resulting in highly artifact-laden reconstructions in certain cases[8]. In this paper, we propose a method to mitigate these effects, solely based on the 2D acquisition images and without any need for further prior knowledge.

Contrary to the previous work of Niebler[37], occurring patient movements are modeled as two separate rigid motions, one for the cranium and an independent mandibular motion. Since the focus lies primarily on reconstructing bone structure and teeth, modeling the patient's movement as two separate rigid motions is well justified.

For that we need to separate $\tilde{x}$ into two parts: *mandible* and *cranium*. We define $\Omega_C \subset \Omega$ as the region of the cranium and $\Omega_M \subset \Omega$ as the mandibular region, so that $\Omega_M \cup \Omega_C = \Omega$, $\Omega_M \cap \Omega_C = \emptyset$. This splits $\tilde{x}$ into

$$\tilde{x}_{C/M}(\mathbf{x}) = \begin{cases} \tilde{x}(\mathbf{x}), & \mathbf{x} \in \Omega_{C/M} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

s.t. $\tilde{x}(\mathbf{x}) = \tilde{x}_C(\mathbf{x}) + \tilde{x}_M(\mathbf{x})\ \forall \mathbf{x} \in \Omega$. For each of the two regions we define *motion fields* $p_C = \{p_C^{(i)}\}_{i=1,\dots,n}$ and $p_M = \{p_M^{(i)}\}_{i=1,\dots,n}$ of the shape $p_{C;M}^{(i)} = (\phi^{(i)}, \theta^{(i)}, \psi^{(i)}, t_x^{(i)}, t_y^{(i)}, t_z^{(i)})$, associated with cranium and mandible, respectively, each describing the 6 degrees of freedom (3 rotational and 3 translational) of a rigid motion at acquisition time i.

Our proposed algorithm aims to solve the motion corrected reconstruction problem for both $\tilde{x}_C$ and $\tilde{x}_M$ using a three-step approach, iteratively splitting Ω into Ω_C and Ω_M (Section II.B.3.), reconstructing the approximate volume $\tilde{x} = \tilde{x}_C + \tilde{x}_M$ using a motion aware reconstruction method with the current motion parameters (Section II.B.1.) and finally updating the motion fields using the latest (imperfect) reconstruction (Section II.B.2.).

II.B. Method Details

II.B.1. Motion-Aware Volume Reconstruction

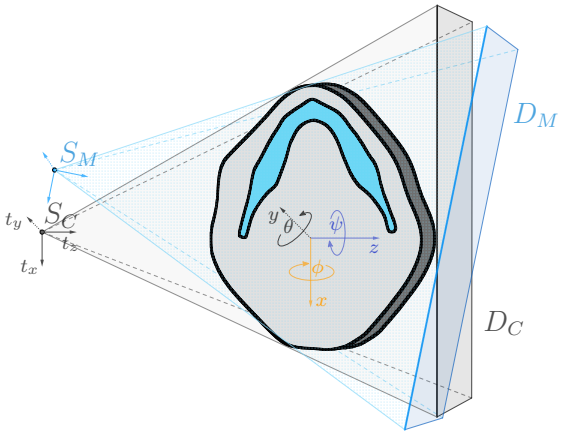


Figure 1: Instead of using one projection, our approach uses two virtual source-detector pairs per X-ray image $b^{(i)}$. (S_C, D_C) only scans $\tilde{x}_C$ (grey area) while (S_M, D_M) scans $\tilde{x}_M$ (blue area). The resulting intensities are then added to generate the final projection.

As we model the movement of the patient by two separate rigid motions, we can express $\tilde{x}$ at frame i as

$$\begin{aligned}\tilde{x}(\mathbf{x}, i) &= \tilde{x}_C(\mathbf{x}, i) + \tilde{x}_M(\mathbf{x}, i) \\ &= T(p_C^{(i)})\tilde{x}_C(\mathbf{x}, 0) + T(p_M^{(i)})\tilde{x}_M(\mathbf{x}, 0)\end{aligned} \tag{4}$$

where $T(p_C^{(i)})$ and $T(p_M^{(i)})$ are some linear maps describing the two separate rotations and translations of $\tilde{x}$, parameterized by p_C and p_M. In our motion-aware reconstruction the patient's motion is incorporated into the system matrix $\mathbf{A}$.

The key observation in finding the motion-aware projection $\mathbf{A}(p_C, p_M)$ comes from Eq. (4), as splitting $\tilde{x}$ into two disjoint regions, each acted upon by only a single rigid motion, allows us to also write $\mathbf{A}(p_C, p_M)$ as two matrices

and rewrite Eq. (1) into

$$\begin{aligned} & A \cdot (T(p_C)\tilde{x}_C + T(p_M)\tilde{x}_M) = b \\ \iff \quad & A(p_C)\tilde{x}_C + A(p_M)\tilde{x}_M = b \\ \iff \quad & (A(p_C) \cdot \mathbb{1}_{\Omega_C} + A(p_M) \cdot \mathbb{1}_{\Omega_M})\,\tilde{x} = b \end{aligned} \tag{5}$$

where $A(\cdot) := AT(\cdot)$ is the motion-aware projection matrix depending on only one of the two motion fields and the indicator $\mathbb{1}$ can be expressed as a matrix multiplication. Defining

$$\mathbf{A}_{\Omega_C,\Omega_M}(p_C, p_M) := A(p_C) \cdot \mathbb{1}_{\Omega_C} + A(p_M) \cdot \mathbb{1}_{\Omega_M} \tag{6}$$

leads to the linear system in the form of Eq. (1), assuming Ω_C, Ω_M, p_C and p_M are known.

II.B.2. Motion Estimation

In this section, we leverage a given reconstruction $\tilde{x}_{approx}$ to find a better set of parameters p_C and p_M by minimizing the residual error between b and the virtual projections of $\tilde{x}_{approx}$

$$\arg\min_{p_C,p_M} E\left(\mathbf{A}_{\Omega_C,\Omega_M}(p_C, p_M) \cdot \tilde{x}_{approx}, b\right). \tag{7}$$

In our implementation we chose E as the $L2$-norm of the residual. There are numerous different suitable metrics for this task suggested by other 2D/3D registration works, e.g. the gradient orientation similarity metric (GO) used by Ouadah[38], or gradient correlation (GC) and normalized gradient information (NGI) as used by Berger et al.[35], however in our experiments we could achieve the most consistent results using $L2$.

As the motion parameters corresponding to two different projection images $b^{(i)}$ and $b^{(j)}$ are independent of each other, we can separate Eq. (7) into n independent optimization problems, one for each frame, for each of which we apply a gradient based solver (for details see Appendix A). We use the nonlinear conjugate gradient method with Polak-Ribière weights[39] for the optimization process of p_C and p_M. The required line search is performed as an exact line search for quadratic functionals, as suggested by van Leeuwen et al.[40] We discard gradients outside of the ROI in general.

II.B.3. Mandible Segmentation

Unfortunately, the unambiguous segmentation into cranium (Ω_C) and mandible (Ω_M) on an estimated reconstruction $\tilde{x}_{approx}$ of earlier iterations, especially in the first iteration, is not possible. This is because $\tilde{x}_{approx}$ still contains artifacts, such as blur or double contours. Even state-of-the-art neural networks such as Anatomy Net[41] are therefore unable to correctly identify the mandible. Instead we perform 2D segmentations of the mandible on the artifact-free projection images $b^{(i)}$ yielding labeled data $l_b^{(i)}$, by using the segmentation network PointRend[42] (see following paragraph **2D segmentation**). From these the 3D label Ω_M is computed. We use a principal component model depending on only a small number of parameters, representing triangle meshes of various shapes of different mandibles. This model's projection is registered with the 2D labels, creating the volumetric label of the mandible within $\tilde{x}_{approx}$ (see paragraph **Creating 3D labels**).

2D segmentation: PointRend is an improvement of Mask-RCNN[43], a network that for a given class (in our case the mandible) provides the localization in the form of a bounding box and a coarse mask. The PointRend variant uses the coarse-resolution pixel mask and increases it to the original input resolution to enhance the segmented objects' edges. For the training set, we used 43 CT datasets representing heads from a previous work[44]. To generate the mask for the mandible, we first extracted the bone structure from the CT by applying the Marching Cubes algorithm, yielding a triangle mesh of the whole skull. We subsequently extracted the mandible from the bony structure by manually separating the meshes at the condyles and teeth area. As a result, we have 43 volumes for our CT set showing the masking of the mandible area. For each volume, we synthetically created a CBCT scan of 516 projection images for the training process. The trained neural net is robust against physical effects like noise, beam hardening and scatter, so we did not need to consider these during data generation.

Creating 3D labels: We compute Ω_M using a principal component analysis (PCA) model. This model is built from the same $k = 43$ triangle meshes of the mandible used for training the 2D segmentation network. We use the $\hat{k}$ largest principal components to formulate a parameterized triangle mesh model $\mathbf{L}_{\hat{k}}(\boldsymbol{\lambda})$. This number is set to $\hat{k} = 5$, since these 5 eigenvalues are sufficient to achieve a coverage of more than half of the underlying data. Finally, we extend $\mathbf{L}$ by the parameters $\mathbf{t}, \mathbf{r}, \mathbf{s} \in \mathbb{R}^3$, the translation, rotation and scale of the resulting mesh, respectively, leading to the final model being parameterized by $\pi = \{\boldsymbol{\lambda}, \mathbf{t}, \mathbf{r}, \mathbf{s}\}$, a total number of 14 parameters. Since the triangle meshes used for $\mathbf{L}$ are static, this model can be computed offline.

We register the forward projection of $\mathbf{L}$ with the labels of the projection data $l_b = \{l_b^{(i)}\}, i \in \{1, ..., n\}$ using the following minimization problem

$$\arg\min_{\pi} E\left(A_{max}(p_M) \cdot \mathbf{L}(\pi), l_b\right) \tag{8}$$

where A_{max} denotes a maximum projection dependent on the parameters p_M. We solve Eq. (8) using the Nelder-Mead algorithm[45]. In our implementation we chose E to be the $L2$-norm of the residual, but other tested metrics such as *Intersection over Union (IoU)* provide very similar results. Note that Eq. (8) is still dependent on p_M, suggesting more exact fits with a good approximation of the jaw's motion parameters in later iterations.

With the optimal set of parameters π computed, Ω_M is defined as the set of voxels inside the triangle mesh $\mathbf{L}(\pi)$; Ω_C is then given implicitly as $\Omega_C = \Omega \setminus \Omega_M$. This approach created very close fits in all of our experiments (Sec. IV.), even for local tomography scenarios where the neural net is not always able to correctly identify the whole mandible (see Fig. 2b). It also has the advantage of being independent of $\tilde{x}_{approx}$, restricting further propagation of occurring reconstruction errors.

We artificially enlarge the computed 3D label of the mandible to account for motions of soft tissue, especially relevant in the chin area, and for possible inaccuracies of $\mathbf{L}$. To avoid growing into the upper tooth row, we do not dilate in the positive y-direction, keeping the upper and lower teeth separated. Using a hard partition into cranium and mandible in Eq. (6) leads to very sharp edges between the two regions within the final result, especially since parts of the soft tissue do not undergo a rigid transformation. Therefore we apply a Gaussian blur to the 3D label itself, smoothing out those edges. Mathematically speaking, we convolve the $\mathbb{1}_{\Omega_M}$ function from Eq. (6) with a three-dimensional Gaussian kernel G_σ of standard deviation σ. This extends $\mathbf{A}$ to

$$\begin{aligned}\mathbf{A}^*_{\Omega_C,\Omega_M}(p_C, p_M) := A(p_C) \cdot (\mathbb{1} - G_\sigma * \mathbb{1}_{\Omega_M}) \\ + A(p_M) \cdot (G_\sigma * \mathbb{1}_{\Omega_M}),\end{aligned} \tag{9}$$

where $\mathbb{1}$ is the identity. In our implementation we chose $\sigma = (5/vp_x, 5/vp_y, 5/vp_z)^T$, where vp is the voxel pitch. This ensures consistent results with differently sized ROIs and resolutions (e.g. in local tomography).

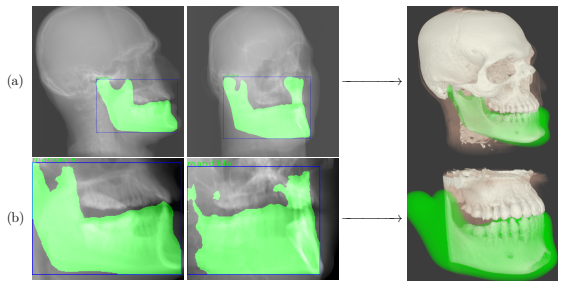


Figure 2: The segmentation process. In each row, the two pictures on the left show the labeling $l_b^{(i)}$ done by our neural net. On the right side, the computed 3D label $\mathbf{L}(\pi)$ can be seen. The top row (a) shows scans of the whole head whereas the bottom row (b) depicts a typical local tomography scenario. Note that in both cases the 3D labels are of high quality.

II.C. Putting everything together

Algorithm 1 combines the described three separate phases. We first create the labels of the mandible in the 2D projections, since they are independent of the motion parameters. Ω_M is then computed in an outer loop as described in Section II.B.3. using the current motion parameters. In earlier iterations of this outer loop, we use a downsampled version b_k of b to speed up calculations. In later iterations we gradually increase the resolution by adjusting the downsampling factor. In an inner loop the reconstruction of $\tilde{x}$ and the search for p_C and p_M are performed alternately according to the sections II.B.1. and II.B.2. The resolution of $\tilde{x}$ also increases with the number of iterations already performed (similar to Sun et al.[46]), since the low resolution versions b_k of the images can only support a limited reconstruction quality. This procedure allows us to decrease the reconstruction time drastically while not sacrificing overall reconstruction quality.

We chose not to update Ω_M in every iteration of the process. Since the updates of p_M are rather small in each iteration, they induce very little change to Ω_M. We find that delaying this calculation has very little impact on quality, while further reducing computational costs.

III. Implementation Details

Rigid Motion Description The position of the source $S^{(i)}$ at acquisition time i is given by applying the rotation

$$\mathbf{R}^{(i)} := \left(R_x(\phi^{(i)})R_y(\theta^{(i)})R_z(\psi^{(i)})\right) \cdot R_y(\delta^{(i)})$$

and the translation

$$\mathbf{t}^{(i)} := \left(t_x^{(i)}, t_y^{(i)}, t_z^{(i)}\right)^T$$

to the initial source position $S^{(0)}$ defined by the specifications of the CBCT device, resulting in

$$S^{(i)} = \mathbf{R}^{(i)} \cdot S^{(0)} + \mathbf{R}^{(i)} \cdot \mathbf{t}^{(i)},$$

Algorithm 1: The complete reconstruction algorithm. We set $\alpha := p_C,\ \beta := p_M$ for notation purposes.

Input : raw projections b, regularization parameter λ, PCA-model $\mathbf{L}$, number of max iterations N_{inner}, N_{outer}, stopping criteria

Output: motion corrected reconstruction x

$x_0, \alpha_0, \beta_0 \leftarrow 0$

/* use neural net to find 2D segmentations */

$l_b \leftarrow$ `IdentifyJaw2d(`b`)`

for $k = 1, \dots, N_{outer}$ **do**

  /* downsample the resolution of each image */

  $b_k \leftarrow \downarrow_k b$

  /* compute Ω_M and Ω_C */

  $\pi_k \leftarrow \arg\min_{\pi} E\left(A_{max}(\beta_{k-1}) \cdot \mathbf{L}(\pi), l_b\right)$

  $\Omega_{Mk} \leftarrow \{v \mid v \text{ is inside } \mathbf{L}(\pi_k)\}$

  $\Omega_{Ck} \leftarrow \Omega \setminus \Omega_{Mk}$

  $\hat{\alpha}_0, \hat{\beta}_0 \leftarrow \alpha_{k-1}, \beta_{k-1}$

  for $t = 1, \dots, N_{inner}$ **do**

    /* reconstruct x with current motion parameters using CGLS */

    $\hat{x}_t \leftarrow \arg\min_{x}(\|\mathbf{A}^*_{\Omega_{Ck},\Omega_{Mk}}(\hat{\alpha}_{t-1}, \hat{\beta}_{t-1})x - b_k\|_2^2 + \lambda R(x))$

    if *stopping criteria met* **then**

      break

    end

    /* find optimal motion parameters for each image independently */

    for $i = 1, \dots, n$ **do**

      $\hat{\alpha}_t^{(i)}, \hat{\beta}_t^{(i)} \leftarrow \arg\min_{\alpha,\beta} \|\mathbf{A}^{*(i)}_{\Omega_{Ck},\Omega_{Mk}}(\alpha, \beta)\hat{x}_t - b_k^{(i)}\|_2^2$

    end

    /* concatenate individual poses to motion field */

    $\hat{\alpha}_t, \hat{\beta}_t \leftarrow (\hat{\alpha}_t^{(1)}, \dots, \hat{\alpha}_t^{(n)}), (\hat{\beta}_t^{(1)}, \dots, \hat{\beta}_t^{(n)})$

  end

  /* update reconstruction and parameters in outer loop */

  $x_k \leftarrow \hat{x}_t$

  $\alpha_k, \beta_k \leftarrow \hat{\alpha}_t, \hat{\beta}_t$

end

return $x_{N_{outer}}$

see Fig. 1 for details. As the relative positions of source and detector are fixed, the position and rotation of the detector can be computed in exactly the same way. The angle $\delta^{(i)}$, the gantry rotation, is implicitly given by the device geometry. Often, the default positions are equidistantly placed on a circle around the scanned object.

The reason for using the local coordinates of the source-detector pair for $\mathbf{t}$, i.e. translating by $\mathbf{R}\cdot\mathbf{t}$ instead of only $\mathbf{t}$, is that translations perpendicular to the detector plane cannot be reconstructed reliably. The narrow field of view (FOV) of CBCT machines (in our test cases about 18° horizontal and 14° vertical cone angles) makes the recovery of this dimension nearly impossible. Using the above notation ensures that only t_z is affected by this uncertainty. In our implementation we drop this coordinate for the optimization process.

Stopping Criteria We stop the motion-aware *CGLS* reconstruction from Sec. II.B.1. after 30 iterations since we found a higher number to only increase the noise within the reconstruction as well as negatively affecting the runtime, but less iterations may yield a blurry 3D volume. We stop Newton's method for Eq. (7) if there is no significant improvement of the residual anymore, i.e. $1 - \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} < \varepsilon^2$, which takes about three to four iterations on average. In Sec. II.B.3. we stop the computation of Eq. (8) if $2 \cdot \frac{\|r_{best} - r_{worst}\|_2^2}{\|r_{best} + r_{worst}\|_2^2} < 10^{-5}$ or after a fixed number of 800 iterations. For all our experiments we set the maximum number of outer iterations in Alg. 1, N_{outer}, to three and inner iterations, N_{inner}, to five. As an additional stopping criterion we test whether $1 - \frac{\|\mathbf{A}\hat{x}_t - b_k\|_2}{\|\mathbf{A}\hat{x}_{t-1} - b_k\|_2} < \varepsilon$ and if so, we continue the outer loop with $k+1$. We set $\varepsilon = 0.025$.

IV. Results

We conducted several experiments with both synthetic and real data to evaluate our motion correction algorithm. The synthetic data stems from a volumetric density model[44], on which different magnitudes of motion were tested. We are aware that this data may overlap with data used to train the neural net and the PCA-model from Sec. II.B.3. Therefore we also verify the labeling of the 2D projection images and creation of the 3D labels on data sets of real patients which were not included in this process.

IV.A. Synthetic Data

In this section we evaluate our methods using synthetically generated projection data. To assess the quality of our motion correction algorithm in various scenarios, we created projections with three differently sized ROIs, one of the whole head, one representing the biggest possible volume of the Accuitomo 170 CBCT device (170 mm diameter, 120 mm height), and a local tomography setup with a very limited ROI (80 mm diameter, 65 mm height), see Fig. 3. For each scenario we simulated the two separate motions of cranium and mandible using different motion profiles.

Motion of the cranium We used three distinct motion profiles for the cranium. For two of them, a random walk was performed for each motion parameter, i.e. degree of freedom, one of low (up to 5° rotation, 2 mm translation) and one of high (up to 15° rotation, 6 mm translation) amplitudes. These motion profiles were taken from the work of Niebler et al.[37] and provide an unpredictable, challenging environment with multiplanar patient motions. In the third profile the patient performed a single sudden movement after 200 frames (3° rotation on each axis, 2 mm translation on each axis) and held this position for the rest of the scan. Spin-Neto et al.[47] put an approximate threshold of 3 mm movements that significantly increases the likelihood of images being not interpretable, but acknowledge that

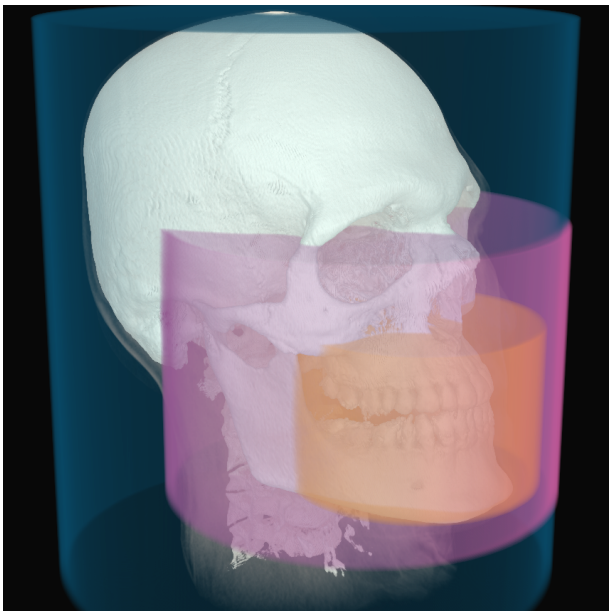


Figure 3: The cylindrical ROIs of the three tested setups are shown in different colors. The biggest is a scan of the whole head (blue), the purple one resembles the biggest ROI of the Accuitomo 170 device and our local tomography scenario can be seen in orange.

the majority of present patient movements are $\leq 2\,\text{mm}$. Those smaller movements are covered by our first profile while our second and third ones cover motions, that potentially render reconstructions unusable.

Motion of the mandible Using 50 volumes of the same subject but with different positions of the mandible, we simulate a patient opening the mouth during an acquisition. For the low and high amplitude motion profiles the mouth is being opened steadily by a downward motion during the whole acquisition (5° and 3 mm in total), in the scenario with the sudden movement the mandible is moved once between frames 258 and 259 (also 5° and 3 mm). All other motions (and thus degrees of freedom) are simulated implicitly since the mandible is additionally moved with the cranium. In the motion reconstruction process we treat p_C and p_M as two independent parameter sets.

With the given motion parameters and CBCT device geometry we generate the 2D acquisition images from the synthetically generated volumes (550×625×550 voxels) by virtually applying rotation and translation in the forward projection. We make sure to always scan the whole head (i.e. matter outside of the reconstruction radius) to achieve high authenticity of our synthetic data. This is especially important in the local tomography scenario.

	Whole head	Accuitomo 170	Local Tomography
Low amplitude	0.056 / 0.011 0.81 / 0.93	0.051 / 0.015 0.74 / 0.90	0.050 / 0.026 0.66 / 0.76
High amplitude	0.093 / 0.014 0.69 / 0.92	0.066 / 0.016 0.63 / 0.87	0.072 / 0.029 (∗) 0.60 / 0.72
Sudden motion	0.057 / 0.02 0.78 / 0.92	0.058 / 0.02 0.71 / 0.89	0.071 / 0.029 0.63 / 0.78

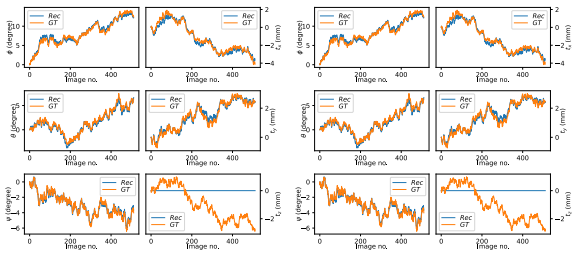
Table 1: Comparisons of the relative projection errors $\|\mathbf{A}\tilde{x}-b\|_2/\|b\|_2$ (upper rows) and SSIM[48] values (bottom rows) of the uncorrected reconstruction (first value) and the output of our algorithm (second value). SSIM values stem from comparing the reconstructions to the ground truth, restricted to their respective ROIs, with a window of 7 voxels. Our procedure was able to drastically improve the projection error (up to 85%) as well as increase the similarity between reconstruction and ground truth in every case. Due to the high motion amplitude and narrow ROI however, the result of the marked scenario (∗) would still be unusable in a clinical application.

Figure 4: Using the proposed method we were able to increase the visual quality of the reconstruction of our synthetic model drastically. In the top image ($[-1000\,\mathrm{HU}, 1600\,\mathrm{HU}]$), one can see a slice of the initial, uncorrected reconstruction while the bottom image ($[-1000\,\mathrm{HU}, 2200\,\mathrm{HU}]$) shows the same slice of our motion correction algorithm output. This motion parameters of this scenario are shown in Fig. 5.

Whole head In this paragraph we evaluate the quality of our proposed algorithm in a scenario, where the limitations and effects of local tomography setups, in particular the truncation of matter, do not occur. Tab. 1 shows a quantitative analysis of similarities to the ground truth and residual errors of this and the following scenarios. An overview over volume and image dimensions can be found in Tab. 3.

Accuitomo 170 Fig. 5 shows the comparison of both the computed cranial (5a) and mandibular (5b) motions to their respective ground truths in the case of high motion amplitudes. Typically, the computed motion parameters and the ground truth would not align. This happens because the pose of the reconstruction itself can differ from the ground truth, for example the whole reconstruction could be shifted upwards on the y-axis. To achieve the maximal comparability of the motion parameters, we rotated and shifted the coordinate system so that it aligns with the first projection image. Even in this case of severe motion, our method could identify the present motion with high precision and increase the overall reconstruction quality (Fig. 4, Tab. 1).

The figure also suggests that finding the cranium's motion works better than reconstructing movements of the mandible. Since the influence of p_C on the cost function E from Eq. (7) is usually greater than the influence of p_M (i.e. $E(p_C+\varepsilon, p_M) > E(p_C, p_M+\varepsilon)$ for optimal p_C, p_M), this phenomenon would be expected. As already mentioned, the translation parameter t_z cannot be computed reliably and is therefore fixed at 0. The relative projection error of this scenario can always be kept below 2% and we achieve an SSIM value of at least 0.86.



(a) Cranium (b) Mandible

Figure 5: Comparison of the reconstructed motion parameters (*Rec*, blue curve) with the ground truth (*GT*, orange curve) for the Accuitomo 170 scenario of both the cranium (a) and mandible (b). The projection data was created using motions from the data set of high motion amplitudes.

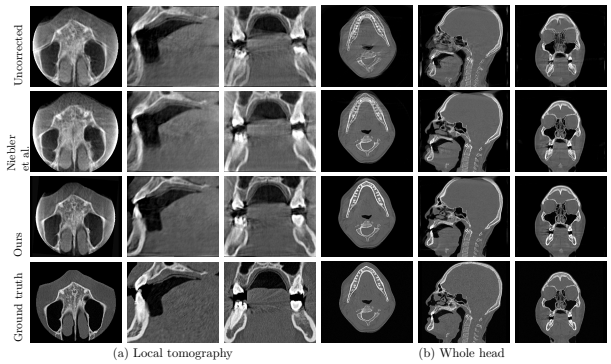


Figure 6: Qualitative comparison of the local tomography setup using low intensity motion amplitudes (a) and the whole head scenario using the sudden motion (b). Even for small movements, the uncorrected reconstruction bears artefacts like double contours, concentric ring patterns and general fuzziness (as already described by Schulze et al.[8]). Compared to the previous approach[37], our method can improve the reconstruction even further. In the local tomography scenario, the improvements are not limited to the region of the lower jaw only, but the quality (visual as well as in least squares sense) of the remaining image is also increased by using our mathematical model of two separate rigid motions. In scenario (b) the definition of the lower jaw is clearly enhanced compared to the previous approach. All shown Hounsfield units are within $[-1000\,\mathrm{HU}, 2200\,\mathrm{HU}]$.

Local Tomography The local tomography problem is an especially challenging one for our algorithm. This type of scan normally only allows reliable reconstructions inside a very limited cylindrical ROI (see Fig. 3). Since the forward projection of our algorithm has to mimic the X-ray projections of the CBCT device (to correctly find the solution of Eq. (7)), we also need to project matter outside of the reconstruction radius. Therefore we are forced to reconstruct data outside of the ROI, which of course automatically induces some error due to incorrect attenuation values at those voxel positions. Accurately identifying the mandible within these truncated 2D projection images poses another challenge for our algorithm. However, even in this setup it was still possible to create fitting 3D labels for the mandible. In Fig. 6 we compare the output of the proposed method with the uncorrected reconstruction, the previous work[37], and the ground truth in our local tomography setup (6a) and for the whole head (6b).

PCA-model L Fig. 7 shows the voxels inside the triangle mesh resulting from Eq. (8), i.e. $\mathbf{L}(\pi)$, for three examples, before we artificially enlarge this area and soften its edges to create the final label Ω_M. The three depicted volumes were neither involved in the training data for our neural net, nor are they included in the data for the PCA-model $\mathbf{L}$

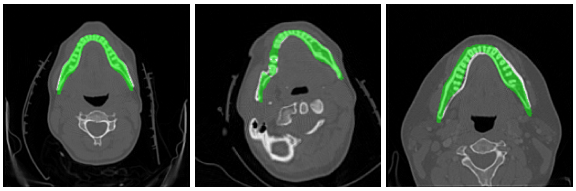


Figure 7: Three examples of our PCA-model with scans of real patients. The model provides a good fit in all cases, albeit a bit offset in the last image. Since the labels are artificially enlarged afterwards, small misalignments of only a few voxels usually do not matter. The second image shows a patient with a tilted head, in this case the model finds the correct rotation parameters and can achieve a close fit, too.

	Whole head	Accuitomo 170	Local Tomography
DSC	0.95	0.95	0.84

Table 2: The dice-coefficient *DSC* between ground truth and the segmentation in the 2D projection data provided by our neural network. The projections stem from the data sets of the high motion amplitude category.

(see Sec. II.B.3.). In Tab. 2 we compare the output of the neural net with the ground truth of the labeled projection data. The network provides segmentations of very high quality, whenever the whole mandible is visible, i.e. for the whole head and Accuitomo 170 scenarios, and some reduced quality in the case of local tomography.

Regularization The choice of the regularization term R and the parameter λ in Eq. (2) can have a drastic effect on the reconstructions $\hat{x}_t$, and with that especially on the motion estimation in Eq. (7). It is important to note that reconstructions showing high visual quality (which can for example be achieved by penalizing $\|\nabla x\|_2^2$) are not necessarily those, which are most useful for our motion estimation. Fig. 8 shows a study on different regularization terms. We find that strongly penalizing negative attenuation values, i.e. setting $R(x) = \|x^-\|_2^2$, which in theory breaks the linearity of Eq. (2), forces the conjugate gradients in a "more positive" direction and has a beneficial effect on our whole algorithm. All shown results were created using this regularization term with $\lambda = 10.000$ during the motion correction algorithm and $R(x) = \|\nabla x\|_2^2$ with $\lambda = 100$ for the final, depicted reconstructions.

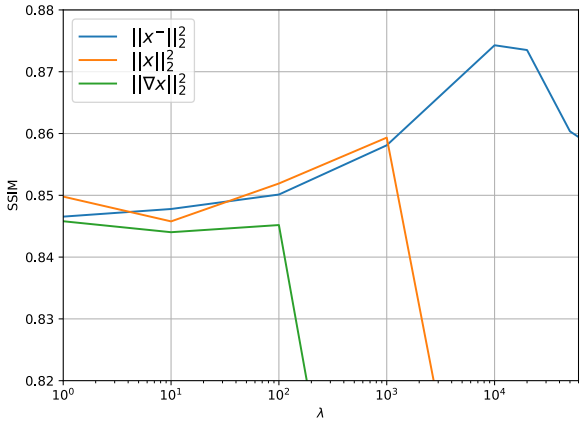


Figure 8: Influence of the regularization term on the quality of our method using the example of the Accuitomo 170 data set with high motion amplitudes. The plot shows the SSIM value between the final reconstruction and the ground truth. All intermediate reconstruction were obtained with the depicted regularization term and for the final reconstruction we applied the same regularization to all test cases ($R(x) = \|\nabla x\|_2^2$ with $\lambda = 100$).

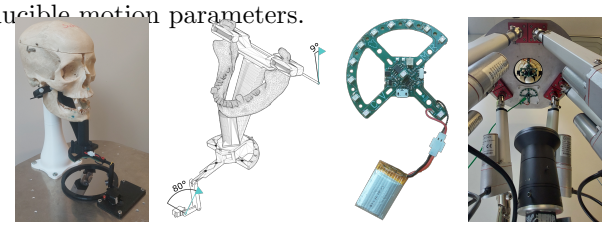
Runtimes Optimizing the runtime of our algorithm was not the primary focus of this work. Nevertheless we want to give a brief summary of the overall execution times of the previously discussed scenarios, which can be seen in Tab. 3. Our algorithm was run on a test system with an Intel i7-11700K processor and 64 GB of RAM equipped with an Nvidia RTX 3090 GPU. The resulting execution time is directly dependent on the number of input pixels and the dimension of the reconstructed volume. Using the resampling approach decreased the number of pixels and voxels along each dimension by 50% and 30% in the two first outer iterations, respectively, cutting the execution time roughly in half without negatively impacting the final reconstruction. Note that the output motion parameters are independent of image dimensions and volume sizes, so they can be used to create arbitrarily large reconstructions afterwards.

	Volume size	Image size	Runtime
Whole head	$300 \times 300 \times 300$	$300 \times 300 \times 516$	9 min
Accuitomo 170	$450 \times 300 \times 450$	$465 \times 370 \times 512$	23 min
Local Tomography	$450 \times 300 \times 450$	$357 \times 285 \times 600$	21 min

Table 3: Runtimes and dimensions of the different test scenarios. For this table we performed all three outer and five inner iterations without the usage of other stopping criteria to achieve a conservative runtime estimation.

IV.B. Real Data

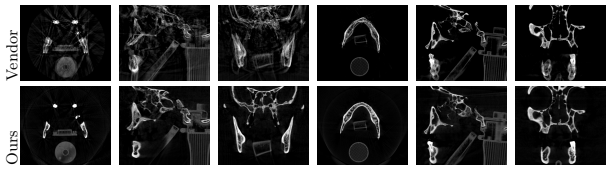
Figure 9: The skull is placed on top of the hexapod, at the height of a real patient's head. The platform moves during the acquisition, creating reproducible motion parameters.



(a) (b) (c) (d)

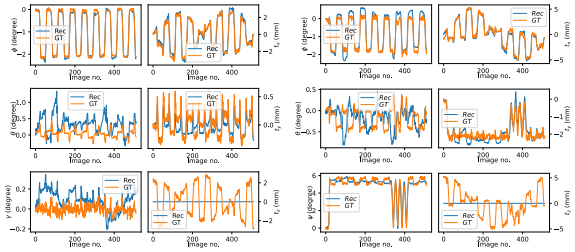
Figure 10: Individual components of the motion apparatus. Skull suspension and mandibular support (a), freedom of movement of the mandible (b), HSRM marker (c), placement of marker and camera on the platform (d).

Skull To verify the practicality of our approach with real CBCT machines, we conducted several scans of a skull placed on a robot platform. We simulate the patient's head movement utilizing a Stewart-Platform [49], often also called a hexapod. A hexapod is a composition of two platforms connected by six linear actuators. The position and orientation of the upper platform can be adjusted by changing the length of the individual actuators. Fig. 9 shows our setup positioned under the CBCT machine during a test run. We use a commercial solution for the Stewart-Platform [50], which we adapt and expand for our usage. In addition, we developed a construction for independent movement of the skull's lower jaw (Fig. 10a). We simulate the complex mandible movement as a simple rotation around the laterally



(a) Cranium and mandible (b) Mandible only

Figure 11: With our setup we were able to precisely control the movements (cranium and mandible) of a human skull while performing an actual CBCT acquisition. Here we show the reconstructions of the machine's manufacturer (top) and the output of our algorithm (bottom). The shown slices of the two scenarios slightly differ to properly highlight the affected regions. In scenario (a) the vendor's reconstruction suffers heavily from motion-induced artifacts, while our method was able to mitigate those up to a point where they are barely even noticeable. In the right half of this figure, reconstruction errors of the mandible are clearly observable in the top row, while the reconstruction of the cranium worked perfectly. By modeling the patient's mandibular motions independently, we were able to create an error-free reconstruction of the mandible in this case, too. $[-1000\,\mathrm{HU}, 2400\,\mathrm{HU}]$



(a) Cranium (b) Mandible

Figure 12: From the tracked positions and orientations of cranium and mandible we compute the angles and translations in the local coordinate system of the source-detector pair, as described in Sec. III. This figure compares the tracked motions (GT, orange curve) with the motion parameters found by our algorithm (Rec, blue curve). Again, the translation in the z-direction is kept at 0.

oriented axis positioned in the condyles area as a first approximation. Using a standard step motor, we can perform jaw movements with a rotation angle of around 9° in steps of 0.1° (Fig. 10b). To track the current position and orientation of cranium and mandible, we use an in-house built tracking system (HSRM Tracking[51]). As seen in Fig. 10d, we placed one marker on the backside of the platform and attached an additional one to the construction of the mandible holder. With a camera fixed to the lower platform, we can simultaneously track the current positions of cranium and mandible.

We run different motion profiles, concurrently moving the skull's cranium and the lower jaw. Fig. 11 shows the results of two different such profiles. For Fig. 11a we performed a (quite strong) periodic motion of the cranium while simultaneously moving the mandible (also see Fig. 12). In total we measured a movement of about 6 mm in the middle and 7.4 mm in the upper part of the reconstructed area. Fig. 11b shows a scenario where the lower jaw of the skull was opened and closed several times and then remained in a different position for the rest of the scan. With this motion profile the chin moved 9.5 mm between start and end position.

Patient Lastly we verify our method on a CBCT acquisition of a patient in a real clinical application. During this acquisition the patient performed a strong motion (about 1 cm displacement), rendering the uncorrected reconstruction unsuitable for further clinical usage (see Fig. 13a). In this case the patient had to undergo another CBCT-examination. The result of our method can be seen in Fig. 13c. Compared to the vendor's reconstruction, the quality could be

(a) Vendor ([−1000 HU, 1750 HU]) (b) Niebler et al. ([−1000 HU, 1750 HU]) (c) Ours ([−1000 HU, 1750 HU])

Figure 13: Reconstructions of motion-impaired data from a real patient. Written informed consent was obtained prior to publishing this image.

clearly enhanced and existing motion artifacts are mitigated. Due to the severe motion however, we increased N_{outer} and N_{inner} and performed a total number of 34 iterations. The image in the middle (Fig. 13b) shows the result when applying the method described by Niebler et al.[37] Our reconstruction shows further improvement of the overall quality, especially noticeable as sharper edges throughout the image, for example the transition between tooth and air is less blurry. The visible artifacts in the region of the lower jaw are due to a metal implant in one of the patient's teeth.

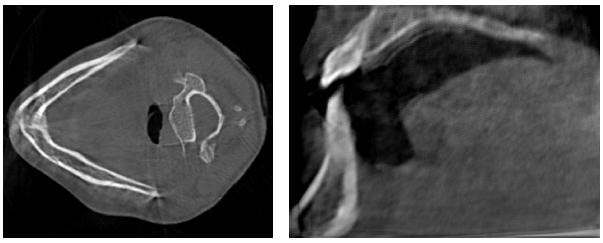
V. Discussion

Patient motion during the 10 to 40 s long exposure in maxillofacial CBCT is a frequent finding. Depending on the assessment method, patient motion was detected between 24%[52,53] and up to 78% of the CBCT-examinations[24]. Typical image degrading effects caused by such motion are motion blur, i.e. reduced spatial resolution and typical artefacts like stripe- and ring-patterns[7,8,54]. Since the long exposure times are due to hardware limitations and will most likely not be reduced considerably in the near future, the patient motion issue will also persist. We propose a marker-free method capable of enhancing the quality of motion-beset maxillofacial CBCT-data *a posteriori*. As a novelty, the method also reconstructs separate motion of the mandible relative to the cranium. This motion pattern has been observed in patient CBCT-examinations[24]. Even in cases with unrealistically large motion amplitudes of 6 mm the proposed method worked rather well. Using a convolutional neural net dealing with segmentations in the artefact-free 2D projection images from the CBCT-scan, the mandible is very reliably segmented from the remaining skull. Based on the previous work from Niebler and colleagues[37] our motion aware reconstruction based on the *CGLS* algorithm models patient motion by two separate rigid motions, i.e. that of the mandible and that of the cranium. This consideration is the main difference to previous works, e.g. Niebler[37] and the similar approach of Sun et al.[46], providing a much more versatile motion correction algorithm. Our real-world clinical case (Fig. 13) also proves the enhancement of the resulting volume reconstruction in an actual clinical application. In theory, it is even possible to apply a different reconstruction algorithm, for example the widely used *FDK* algorithm, but we found *CGLS* to produce results of higher quality.

It is also interesting to note, that the proposed method is capable of reliably estimating the motion occurring in a CBCT. This could also be used to track motion in existing data, whenever the 2D projection radiographs and geometric machine parameters are available. In a clinical or scientific context, estimation of the true patient motion yields helpful information for both CBCT image acquisition as well as further enhancement of the machines.

V.A. Limitations

All of our test cases assumed an aligned detector of the CBCT device and a full 360° rotation scan. Contrary, some clinical CBCT devices employ different strategies, e.g. by using a lateral-offset detector or by applying short scan



(a) (b)

Figure 14: Limitations of our algorithm: (a) A typical flawed output when using scan geometries with lateral-offset detectors. (b) Reconstruction of the scenario marked in Tab. 1 after motion correction. In this case the combination of high motion amplitudes and small ROI made the recovery of the true motion parameters impossible.

protocols in certain acquisition modes. In such cases, the patient is scanned mostly, or even completely, for only 180° plus cone angle. Similar to the methods analyzed by Santaella et al.[55] we find that our motion correction algorithm produces results of lower quality in setups with lateral-offset detectors, and it fails to output acceptable results for short scan data. In lateral-offset setups, recovery of motions of the cranium is still possible in many cases (albeit usually with less precision), but correcting separate mandibular motions was impossible in our test cases. A typical output with this setup can be seen in Fig. 14a.

When motion amplitudes are excessive, our proposed method cannot reconstruct the correct motion parameters because the bad quality of the initial reconstruction does not allow useful pose estimations. Usually it is still possible to reduce the projection error $\|A\tilde{x} - b\|_2$ and enhance the reconstruction quality, but the result may still be unusable for medical diagnosis. For example, with the local tomography scenario using the tested high motion amplitudes, our algorithm can improve both the projection error as well as the SSIM compared to the ground truth (Tab. 1), however the visual quality of the result is still not satisfactory and unsuitable for further usage (Fig. 14b). Unfortunately there is no hard threshold of when motion amplitudes are too high, since the quality of our results also depends on the given setup. The motion estimation works best for bigger ROIs and therefore fails earlier in narrow local tomography setups, as already seen in Sec. IV.

Real-world CBCT data is hard to come by. We could demonstrate our approach on one clinical data set and several scans of a moving skull, but the evaluation of our method would still benefit from more clinical data. This also applies to the mandible segmentation within the 3D volume. In our test cases, the label of the mandible did not grow into the region of the upper jaw. However, our segmentation method currently does not strictly enforce this. In such cases separate motion correction for cranium and mandible would not be possible or artifacts like double contours within the tooth row can arise. In these situations a tooth-type trained network could be helpful to ensure a more exact segmentation of the lower and upper teeth. We hope to perform further studies on more measurements in the future.

VI. Conclusion and Future Work

We presented a motion estimation and motion correction method for 3D-CBCT, based only on the 2D radiographic images. As a novelty, our method considers separate cranial and mandibular motions. The experiments showed that our algorithm is capable of consistently enhancing reconstruction quality. Quantitative comparisons of synthetically generated data from different scenarios (including local tomography) and qualitative comparisons of real acquisitions are provided. We found, that the proposed method was able to improve visual quality as well as the SSIM to the

ground truth in every case. In many cases previously unusable CBCT scans could be enhanced to allow for further clinical usage. Since some manufacturers use lateral-offset detectors to save hardware costs, our research group will focus future work on further improving results with these kinds of scanning geometries. Future research will be also directed towards refinement of the methodology and potential implementation in clinical work.

VII. Disclosure of Conflicts of Interest

The authors have no relevant conflicts of interest to disclose. The work was funded by a grant from the German Research Foundation DFG (SCHU1496/7-1). We gratefully acknowledge the staff of the division of oral radiology, University Medical Center of Mainz, Germany, for their help in acquiring the experimental CBCT data.

References

[1] P. Mozzo, C. Procacci, A. Tacconi, P. Martini, and I. Andreis, A new volumetric CT machine for dental imaging based on the cone-beam technique: preliminary results, Eur Radiol **8**, 1558–1564 (1998).

[2] Y. Arai, E. Tammisalo, K. Iwai, K. Hashimoto, and K. Shinoda, Development of a compact computed tomographic apparatus for dental use, Dentomaxillofac Radiol **28**, 245–248 (1999).

[3] A. O'Connell, D. Conover, Y. Zhang, P. Seifert, W. Logan-Young, C. Lin, L. Sahler, and R. Ning, Cone-beam CT for breast imaging: Radiation dose, breast coverage, and image quality, Am J Roentgenol **195**, 496–509 (2010).

[4] W. Hohenforst-Schmidt, P. Zarogoulidis, T. Vogl, J. Turner, R. Browning, B. Linsmeier, H. Huang, Q. Li, K. Darwiche, L. Freitag, M. Simoff, I. Kioumis, K. Zarogoulidis, and J. Brachmann, Cone Beam Computertomography (CBCT) in Interventional Chest Medicine - High Feasibility for Endobronchial Realtime Navigation, J Cancer **5**, 231–241 (2014).

[5] J. Tonetti, M. Boudissa, G. Kerschbaumer, and O. Seurat, Role of 3D intraoperative imaging in orthopedic and trauma surgery, Orthop Traumatol Surg Res **106**, S19–S25 (2020).

[6] C. Hodez, C. Griffaton-Taillandier, and I. Bensimon, Cone-beam imaging: Applications in ENT, Eur Ann Otorhinolaryngol Head Neck Dis **128**, 65–78 (2011).

[7] R. Pauwels, K. Araki, J. H. Siewerdsen, and S. S. Thongvigitmanee, Technical aspects of dental CBCT: state of the art, Dentomaxillofac Radiol **44**, 20140224 (2015).

[8] R. Schulze, U. Heil, D. Gross, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, Artefacts in CBCT: a review., Dentomaxillofac Radiol **40 5**, 265–273 (2011).

[9] R. Spin-Neto, L. Matzen, L. Schropp, E. Gotfredsen, and A. Wenzel, Factors affecting patient movement and re-exposure in CBCT examination, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology **119** (2015).

[10] L. Feldkamp, D. LC, and J. Kress, Practical cone-beam algorithm, J Opt Soc Am , 612–619 (1984).

[11] Y. A. Shepp and B. F. Logan, The Fourier reconstruction of a head section, IEEE Transactions on Nuclear Science **21**, 21–43 (1974).

[12] Z. Zhang, S. Ghadai, O. R. Bingol, A. Krishnamurthy, and L. J. Bond, A framework for 3D x-ray CT iterative reconstruction using GPU-accelerated ray casting, AIP Conference Proceedings **2102**, 030003 (2019).

[13] R. Pauwels, O. Nackaerts, N. Bellaiche, H. Stamatakis, K. Tsiklakis, A. Walker, H. Bosmans, R. Bogaerts, R. Jacobs, K. Horner, and S. P. Consortium, Variability of dental cone beam CT grey values for density estimations, Br J Radiol **86**, 20120135 (2013).

[14] R. Pauwels, R. Jacobs, S. Singer, and M. Mupparapu, CBCT-based bone quality assessment: are Hounsfield units applicable?, Dentomaxillofac Radiol **44**, 20140238 (2014).

[15] J. Sonke, L. Zijp, P. Remeijer, and M. van Herk, Respiratory correlated cone beam CT, Med Phys **32**, 1176–1186 (2005).

[16] R. Li, J. Lewis, X. Jia, T. Zhao, W. Liu, S. Wuenschel, J. Lamb, D. Yang, D. Low, and S. Jiang, On a PCA-based lung motion model, Phys Med Biol **56**, 6009–6030 (2011).

[17] S. Dhou, M. Hurwitz, P. Mishra, W. Cai, J. Rottmann, R. Li, C. Williams, M. Wagar, R. Berbeco, D. Ionascu, and J. Lewis, 3D fluoroscopic image estimation using patient-specific 4DCBCT-based motion models, Phys Med Biol **60**, 3807–3824 (2015).

[18] G. Chee, D. O'Connell, Y. Yang, K. Singhrao, D. Low, and J. Lewis, McSART: an iterative model-based, motion-compensated SART algorithm for CBCT reconstruction, Phys Med Biol **64**, 095013 (2019).

[19] M. Ranjbar, P. Sabouri, S. Mossahebi, A. Sawant, P. Mohindra, G. Lasio, and L. Topoleski, Validation of a CT-based motion model with in-situ fluoroscopy for lung surface deformation estimation, Phys Med Biol **66**, 045035 (2021).

[20] D. Thomas, J. Lamb, B. White, S. Jani, S. Gaudio, P. Lee, D. Ruan, M. McNitt-Gray, and D. Low, Validation of a CT-based motion model with in-situ fluoroscopy for lung surface deformation estimation, Int J Radiat Oncol Biol Phys **89**, 191–8 (2014).

[21] T. Zhang, H. Keller, M. O'Brien, T. Mackie, and B. Paliwal, Application of the spirometer in respiratory gated radiotherapy, Med Phys **30**, 3165–3171 (2003).

[22] D. Low, M. Nystrom, E. Kalinin, P. Parikh, J. Dempsey, J. Bradley, S. Mutic, S. Wahab, T. Islam, and G. Christensen, A method for the reconstruction of four-dimensional synchronized CT scans acquired during free breathing, Med Phys **30**, 1254–1263 (2003).

[23] H. Yan, X. Wang, W. Yin, T. Pan, M. Ahmad, X. Mou, L. Cerviño, X. Jia, and S. Jiang, Extracting respiratory signals from thoracic cone beam CT projections, Phys Med Biol **58**, 1447–1464 (2013).

[24] R. Spin-Neto, L. Matzen, L. Schropp, E. Gotfredsen, and A. Wenzel, Detection of patient movement during CBCT examination using video observation compared with an accelerometer-gyroscope tracking system, Dentomaxillofacl Radiol **46**, 201602 (2017).

[25] J. Hahn, H. Bruder, C. Rohkohl, T. Allmendinger, K. Stierstorfer, T. Flohr, and M. Kachelrieß, Motion Compensation in the Region of the Coronary Arteries based on Partial Angle Reconstructions from Short Scan CT Data, Medical Physics **44** (2017).

[26] C. Rohkohl, H. Bruder, K. Stierstorfer, and T. Flohr, Improving best-phase image quality in cardiac CT by motion correction with MAM optimization, Medical Physics **40**, 031901 (2013).

[27] A. Sisniega, J. Stayman, J. Yorkston, J. Siewerdsen, and W. Zbijewski, Motion compensation in extremity cone-beam CT using a penalized image sharpness criterion., Physics in medicine and biology **62 9**, 3712–3734 (2017).

[28] H. Huang, J. H. Siewerdsen, W. Zbijewski, C. R. Weiss, M. Unberath, T. Ehtiati, and A. Sisniega, Reference-free learning-based similarity metric for motion compensation in cone-beam CT, Physics in Medicine & Biology **67**, 125020 (2022).

[29] M. Berger, Y. Xia, W. Aichinger, K. Mentl, M. Unberath, A. Aichert, C. Riess, J. Hornegger, R. Fahrig, and A. Maier, Motion compensation for cone-beam CT using Fourier consistency conditions, Physics in Medicine & Biology **62**, 7181–7215 (2017).

[30] A. Preuhs, A. Maier, M. Manhart, M. Kowarschik, E. Hoppe-Preuhs, J. Fotouhi, N. Navab, and M. Unberath, Symmetry prior for epipolar consistency, International Journal of Computer Assisted Radiology and Surgery **14** (2019).

[31] A. Aichert, M. Berger, J. Wang, N. Maass, A. Doerfler, J. Hornegger, and A. K. Maier, Epipolar Consistency in Transmission Imaging, IEEE Transactions on Medical Imaging **34**, 2205–2219 (2015).

[32] T. Lossau (née Elss), H. Nickisch, T. Wissel, R. Bippus, H. Schmitt, M. Morlock, and M. Grass, Motion estimation and correction in cardiac CT angiography images using convolutional neural networks, Computerized Medical Imaging and Graphics **76**, 101640 (2019).

[33] J. Maier, S. Lebedev, J. Erath, E. Eulig, S. Sawall, E. Fournié, K. Stierstorfer, M. Lell, and M. Kachelrieß, Deep learning-based coronary artery motion estimation and compensation for short-scan cardiac CT, Medical Physics **48**, 3559–3571 (2021).

[34] A. Preuhs, M. Manhart, P. Roser, E. Hoppe, Y. Huang, M. Psychogios, M. Kowarschik, and A. Maier, Appearance Learning for Image-Based Motion Estimation in Tomography, IEEE Transactions on Medical Imaging **39**, 3667–3678 (2020).

[35] M. Berger, K. Müller, A. Aichert, M. Unberath, J. Thies, J.-H. Choi, R. Fahrig, and A. Maier, Marker-free motion correction in weight-bearing cone-beam CT of the knee joint, Medical Physics **43**, 1235–1248 (2016).

[36] B. Flach, M. Brehm, S. Sawall, and M. Kachelrieß, Deformable 3D–2D registration for CT and its application to low dose tomographic fluoroscopy, Physics in Medicine and Biology **59**, 7865–7887 (2014).

[37] S. Niebler, E. Schömer, H. Tjaden, U. Schwanecke, and R. Schulze, Projection-based improvement of 3D reconstructions from motion-impaired dental Cone beam CT data, Medical Physics **46** (2019).

[38] S. Ouadah, M. Jacobson, J. W. Stayman, T. Ehtiati, C. Weiss, and J. H. Siewerdsen, Correction of patient motion in cone-beam CT using 3D–2D registration, Physics in Medicine & Biology **62**, 8813–8831 (2017).

[39] E. Polak and G. Ribiere, Note sur la convergence de méthodes de directions conjuguées, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique **3**, 35–43 (1969).

[40] T. van Leeuwen, S. Maretzke, and K. J. Batenburg, Automatic alignment for three-dimensional tomographic reconstruction, Inverse Problems **34**, 024004 (2018).

[41] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy, Medical Physics **46**, 576–589 (2019).

[42] B. Cheng, O. Parkhi, and A. Kirillov, Pointly-Supervised Instance Segmentation, 2021.

[43] K. He, G. Gkioxari, P. Dollár, and R. Girshick, Mask R-CNN, in *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[44] T. Gietzen, R. Brylka, J. Achenbach, K. zum Hebel, E. Schömer, M. Botsch, U. Schwanecke, and R. Schulze, A method for automatic forensic facial reconstruction based on dense statistics of soft tissue thickness, PloS onet , e0210257 (2019).

[45] J. A. Nelder and R. Mead, A Simplex Method for Function Minimization, The Computer Journal **7**, 308–313 (1965).

[46] T. Sun, R. Jacobs, R. J. Pauwels, E. Tijskens, R. R. Fulton, and J. Nuyts, A motion correction approach for oral and maxillofacial cone-beam CT imaging, Physics in Medicine & Biology (2021).

[47] R. Spin-Neto, C. Costa, D. M. Salgado, N. R. Zambrana, E. Gotfredsen, and A. Wenzel, Patient movement characteristics and the impact on CBCT image quality and interpretability, Dentomaxillofacial Radiology **47**, 20170216 (2018), PMID: 28872352.

[48] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing **13**, 600–612 (2004).

[49] D. Stewart, A Platform with Six Degrees of Freedom, Aircraft Engineering and Aerospace Technology **38**, 30–35 (1966).

[50] ACROME Stewart Pro Platform, https://acrome.net/products/stewart-pro, Accessed: 2022-05-01.

[51] H. Tjaden, U. Schwanecke, F. Stein, and E. Schömer, High-Speed and Robust Monocular Tracking, VISAPP 2015 - 10th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings **3**, 462–471 (2015).

[52] R. Spin-Neto, C. Costa, D. Salgado, N. Zambrana, E. Gotfredsen, and A. Wenzel, Head motion during cone-beam computed tomography: Analysis of frequency and influence on image quality, Dentomaxillofac Radiol , 20170216 (2018).

[53] J. Moratin, M. Berger, T. Rückschloss, K. Metzger, H. Berger, M. Gottsauner, M. Engel, J. Hoffmann, C. Freudlsperger, and O. Ristow, Head motion during cone-beam computed tomography: Analysis of frequency and influence on image quality, Imaging Sci Dent , 227–236 (2020).

[54] R. Spin-Neto and A. Wenzel, Patient movement and motion artefacts in cone beam computed tomography of the dentomaxillofacial region: a systematic literature review, Oral Surg Oral Med Oral Pathol Oral Radiol , 425–433 (2016).

[55] G. M. Santaella, A. Wenzel, F. Haiter-Neto, P. L. Rosalen, and R. Spin-Neto, Impact of movement and motion-artefact correction on image quality and interpretability in CBCT units with aligned and lateral-offset detectors, Dentomaxillofacial Radiology **49**, 20190240 (2020), PMID: 31530012.

List of Figures

List of Tables

List of Algorithms

A Derivation of the Gradient

Here we derive the gradient of Eq. (7) with respect to the motion parameters p_C and p_M. As we separate Eq. (7) into n independent optimization problems and choose E as the $L2$-norm of the residual, we can write it for each image $b^{(i)}$ as

$$E(\tilde{x}, p_C^{(i)}, p_M^{(i)}) = \|\mathbf{A}^{(i)}(p_C^{(i)}, p_M^{(i)})\tilde{x} - b^{(i)}\|_2^2$$

and with that the gradient is given by

$$\nabla_p E = 2(\nabla_p \mathbf{A}^{(i)}(p_C^{(i)}, p_M^{(i)})\tilde{x})^T(\mathbf{A}^{(i)}(p_C^{(i)}, p_M^{(i)})\tilde{x} - b^{(i)}).$$

To compute $\nabla_p \mathbf{A}^{(i)}(p_C^{(i)}, p_M^{(i)})\tilde{x}$, we first look at every pixel c of the forward projected (discrete) volume, which is given by summing up the attenuation values along the discretized ray γ_c:

$$\begin{aligned}(\mathbf{A}(p_C, p_M)\tilde{x})_c = &\sum_{\mathbf{x}\in\gamma_c} [(1 - \lambda(T(p_C)(\mathbf{x}))) \cdot \tilde{x}(T(p_C)(\mathbf{x})) \\ &+ \lambda(T(p_M)(\mathbf{x})) \cdot \tilde{x}(T(p_M)(\mathbf{x}))] \\ = &\sum_{\mathbf{x}\in\gamma_c} (1 - \lambda(T(p_C)(\mathbf{x}))) \cdot \tilde{x}(T(p_C)(\mathbf{x})) \\ + &\sum_{\mathbf{x}\in\gamma_c} \lambda(T(p_M)(\mathbf{x})) \cdot \tilde{x}(T(p_M)(\mathbf{x})),\end{aligned}$$

where $\lambda(\mathbf{x})$ is the value of the 3D label at position $\mathbf{x} \in \mathbb{R}^3$ and T being the rigid transformation on the voxel position.

This results in the (separated) gradients

$$\begin{aligned}&\nabla_{p_C}(\mathbf{A}(p_C, p_M)\tilde{x})_c \\ &= \sum_{\mathbf{x}\in\gamma_c} [-J_T(p_C)(\mathbf{x}) \cdot \nabla_x \lambda(T(p_C)(\mathbf{x})) \cdot \tilde{x}(T(p_C)(\mathbf{x})) \\ &+(1 - \lambda(T(p_C)(\mathbf{x}))) \cdot J_T(p_C)(\mathbf{x}) \cdot \nabla_x \tilde{x}(T(p_C)(\mathbf{x}))] \\ &= \sum_{\mathbf{x}\in\gamma_c} J_T(p_C)(\mathbf{x}) \cdot [-\nabla_x \lambda(T(p_C)(\mathbf{x})) \cdot \tilde{x}(T(p_C)(\mathbf{x})) \\ &+(1 - \lambda(T(p_C)(\mathbf{x}))) \cdot \nabla_x \tilde{x}(T(p_C)(\mathbf{x})))] \\ &\nabla_{p_M}(\mathbf{A}(p_C, p_M)\tilde{x})_c \\ &= \sum_{\mathbf{x}\in\gamma_c} J_T(p_M)(\mathbf{x}) \cdot [\nabla_x \lambda(T(p_M)(\mathbf{x})) \cdot \tilde{x}(T(p_M)(\mathbf{x})) \\ &+\lambda(T(p_M)(\mathbf{x})) \cdot \nabla_x \tilde{x}(T(p_M)(\mathbf{x}))].\end{aligned}$$

$J_T \in \mathbb{R}^{6\times 3}$ can be obtained by computing the Jacobian of the rigid transformation $T(p)$ with respect to the parameter $p \in \mathbb{R}^6$ and $\nabla_x \tilde{x}, \nabla_x \lambda \in \mathbb{R}^3$ as the spatial gradient of the reconstruction and label, respectively. In our implementation we use linear interpolation on both volumes, which enables a fast computation of these gradients as they can be implemented via texture lookups on the GPU.