

CBCT of a Moving Sample from X-rays and Multiple Videos

Julien Pansiot, Edmond Boyer

▶ **To cite this version:**

Julien Pansiot, Edmond Boyer. CBCT of a Moving Sample from X-rays and Multiple Videos. IEEE Transactions on Medical Imaging, 2019, 38 (2), pp.383-393. 10.1109/TMI.2018.2865228 . hal-01857487

HAL Id: hal-01857487

https://inria.hal.science/hal-01857487

Submitted on 16 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CBCT of a Moving Sample from X-rays and Multiple Videos

Julien Pansiot and Edmond Boyer

***Abstract*—In this paper we consider dense volumetric modeling of moving samples such as body parts. Most dense modeling methods consider samples observed with a moving X-ray device and cannot easily handle moving samples. We propose instead a novel method to observe shape motion from a fixed X-ray device and to build dense in-depth attenuation information. This yields a low-cost, low-dose 3D imaging solution, taking benefit of equipment widely available in clinical environments. Our first innovation is to combine a video-based surface motion capture system with a single low-cost/low-dose fixed planar X-ray device, in order to retrieve the sample motion and attenuation information with minimal radiation exposure. Our second innovation is to rely on Bayesian inference to solve for a dense attenuation volume given planar radioscopic images of a moving sample. This approach enables multiple sources of noise to be considered and takes advantage of very limited prior information to solve an otherwise ill-posed problem. Results show that the proposed strategy is able to reconstruct dense volumetric attenuation models from a very limited number of radiographic views over time on synthetic and in-situ data.**

***Index Terms*—radiography, video, tomography, CBCT, Bayesian**

I. Introduction

THE ability to capture intrinsic body structure in motion is of interest in a number of fields related to medical imaging such as computer-assisted surgery, anatomy, biomechanics, and sports science. Most existing applications consider video or depth cameras and infer internal structure information, e. g. skeletal motion from surface observations using prior models. However, this strategy does not provide real measures on the internal structure and the actual bone structure can be far from the prediction due to multiple factors such as inaccurate model and complex elastic tissue motion [1]. With the aim to provide means to observe intrinsic structures in motion, we investigate in this paper a new strategy that recovers dense 3D volumetric models of moving samples, as illustrated in Figure 1.

To this purpose, we combine a video-based surface motion capture system that provides motion cues, with a X-ray imaging apparatus that captures the inner structure. As a preliminary step towards unconstrained three-dimensional volumetric motion capture, we investigate first rigidly moving samples, assuming limited prior knowledge on the captured samples. A key idea of our approach compared to traditional computed tomography is that we do not consider sample motion as low-amplitude noise to be corrected, but at the contrary as a source of information, ensuring the capture of X-ray images from multiple viewpoints.

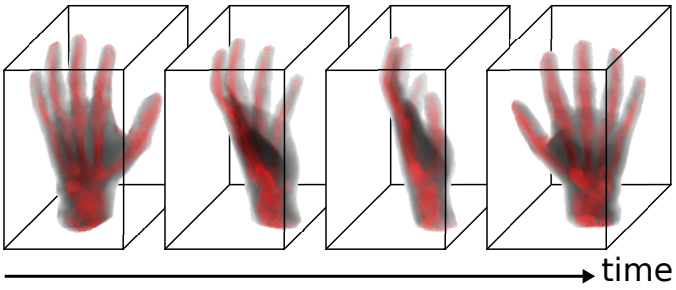


Fig. 1. Dense volumetric attenuation reconstruction from a rigidly moving sample captured by a single planar X-ray imaging device and a surface motion capture system. Higher attenuation (here bone structure) is highlighted in red.

The X-ray imaging apparatus can include one or more X-ray sensors in a similar fashion. Multiple sensors, albeit more complex setups, enable a wider range of sample motion and shorter acquisition time. Yet less accurate than a CT-scanner with static subjects, it nevertheless yields a less expensive low-dose solution, taking benefit of equipment widely available in clinical environments. We present in this paper a generic capture and reconstruction method potentially suited to various scenarii. While clinical extremity imaging, e. g. hand or weight-bearing foot motion, are obvious applications, the approach could also handle other moving shapes.

Our volumetric reconstruction method builds on image super-resolution techniques [2] to optimally exploit X-ray samples and infer 3D attenuation. It relies on an X-ray image formation model (Figure 2) accounting for 2D sensor noise as well as 3D geometric errors. This model is associated with a volumetric L_1 smoothness prior to constrain the reconstruction, hence allowing for a limited number of input views. All these elements are integrated within a Bayesian framework for backward inference.

To summarize, our approach is based on two key contributions. First, we have proposed a novel imaging process by combining 2D X-ray and 3D video imagery for tomographic reconstruction. And secondly we have introduced a practical Bayesian method to 3D imaging of a moving sample which accounts for both sensor and calibration inaccuracies using a generative model.

The remainder of this paper is organised as follows: related

Julien Pansiot and Edmond Boyer are with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France.
*Institute of Engineering Univ. Grenoble Alpes.
{julien.pansiot,edmond.boyer}@inria.fr

Copyright © 2018 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

work is presented in Section II, Bayesian image formation model in Section III, model estimation in Section IV, experimental design in Section V, results in Section VI, discussion in Section VII, before concluding in Section VIII.

II. Related Work

Our objective in this paper is to reconstruct a dense 3D volumetric attenuation model of a rigidly moving sample. Currently, the two well-established classes of radiographic imaging methods are planar radiography and Computed Tomography (CT).

On the one hand, CT methods reconstruct accurate dense 3D volumetric attenuation models, however they require sample immobility, generally expose patients to high ionising radiations [3], and present high costs. The fastest Multi-detector CT (MDCT) are only able to capture in 100-200ms with significant limitations on the capture volume access. Hence, generic 4D-CT capture of non-cyclic motion such as extremity musculoskeletal features can only be performed reliably when the sample motion speed is a fraction of the scanner rotation speed [4]. In the general case, motion artefacts such as ghosting, blurring, or streaks depend on the relative axis of motion and cannot be entirely avoided during capture [5]. Motion artefact correction can be performed in post-processing, and has been widely covered by relying on either specific anatomical models [6] or more generic approaches such as optical-flow [7]. Conceptually closer to our purpose, this issue has also been tackled by combining (RGB)D images to X-rays for motion tracking using either RGBD SLAM [8] or probabilistic ICP [9]. These methods only succeed on relatively minor, local motion artefacts. With the large-scale unpredictable motions that we aim at imaging, reconstruction without considering motion would simply not lead to any meaningful 3D image. Our strategy differs since we consider motion as a feature, not as noise. This could consequently significantly simplify acquisition setups and help capturing moving samples.

On the other hand, planar cine-radiography does not suffer from these drawbacks but raw data is limited to two-dimensional projected and integrated information.

In order to find a compromise between these two classes of radiographic methods, a new type of procedure, few-views Cone-Beam CT (CBCT) [10], has gained interest. Several methods have been implemented to reconstruct 3D attenuation models from a limited number of cone-beam images [11], using for example an isocentric C-arm [3]. With a reduced number of input images, relatively accurate attenuation models can be reconstructed for e. g. breast tomosynthesis [12], [13]. Nevertheless these methods are still limited to static samples.

A significant amount of research has also focused on recovering features from bi-planar radiography, based on prior models [14], [15], [16], which can then be used to capture moving samples [17]. While these methods produce medically relevant results, they require strong prior anatomic models and usually some amount of manual intervention. With no anatomical model, sparse data can nevertheless be captured such as the position of purposefully implanted markers [18], [19] or specific, well-segmented features [20], [21]. In both cases however, these markers and features are in fact prior models.

A method to reconstruct 3D attenuation from a limited number of arbitrary X-ray views was proposed in [22], but assumes reasonably good calibration. Our earlier approach [23] introduced volumetric attenuation reconstruction process from a limited number of X-ray views of a moving sample which motion is estimated using videos. However this approach was not grounded on a formal model and addresses calibration or motion inaccuracies in an ad hoc fashion. Practical cross-calibration of a RGB(D) system with X-rays has been addressed in a number of previous works using 2D objects such as a visible checker-board with X-ray opaque balls [24] or a metal sheet checker-board [8], as well as 3D objects such as markers on a 3D cylinder [25], 3D rings and spheres [26], or a hex-faced marker [27].

In contrast, we introduced in [28] a novel strategy based on a generative image formation model that can easily handle uncertainties within a Bayesian framework to solve the tomographic problem. This strategy takes inspiration from image super resolution [2] to optimally use observations from sparse sets of views. So far, most Bayesian approaches to tomography have been limited to lower-dimensionality problems such as 2D slices [29], lower resolution imaging modalities such as emission tomography [30], as well as sparse (e. g. feature-based) or discrete [31] reconstructions. This work extends them to dense 3D volumetric attenuation models.

In [32] we also considered non-rigid samples with, however, a simplified reconstruction approach. The purpose of this article is to thoughtfully evaluate the Bayesian strategy in the CBCT context.

In this paper, we extend [28] with a thorough result analysis, in particular in the presence of multiple noise sources, as well as varied capture conditions.

III. Generative X-ray image model

As mentioned, our generative model builds on an image formation model [2] to explain the X-ray observations given the 3D model. Our method takes as input a set of X-ray images of a rigidly moving sample. The images are first registered in a common framework using the motion estimated by a multi-view capture system. A dense attenuation model of the moving sample, represented as a voxel grid, is then reconstructed using the entire X-ray image sequence.

We detail below the main components of this model. In order to account for the multiple sources of noise present in the acquisition process, we introduce a generative image formation model, as illustrated in Figures 2 and 3. This model is associated with a sparse prior, i. e. a TVL_1 norm, on the attenuation voxel grid.

A. Image formation

Given the sample 3D attenuation, the X-ray image is formed by projection and integration of the X-rays attenuated through the partly transparent sample by photoelectric absorption. We discretise the continuous absorbance problem in 3D as a

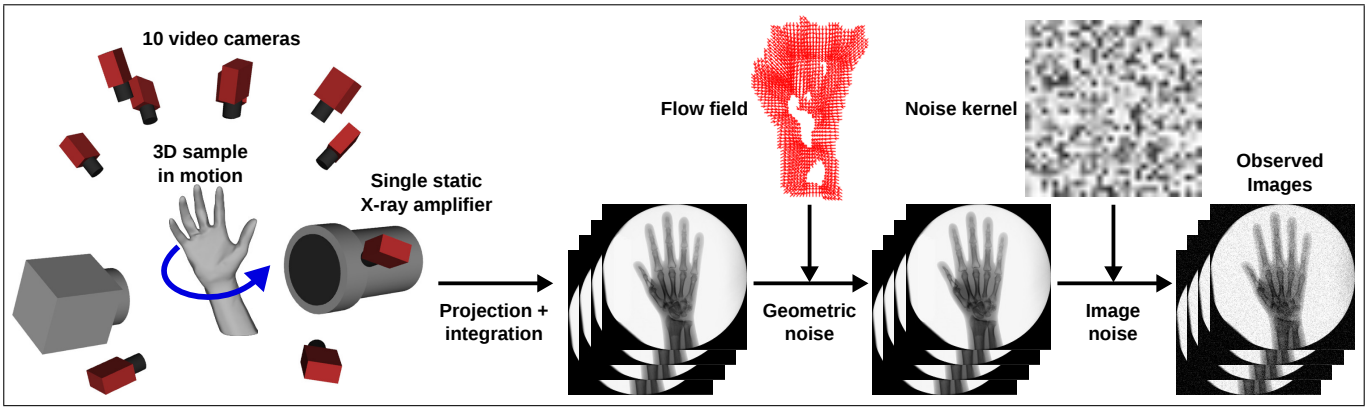


Fig. 2. X-ray image formation model for a moving sample observed by a single static planar X-ray amplifier. Video cameras are used to recover the sample surface motion. The formation model is composed of 3D attenuation projection and integration into a 2D image, geometric noise, and image noise.

weighted sum over the voxels v_j along the given ray ρ. d_j is the distance covered within the voxel v_j and μ_j the attenuation assumed uniform within v_j. We can then define the absorbance $I(\rho)$ in function of the emitted and transmitted intensities L_0 and $L(\rho)$:

$$I(\rho) = -\log \frac{L(\rho)}{L_0} = \sum_{j \in \rho} d_j \mu_j. \tag{1}$$

In the reminder of this paper, we assume that L_0 is known, and therefore work only on absorbance images, in which each pixel value is equal to the absorbance $I(\rho)$ over its corresponding ray ρ given the camera projection model.

The theoretical absorbance model described here-above describes how X-ray radiation interacts with a 3D sample to form a 2D image. In real scenarii however, several sources of noise affect the image formation, and therefore a more comprehensive image formation model must be devised as illustrated in Figure 2. We consider a sequence of N images $\{I_i\}$ as a vector juxtaposition $I = [I_0|\dots|I_N]$ acquired from a volume discretised as a voxel grid with attenuations $V = \{\mu_j\}$. For each image I_i, this model includes:

1) A known projection and integration matrix P_i composed of the coefficients d_j obtained from motion estimation. In the ideal case, we would have $P_i V = I_i$. We denote the projection matrix juxtaposition for all images $P = [P_0|\dots|P_N]$.
2) Geometric noise, i. e. the errors in the projection matrix P_i. This includes the inaccuracy in the motion and projection matrix (calibration) estimation as well as the deviation from purely rigid motion. It is modeled by a warping matrix F_i implemented as a discrete optical flow w_i.
3) A 2D Gaussian image noise level θ_i accounting for the radiation source, the sample (scattering), the amplifier, and the imaging sensor. While the actual noise in the X-ray image is a Poisson process, a Gaussian approximation yields nonetheless good results. This is especially important since we rely on low-dose cine-radioscopic sequences, achieved through fast shutter and low power.

The complete image formation model for an image I_i, summarised in Figure 3, is the following:

$$I_i = F_i(w_i) P_i V + \mathcal{N}(\theta_i) \tag{2}$$

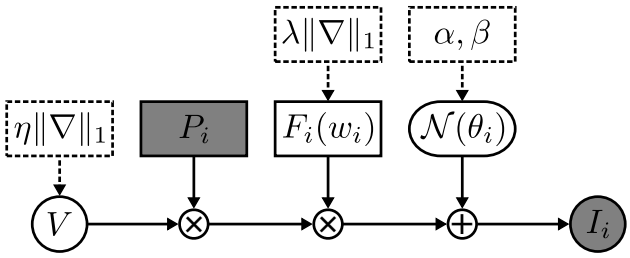


Fig. 3. Image formation model for X-ray absorbance imaging for a single image I_i. Vectors are represented by circles/rounded boxes, matrices by solid boxes, and priors by dashed boxes. Input data to the tomographic problem is filled in grey. The volume V is projected by P_i, warped by F_i (encoded as an optical flow w_i), and noise is added by a Gaussian kernel $\mathcal{N}$ of variance θ_i. $\|\nabla\|_1$ represents the norm of the gradient, and η, λ, α, and β are relative contribution weights.

B. Bayesian model

Our aim is to recover the 3D attenuation V given the absorbance image sequence I and the projection matrices P, i. e. to invert the model described previously. For this purpose we rely on a MAP estimation to find the optimal solution in terms of attenuation and noise:

$$\{V^*, \{F_i\}^*, \{\theta_i\}^*\} = \underset{V, \{F_i\}, \{\theta_i\}}{\operatorname{argmax}} \; p(V, \{F_i\}, \{\theta_i\}) | \{I_i\}), \tag{3}$$

Following Bayes law and assuming statistical conditional independence between images given the attenuation model:

$$p(V, \{F_i\}, \{\theta_i\}) | \{I_i\}) \propto p(V) \prod_i p(F_i) \prod_i p(\theta_i) \prod_i p(I_i | V, F_i, \theta_i). \tag{4}$$

C. Priors

Geometric noise appears in the acquisition process as a result of, primarily, calibration errors and object motions that

are not exactly rigid. We model this noise effect by a warping function F_i, estimated using the optical flow w_i [33] between the observed image I_i and the generated one P_iV.

As the inverse problem (4) is ill-posed and noise-ridden, we introduce noise and model priors. Given the nature of the data typically observed, the sparsity of the derivative responses is used as a prior for the 3D attenuation volume as in [2]:

$$p(V) = \eta^{dim(V)} e^{-\eta \|\nabla V\|_1}, \tag{5}$$

where η is the gradient weight and the following gradient notation is used, with q the voxel index and $V_k = \partial V / \partial k$:

$$\|\nabla V\|_1 = \sum_q (\|V_x(q)\|_1 + \|V_y(q)\|_1 + \|V_z(q)\|_1). \tag{6}$$

The minimisation of the L_1 norm of the gradient, or Total Variation TVL_1, favours continuous volumes separated by potentially high, albeit localised gradients. This is commonly observed in human tissues, often exhibiting relatively smooth tissues separated by larger local variations.

D. Image likelihood

The likelihood distribution is modeled as an exponential distribution:

$$p(I_i|V, F_i, \theta_i) = \theta_i^{\dim(I_i)} e^{-\theta_i \|I_i - F_i P_i V\|_1}, \tag{7}$$

where the 2D image noise level θ_i follows a Gamma distribution [2]:

$$p(\theta_i; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} e^{-\theta_i \beta}. \tag{8}$$

IV. Model estimation

In order to solve for the parameters in the MAP estimation (3), we use a coordinate descent scheme [2], that iteratively cycles through the independent estimation of each parameter group: the original volume V, the image noise level θ, and the motion noise warp F as detailed previously. The volume V is initialized using the ART+TVL_1 method [23] applied with P and I. We also assume that, initially, $\theta_i = 1$ and $F = Id$. The corresponding algorithm is summarised below.

Algorithm 1: Bayesian attenuation model estimation

In : (P, I)
Out: V
$V = ART + TVL_1(P, I)$ ▷ [23]
$\theta = 1, F = Id$
▷ Main loop
repeat over *nb_main_iterations*
  $F = estimate_flow(P, V, I, \theta)$ ▷ [33]
  $\theta = estimate_noise(P, V, I, F)$ ▷ (14)
  ▷ Volume estimation with IRLS
  repeat over *nb_IRLS_iterations*
    $W_{g,i} = weights(F, P, V, I)$ ▷ (11) (12)
    $V = solve_conj_grad(P, I, F, \theta, W_{g,i})$ ▷ (10)

The following subsections detail the individual parameter estimation.

A. Sample pose estimation

The rigid motion of the sample is estimated over the time sequence prior to the reconstruction. The rigid motion matrix at time i is multiplied to the constant X-ray device projection matrix to obtain the projection matrix P_i.

In order to estimate the sample rigid motion, we use the multi-view colour images and a silhouette-based approach [34] to generate a 3D polyhedral surface of the sample at each frame. Other approaches with better precision could be considered here, however the silhouette-based approach is robust and our purpose is not to estimate locally precise 3D models but globally robust 3D motions. Given the 3D mesh models over the sequence, we register them to a reference frame, for example the first frame, with a robust Iterative Closest Point (ICP) method with outlier detection [35]. Residual geometric artefacts are corrected in the global optimisation as detailed in Sec. IV-D.

B. Attenuation Volume

Given the current estimates for the flow field w_i and noise θ_i, we estimate the volume V based on the image set $\{I_i\}$ and the gradient prior η by minimizing:

$$V^* = \underset{V}{\operatorname{argmin}} \sum_i \theta_i \|F_i P_i V - I_i\|_1 + \eta \|\nabla V\|_1. \tag{9}$$

To this aim, we rely on Iteratively Reweighted Least Squares (IRLS) [2] minimisation. Despite its somewhat misleading name, IRLS allows to minimise with any L_N norm. The idea behind IRLS is to iteratively update the weights of a Weighted Least Squares minimisation in order to account for the residual between a target norm, here L_1, and the minimised L_2 norm. Hence, in order to minimise $|b - Ax|$ with L_1, we iteratively compute the residual $e = (b - Ax)$ and the derived weights $W^2 = diag(|e_i|^{-1})$ and minimise $\|W(Ax - b)\|^2$, i. e. solve for $A^T W^2 Ax = A^T W^2 b$. We therefore iterate between the two following steps.

First, we solve the following Weighted Linear Least Squares problem using the conjugate gradient method:

$$\left[\eta(D_x^T W_g D_x + D_y^T W_g D_y + D_z^T W_g D_z) + \sum_i \theta_i P_i^T F_i^T W_i F_i P_i \right] V = \sum_i \theta_i P_i^T F_i^T W_i I_i, \tag{10}$$

where D_x, D_y, and D_z are the derivative operators expressed as matrices. The 3 weighted terms in eq. (10) are the TVL_1 minimisation prior, the norm residual, and the observation term, respectively.

Second, we estimate the diagonal weight matrices as follows:

$$W_g = \text{diag}(\Phi(\nabla V)), \tag{11}$$

$$W_i = \text{diag}(\Phi(F_i P_i V - I_i)). \tag{12}$$

The following linear approximation for the L_2 norm for a given ϵ is used to enable gradient-based methods, with the negative exponent coming from the residual weight form:

$$\Phi(V)(q) = \left(V(q)^2 + \epsilon^2\right)^{-1/2}. \tag{13}$$

C. Image/sensor noise

Given the current volume V of M voxels and flow field F_i we estimate the image noise level θ_i, based on the assumption that the residual error follows a gamma distribution [2]:

$$\theta_i = \frac{\alpha + M - 1}{\beta + \overline{x}} \quad \text{with} \quad \overline{x} = \sum_{q=1}^{M} |(I_i - F_i P_i V)(q)|. \quad (14)$$

Since precise measurements of the actual actual noise levels are not necessarily available, we assume a conservative $(\alpha, \beta) = (1, 1)$.

In theory, the image noise is due to a number of factors in the X-ray imaging pipeline, as in other imaging modalities, and therefore does not strictly follow a Poisson distribution and even less so a Gaussian one. Nevertheless, our experiments show that this assumption is sufficient for the considered problem.

D. Geometric correction

The residual motion is estimated using the optical flow w_i [33] between the observed image I_i and the projected volume $P_i V$, as illustrated in Figure 4. This multi-resolution differential approach to optical flow combines local and global strategies, i. e. Lucas-Kanade and Horn-Schunck approaches. This allows for both global smoothness and local detail preservation.

Given the current volume V and the noise level θ_i, we estimate the flow w_i associated to the warp matrix F_i. We then update the warped image vector $I'_i = F_i^{-1} I_i$ using the flow and reformulate the data term in (9) as:

$$\sum_i \theta_i \| P_i V - F_i^{-1} I_i \|_1. \quad (15)$$

The relative gradient (TVL_1) weight η and optical flow weight λ ratio is a trade-off between edge preservation and high-frequency noise reduction, and will depend on input data noise levels. Without any attempt for fine-tuning, we have settled for $(\eta, \lambda) = (2, 1)$ for all our experiments (synthetic and real), which demonstrates that while important, this ratio is not critical.

E. Performance optimisation

Given the dimensionality of the problem at hand, several steps are required to ensure practicality and performance of the described method. First, only the pixels within the X-ray image silhouettes and only the voxels within the X-ray visual hull they define are considered. In the experiments described hereafter, this reduces the pixel count by a factor 2 to 5 and the voxel count by a factor 10. Yet still, the projection matrix P sparsity (i. e. 0.01% of non-zero coefficients) must be taken advantage of with a specific implementation [36]. Lastly, this matrix is pre-conditioned for improved performance and memory usage by dual voxel/ray sorting according to their respective 3D localization.

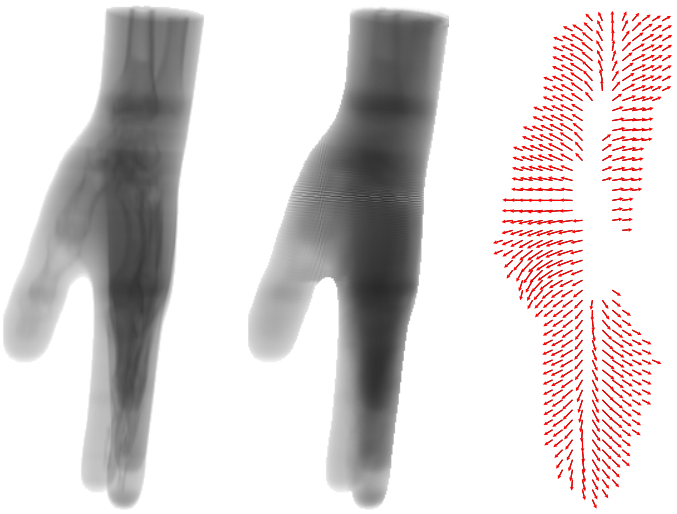


Fig. 4. Geometric correction by optical flow at an early iteration (volume reconstruction not sharp yet). This globally divergent flow denotes the geometric correction of an apparent scale factor, most likely due to geometric noise along the optical axis.

V. Experimental design

Two sets of experiment were carried out to validate the proposed framework. First the CT scan data of a phantom model was used to simulate image observations. This allows to evaluate the proposed approach under varied conditions with readily available ground truth. Secondly, the actual phantom model was placed into our hardware platform. This experiment allows for ground truth comparison in an actual capture platform.

A. Capture platform

We first present the platform used to capture the phantom dataset.

1) Hardware configuration: The platform is composed of ten colour video cameras capable of up to 100 frames per second (fps) at 2048×2048 pixels and a single Siemens ARCADIS Avantic X-ray C-arm (1024×1024 pixels) left static in our experiments, and capable of up to 30fps. The reconstruction was performed in a volume right next to the centre of the X-ray amplifier (diametre: 33cm), which therefore allows for a maximal theoretical radiographic resolution of 0.32mm. In practice, the reconstruction was performed in a cube of 26×26×26 cm sampled by 256×256×256 voxels, i. e. a theoretical resolution of 1.0mm. An example of such input data in given in Figure 5.

The video sub-system ensures sub-millisecond synchronisation between the RGB cameras. However, the X-ray amplifier C-arm could only be loosely synchronised with the video sub-system, both system running nevertheless at the same frequency of 30fps. Hence the synchronisation is accurate within half-a-frame (16.7ms).

2) Cross-calibration: In order to proceed with three-dimensional computation, the entire platform must be calibrated within the same coordinate system. For all ten cameras

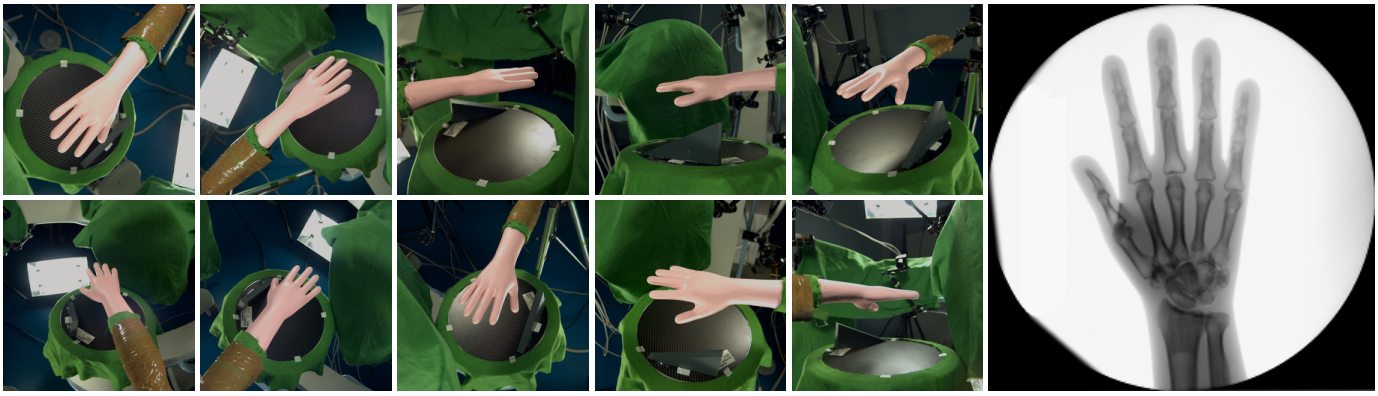


Fig. 5. Example of input at a given frame on the forearm phantom. Left: ten video images (2048×2048) from different view points, right: one radiographic image (1024×1024). Green backdrop has been installed on specular material to improve automated video segmentation.

and the X-ray amplifier, the pose, projection matrix, and distortion coefficients are calculated. The calibration is performed in two stages to optimise the X-ray usage. First, the video subsystem is fully calibrated on its own based on the detection of a moving LED wand followed by the parameter optimisation with a bundle adjustment.

Secondly, the X-ray amplifier is calibrated with respect to the video sub-system using a continuously moving lead ball (diametre: 10mm, weight: 5.5g) covered in orange paint for efficient segmentation, as illustrated in Figure 6. The ball is segmented and identified automatically in both the X-ray and video images. The positions in the multiple video images are then triangulated into a set of 3D positions. The X-ray amplifier is finally calibrated using the 2D/3D correspondences using the conventional approach by Zhang [37], similarly to the method proposed by [24].

ince the ball is moved relatively fast to reduce the operator radiation exposure and the hardware synchronisation is up to one frame, automated time synchronisation and spatial calibration are performed simultaneously in order to improve accuracy. More specifically the time offset (with sub-frame resolution) and optionally the frequency ratio between the two sub-systems are estimated. For this purpose, the conventional 3D calibration is run iteratively while varying hierarchically the offset and frequency so to minimise the reprojection error.

The calibration features reprojection error (RMS) was 1.804 pixels, which corresponds to a sub-voxel accuracy after projection.

B. Synthetic radiographic and video data from CT

In order to get a good overview of our approach performance under varied conditions, we first relied on a synthetic dataset, which allows for direct ground truth comparison.

A forearm phantom, consisting of a real human forearm skeleton cast in resin was first scanned with a regular CT device. A complete capture pipeline was then simulated solely from the phantom scan which was rendered by 10 virtual video cameras and one virtual planar X-ray image using ray-casting [38]. This method only simulates the photoelectric effect and thus the resulting images are free of scattered radiation as well as other sources of noise. They however exhibit 8-bit quantisation errors, added to simulate the effective quantisation of the low-dose C-arm used in the experiments. The phantom scan model was then moved artificially, following roughly a 180 degrees rotation. The phantom motion was estimated using ICP on the mesh computed from the video input for comparison with [23]. The proposed reconstruction method was then applied to the synthetic data.



Fig. 6. Calibration ball. Left: cropped video input from a selected camera; right: radiographic image. While 2 C-arms are visible in this image, a single one is used in this work.

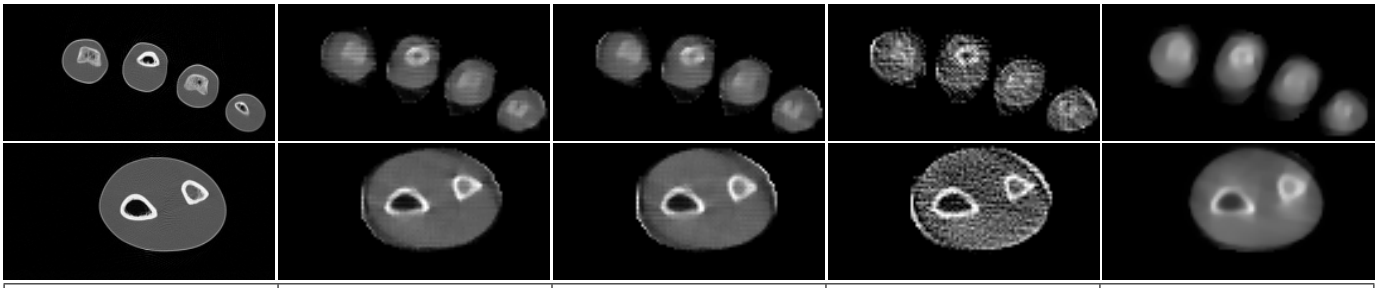
C. In-situ forearm phantom

The forearm phantom presented here above was placed into the capture platform described before. Such phantom allows for testing the actual hardware platform, with partial ground truth available.

It was then moved manually to follow roughly a 180 degree rotation. An example of the captured data at a given frame is provided in Figure 5.

This sequence was captured twice with diametrically opposed sets of X-ray tube parameters. A first set with long exposure and low amperage (referred to as DR) and a second set with short exposure and high amperage (referred to as DCM), as illustrated in Figure 8 and summarised in table I.

A green backdrop has been installed on specular material to improve automated video segmentation. Without this addition the segmentation might be slightly noisier, which would lead to a noisier 3D surface reconstruction. This is equivalent to sample motion noise which impact is covered in Sec. VI-A4.



RMS error	**0.063**	0.070	0.090	0.072
MI score	**0.442**	0.425	0.318	0.309
Ground-truth CT	**Proposed method**	**Without optical flow**	**Without TVL_1 prior**	**ART+TVL_1 [23]**

Fig. 7. Results on synthetic data: RMS and Mutual Information (MI) score for 2 selected slices, with best scores in bold. Without TVL_1 prior, the proposed algorithm does not converge. The contrast is better with the proposed approach (better MI as compared to ART+TVL_1) even though artefacts appear on the edges as a result of aliasing during the data simulation process (higher RMS as compared to ART+TVL_1). ART+TVL_1 performs relatively well in part due to the fact that synthetic data are close to the noiseless theoretical model, as well as minimising to an L_2 norm, which is coincidently the metric used in RMS evaluation.

TABLE I
X-RAY TUBE CONFIGURATIONS AND RELATIVE EXPOSURE

Dataset	kV(p)	X-ray current	Pulse width	Expos. /frame	Frames	Total expos.
	kV	*mA*	*ms*	*mAs*		*mAs*
DR	54	700	1056	739.2	20	14784
DCM	57	14500	7	101.5	32	3248

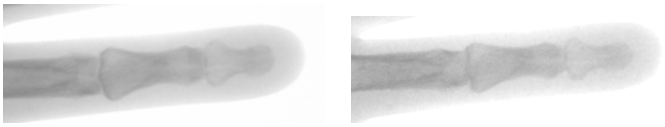


Fig. 8. In-situ phantom dataset input images (cropped). Left: DR, right: DCM. DCM images exhibit more noise.

VI. RESULTS

We present in this section the results of the synthetic dataset reconstruction with a number of variations in the process and the reconstruction of the actual forearm phantom in the platform.

A. Synthetic radiographic and video data from CT

The performance of individual components of the algorithm (optical flow, TVL_1 prior) were analysed independently, as illustrated in Figure 7. We note that on synthetic data, our approach exhibits slightly better contrast than ART+TVL_1 with respect to the Mutual Information (MI) score. We also note that the ART+TVL_1 reconstruction performs relatively well on synthetic data. This can be explained in part because the synthetic data does not exhibit large levels of noise, and hence the observations are close to the expected model. Furthermore, ART+TVL_1 reconstructions show relatively low RMS scores because ART+TVL_1 reconstruction is performed by minimising globally to an L_2 norm (albeit with L_1 regularisation which mostly impacts local areas), which is precisely what the RMS measure is quantifying.

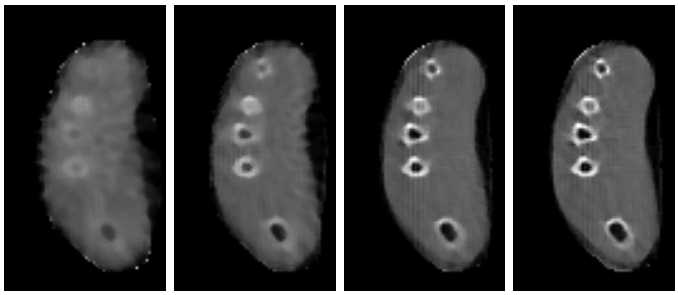


Fig. 9. Results on synthetic data for a selected slice based on varying numbers of input frames. Left-to-right: 8, 16, 32 and 64 frames. Skeletal structures are visible with 16 frames, detailed features require 32 frames, and noise is further reduced with 64 frames.

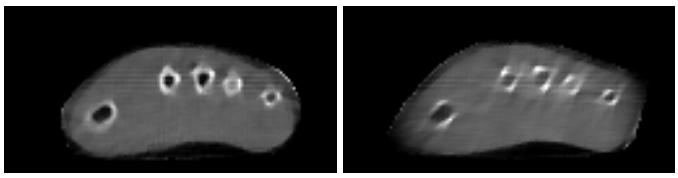


Fig. 10. Results on synthetic data for a selected slice based on varying input angular range. Left: 32 frames roughly distributed over 180 degrees; right: over 90 degrees.

We then ran a set of experiments to determine the range of conditions in which the proposed approach can be applied by purposefully degrading the input data. In the following, we show the impact of the number of input frames (VI-A1), the motion range in the input dataset (VI-A2), the image sensor noise (VI-A3), and the geometric noise (VI-A4). We also evaluated the quality of the reconstruction with the number of iterations (VI-A5).

1) Number of input frames: We first evaluate the sensitivity of our method with respect to the number of input frames, as illustrated in Figure 9. These experiments show that for the given dataset, the main skeletal structures can be recovered with as little as 16 frames. However, finer features such as bone cavities require at least 32 frames.

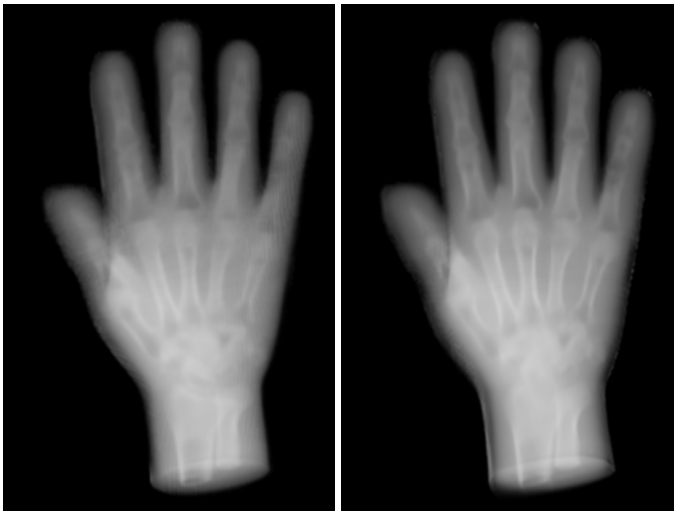


Fig. 11. Results on the synthetic data (raycasting rendering) based on varying input angular range. Left: 32 frames roughly distributed over 180 degrees; right: over 90 degrees. The rendered viewpoint falls within the range of the original 90 degrees motion, but not on an original viewpoint, leading to sharper rendering due to locally denser sampling.

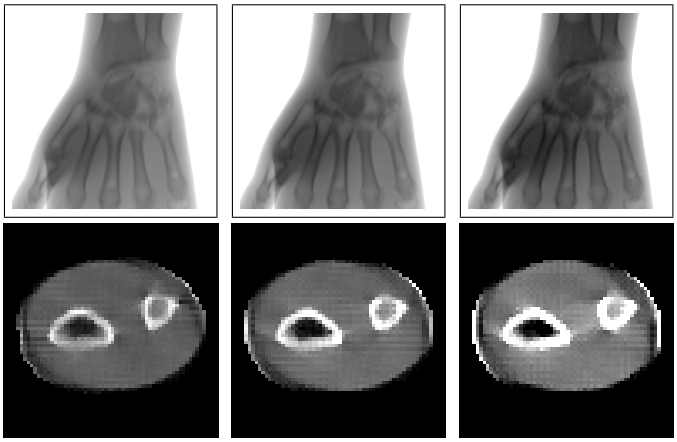


Fig. 12. Results on synthetic data over 32 frames with added image Poisson noise. Top: selected input planar X-ray image (extract), bottom: selected reconstructed slice. Left-to-right: without noise, with light noise (SNR: 10dB), with significant noise (SNR: 7dB). Light noise has little impact on the reconstruction, and stronger one only affects the reconstruction locally.

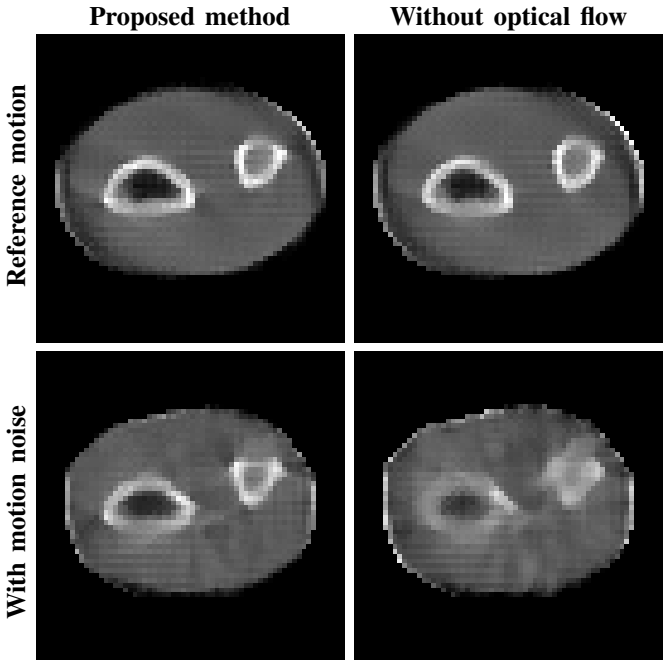


Fig. 13. Results on synthetic data over 32 frames with added sample motion noise (variance: 0.2, intensity: 5.0mm). Geometric noise is greatly reduced by optical flow correction (bottom left). When no noise is present, the flow has limited impact (top left).

2) Motion range: We reduce the motion range from 180 to 90 degrees, which clearly impacts the volumetric estimation quality, as illustrated in Figure 10. Features can nevertheless be measured within suitably selected directions. Raycasting rendering of the volume yields sharper results for poses within the original motion range, due to increased sampling density, as illustrated in Figure 11.

3) Image sensor noise: Poisson noise was artificially added to the input X-ray images at several levels to measure the impact on the reconstruction, as illustrated in Figure 12. This experiment shows that the proposed approach is capable of dealing with relatively high levels of noise. Light noise has limited impact on the reconstruction. With more potent image noise, in conditions where the skeletal structures are severely altered by image noise, the reconstruction still yields relatively good results: the main skeletal structures are well reconstructed, albeit with local noise. The Bayesian approach is key, allowing to fuse probabilities from the different images and hence reducing noise globally.

4) Geometric noise: Geometric noise was simulated in the form of Gaussian displacement in the sample pose estimation in order to estimate the impact on the reconstruction, as illustrated in Figure 13. This experiment demonstrates clearly the ability of the optical flow term to correct for the geometric noise induced by displacing the sample. Indeed, without this correction, the bone structure can barely be distinguished from the rest, whereas only minor artefacts still persist after correction. We also note that when no geometric noise is added, the optical flow correction shows virtually no impact on the reconstruction.

5) Number of iterations: In order to evaluate the convergence speed and stability of the proposed approach, we have computed the RMS error and the MI score with respect to the ground truth at every iteration, as illustrated in Figure 14. This shows that the proposed algorithm converges after about 32 iterations on the synthetic dataset. Figure 15 illustrates visually the reconstruction quality improvement over a number of iterations, demonstrating that in practice the proposed method provides usable results with as little as 8 iterations. We note that the RMS scores in Figure 14 are getting worse with the number of iterations, mostly due to edges artefacts. We also note that the MI score decreases during the first iterations before improving again after 8 iterations. We attribute this behaviour to the fact that the optical flow does not perform well on smooth data such as the volume computed with

ART used for initialisation. This might lead to incoherent motion correction until the volume has been sharpened enough through a few iterations.

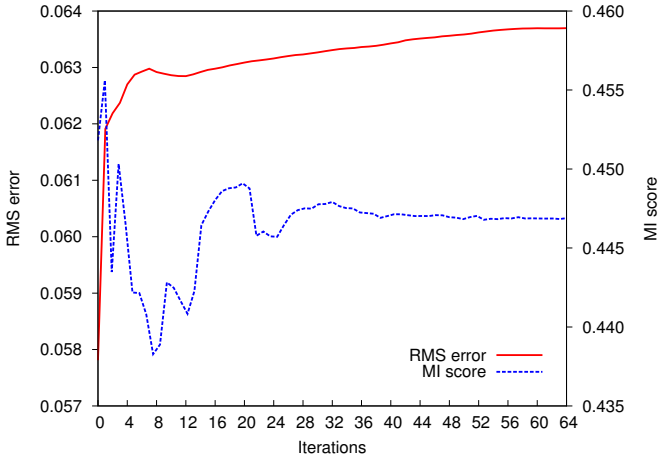
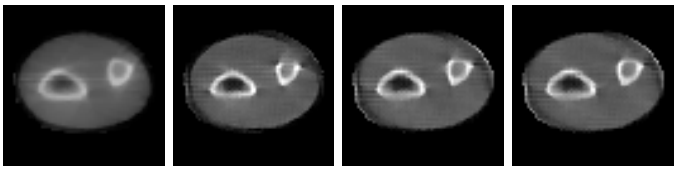


Fig. 14. Convergence of RMS error and MI score over the number of iterations. On the synthetic dataset, the algorithm exhibits first a degradation in both metrics before converging within 32 iterations. We note again that the RMS and MI at initialisation (using ART) show relatively good apparent performance due to low noise conditions.



Initialisation 1 iteration 8 iterations 32 iterations

Fig. 15. Reconstruction results after an increasing number of iterations (selected slice). The reconstruction is already much sharper after a single iteration, and improves later on, e. g. on the 'top' part of the ulna bone (top right corner).

B. In-situ forearm phantom

The volumetric results were compared to the original CT model. Unlike the synthetic experiment, the CT model and the model reconstructed with the X-ray images are in different poses since they correspond to 2 different acquisitions of the phantom, as illustrated in Figure 16. Furthermore, the energy spectrum of the CT scanner and that of the low-dose X-ray amplifier are different. Hence, the two models are first registered using multi-resolution Mutual Information (MI). The MI score is provided for quantitative comparison, being invariant to pose and attenuation spectrum.

Unlike the synthetic case, this experiment shows that the proposed method performs significantly better than ART+TVL_1 for both considered configurations (DCM and DR). In particular, the use of optical flow for motion noise compensation allows to retain a fair level of detail, while the TVL_1 norm prior constrains the ill-posed problem without excessive blurring. For example, Figure 16 demonstrates that thin features such as the bone cavities are better retained when using optical flow.

The ART+TVL_1 [23] method produces noisier results, ridden with streak artefacts due to excessive under-sampling. ART+TVL_1 does not provide any explicit mechanism for geometric noise handling, hence any regularisation scheme will be a trade-off between noise level and feature detail. On the contrary, the proposed method avoids such a trade-off by decorrelating noise level and feature details in two terms: Total Variation TVL_1 and optical flow motion correction.

Qualitatively, the raw projections captured with the DR settings (long exposure at low current) exhibit much less noise, but marginally more motion blur than the DCM configuration as illustrated in Figure 8. However, we observe that this major difference in X-ray tube configuration does not yield significantly different reconstruction results. We note that the results from the DCM dataset are slightly noisier but sharper than those from the DR dataset, while exposing the sample only to a fraction of the radiation as shown in Table I. This demonstrates the proposed method ability to deal with varied input X-ray tube characteristics.

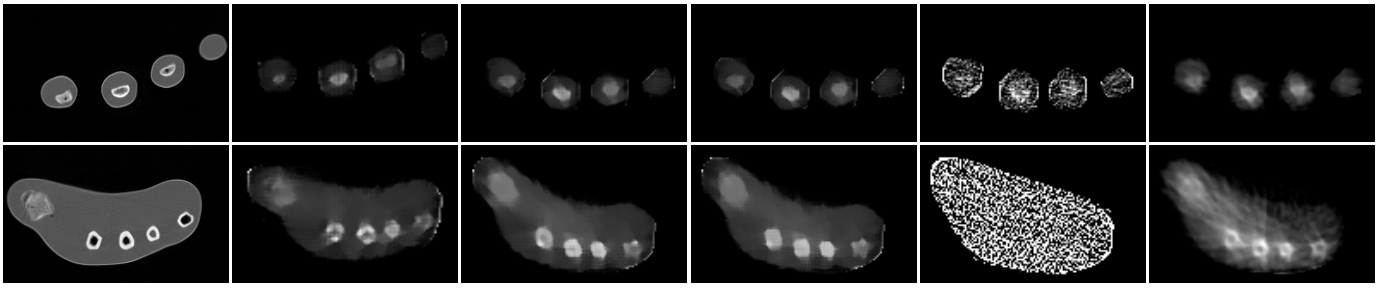
VII. Discussion

In this paper we have presented a novel imaging platform and method for dense CBCT of a sample in near-rigid motion. We have found the TVL_1 regularisation fundamental to ensure convergence of the ill-posed reconstruction problem. We also noted the benefit of the optical flow correction that allows for fine motion correction due to errors in calibration, in pose estimation, and non-rigid motion. A few limitations appeared anyway which are discussed below.

A green backdrop has been deployed to cover highly specular objects, which might not be practical in the operating theatre. While multi-view segmentation can also be performed in natural environments without specific background, segmentation might be less accurate. Similarly, occlusions will inevitably occur in a clinical scenario as they did during the experiments presented in this paper. In this case, some cameras might simply not provide any meaningful information. These two effects will inevitably reduce the quality of the surface reconstruction from video and thus the final dense reconstruction. However, our method is capable of handling a fair amount of geometric noise through the use of optical flow correction as demonstrated with the synthetic data.

We proposed to use multiple video cameras for surface motion tracking. We relied on 10 widely spread cameras, a smaller number could be used in a tighter configuration, that is closer to the gantry so as to minimise inconvenience. The number of required cameras depends on the scene complexity (especially with respect to the sample self-occlusions). We have found that 10 cameras is an upper bound for the most complex finger motions, 8 cameras for simpler hand motions, and that other scenarii would require less, with 4 video cameras being a reasonable lower bound for convex sample geometries.

In theory, given the current capture resolution (1.0mm) and X-ray exposure time in DCM configuration (7ms), the proposed approach allows to capture motion up to 0.14ms^{-1} without significant motion blur artefacts. This is equivalent to



MI score	0.061	**0.146**	0.137	*N/A (failed)*	0.094
Ground-truth CT	**Proposed method (DCM)**	**Proposed method (DR)**	**Without optical flow (DR)**	**Without TVL_1 prior (DR)**	**ART+TVL_1 [23] (DR)**

Fig. 16. Results on the forearm phantom: 2 selected slices and Mutual Information (MI) score for both DR and DCM datasets. MI registration scores are provided for quantitative comparison with differing pose/spectrum parameters. Without optical flow, artefacts are visible, for example in the bone cavities. The ART+TVL_1 method produces much noisier results with obvious streak artefacts due to under-sampling. We also observe that the DCM dataset produces slightly noisier and sharper reconstructions compared to the DR dataset with the proposed method. The MI score for the proposed method (DCM) is lower than that for ART+TVL_1 (DR), however these scores are not to be compared directly as both the dataset and the method differ in these two results.

travelling across the entire field-of-view in 2.3s. In practice, we observe that, given that the video/X-ray synchronisation is up to half-a-frame (16.7ms), the optical flow correction does compensate for motion estimation inaccuracies up to at least 0.34ms^{-1}, i. e. travelling across the field-of-view in only 0.97s.

We consider here that the capture setup is static and the calibration performed only once. If the cameras or the X-ray device were to move during the experiment, an additional relative 6D pose estimation per camera would be required. Those can be obtained relatively easily by exposing a calibration grid for a single shot. Alternatively, natural features such as surgical tools could be used for this purpose.

VIII. Conclusions and Future Work

In this paper we have presented both a novel imaging process combining X-ray and video capture of moving samples and a practical Bayesian approach to solve for a volumetric attenuation model. Our approach takes benefit of sample motion to accumulate evidence on its inner structure using motion tracking and a single static planar X-ray device. A novel generative model has been introduced to estimate a dense volumetric attenuation model. The proposed Bayesian approach optimally exploits X-ray information while enabling for acquisition noise. Our experiments show that the TVL_1 prior on the attenuation volume fundamentally contributed to convergence without excessive blurring, and that geometric noise can be effectively corrected using optical flow. This work considers rigid motion and we are currently investigating Bayesian reconstruction of non-rigidly moving samples as well as relying on a smaller number of depth cameras instead of video.

Acknowledgements

This research was partly funded by the KINOVIS (ANR-11-EQPX-0024) and CaMoPi (ANR-16-CE33-0014) projects.

References

[1] D. L. Miranda, M. J. Rainbow, J. J. Crisco, and B. C. Fleming, "Kinematic differences between optical motion capture and biplanar videoradiography during a jump-cut maneuver," *Journal of Biomechanics*, vol. 46, no. 3, pp. 567–573, 2013.

[2] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *TPAMI*, vol. 36, no. 2, pp. 346–360, Feb 2014.

[3] J. H. Siewerdsen, D. J. Moseley, S. Burch, S. K. Bisland, A. Bogaards, B. C. Wilson, and D. A. Jaffray, "Volume CT with a flat-panel detector on a mobile, isocentric C-arm: pre-clinical investigation in guidance of minimally invasive surgery," *Medical physics*, vol. 32, no. 1, pp. 241–254, 2005.

[4] P. A. Gondim Teixeira, A.-S. Formery, G. Hossu, D. Winninger, T. Batch, A. Gervaise, and A. Blum, "Evidence-based recommendations for musculoskeletal kinematic 4D-CT studies using wide area-detector scanners: a phantom study with cadaveric correlation," *European Radiology*, vol. 27, no. 2, pp. 437–446, Feb 2017.

[5] P. A. Gondim Teixeira, A. Gervaise, M. Louis, A. Raymond, A.-S. Formery, S. Lecocq, and A. Blum, "Musculoskeletal wide-detector CT kinematic evaluation: From motion to image," *Semin Musculoskelet Radiol*, vol. 19, pp. 456–462, 2015.

[6] Q. Zhang, Y.-C. Hu, F. Liu, K. Goodman, K. E. Rosenzweig, K. Goodman, and G. S. Mageras, "Correction of motion artifacts in cone-beam CT using a patient-specific respiratory motion model," *Medical Physics*, vol. 37, no. 6, pp. 2901–2909, 2010.

[7] J. Ehrhardt, R. Werner, T. Frenzel, D. Säring, W. Lu, D. Low, and H. Handels, "Reconstruction of 4D-CT data sets acquired during free breathing for the analysis of respiratory motion," in *Medical Imaging 2006: Image Processing*, vol. 6144. International Society for Optics and Photonics, 2006, p. 614414.

[8] J. Fotouhi, B. Fuerst, W. Wein, and N. Navab, "Can real-time RGBD enhance intraoperative cone-beam CT?" *IJCARS*, vol. 12, no. 7, pp. 1211–1219, 2017.

[9] B. Bier, N. Ravikumar, M. Unberath, M. Levenston, G. Gold, R. Fahrig, and A. Maier, "Range imaging for motion compensation in C-arm cone-beam CT of knees under weight-bearing conditions," *Journal of Imaging*, vol. 4, no. 1, p. 13, 2018.

[10] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *Journal of the Optical Society of America (JOSA) A*, vol. 1, no. 6, pp. 612–619, 1984.

[11] T. Q. Bang and I. Jeon, "CT reconstruction from a limited number of X-ray projections," *World Academy of Science, Engineering and Technology*, vol. 5, no. 10, pp. 488–490, 2011.

[12] G. Yang, J. H. Hipwell, D. J. Hawkes, and S. R. Arridge, "A nonlinear least squares method for solving the joint reconstruction and registration problem in digital breast tomosynthesis," in *MIUA*, 2012, pp. 87–92.

[13] I. Reiser, J. Bian, R. M. Nishikawa, E. Y. Sidky, and X. Pan, "Comparison of reconstruction algorithms for digital breast tomosynthesis," in *The 9th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, Jul. 2007.

[14] A. Baudoin, W. Skalli, J. A. de Guise, and D. Mitton, "Parametric subject-specific model for in vivo 3D reconstruction using bi-planar X-rays: application to the upper femoral extremity," *Medical & Biological Engineering & Computing*, vol. 46, pp. 799–805, 2008.
[15] S. Benameur, M. Mignotte, H. Labelle, and J. A. De Guise, "A hierarchical statistical modeling approach for the unsupervised 3-D biplanar reconstruction of the scoliotic spine," *TBME*, vol. 52, no. 12, pp. 2041–2057, 2005.
[16] B.-M. You, P. Siy, W. Anderst, and S. Tashman, "In vivo measurement of 3-D skeletal kinematics from sequences of biplane radiographs: Application to knee kinematics," *TMI*, vol. 20, no. 6, pp. 514–525, June 2001.
[17] E. L. Brainerd, D. B. Baier, S. M. Gatesy, T. L. Hedrick, K. A. Metzger, S. L. Gilbert, and J. J. Crisco, "X-ray reconstruction of moving morphology (XROMM): precision, accuracy and applications in comparative biomechanics research," *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, vol. 313, no. 5, pp. 262–279, 2010.
[18] A. Abourachid, R. Hackert, M. Herbin, P. A. Libourel, F. Lambert, H. Gioanni, P. Provini, P. Blazevic, and V. Hugel, "Bird terrestrial locomotion as revealed by 3D kinematics," *Journal of Zoology*, vol. 114, no. 6, pp. 360–368, 2011.
[19] D. D. Cox, A. M. Papanastassiou, D. Oreper, B. B. Andken, and J. J. DiCarlo, "High-resolution three-dimensional microelectrode brain mapping using stereo microfocal X-ray imaging," *Journal of Neurophysiology*, vol. 100, pp. 2966–2976, 2008.
[20] M. Yam, M. Brady, R. Highnam, C. Behrenbruch, R. English, and Y. Kita, "Three-dimensional reconstruction of microcalcification clusters from two mammographic views," *TMI*, vol. 20, no. 6, pp. 479–489, June 2001.
[21] M. Hoshino, K. Uesugi, J. Pearson, T. Sonobe, M. Shirai, and N. Yagi, "Development of an X-ray real-time stereo imaging technique using synchrotron radiation," *Journal of synchrotron radiation*, vol. 18, no. 4, pp. 569–574, 2011.
[22] E. Y. Sidky, C.-M. Kao, and X. Pan, "Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT," *Journal of X-ray Science and Technology*, vol. 14, no. 2, pp. 119–139, 2006.
[23] J. Pansiot, L. Reveret, and E. Boyer, "Combined visible and X-ray 3D imaging," in *MIUA*, London, Jul. 2014, pp. 13–18.
[24] X. Wang, S. Habert, M. Ma, C.-H. Huang, P. Fallavollita, and N. Navab, "Precise 3D/2D calibration between a RGB-D sensor and a C-arm fluoroscope," *IJCARS*, vol. 11, no. 8, pp. 1385–1395, 2016.
[25] N. Navab, A. Bani-Kashemi, and M. Mitschke, "Merging visible and invisible: two camera-augmented mobile C-arm (CAMC) applications," in *IWAR*, 1999, pp. 134–141.
[26] S. Habert, J. Gardiazabal, P. Fallavollita, and N. Navab, "RGBDX: First design and experimental validation of a mirror-based RGBD X-ray imaging system," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, Sept 2015, pp. 13–18.
[27] S. Reaungamornrat, Y. Otake, A. Uneri, S. Schafer, D. J. Mirota, S. Nithiananthan, J. W. Stayman, G. Kleinszig, A. J. Khanna, R. H. Taylor, and J. H. Siewerdsen, "An on-board surgical tracking and video augmentation system for C-arm image guidance," *IJCARS*, vol. 7, no. 5, pp. 647–665, Sep 2012.
[28] J. Pansiot and E. Boyer, "3D imaging from video and planar radiography," in *MICCAI*, vol. 9902. Athens: Springer, Oct. 2016, pp. 450–457.
[29] J. Zheng, S. S. Saquib, K. Sauer, and C. A. Bouman, "Parallelizable Bayesian tomography algorithms with rapid, guaranteed convergence," *TMI*, vol. 9, no. 10, pp. 1745–1759, 2000.
[30] P. J. Green, "Bayesian reconstructions from emission tomography data using a modified EM algorithm," *TMI*, vol. 9, no. 1, pp. 84–93, 1990.
[31] T. Schüle, C. Schnörr, S. Weber, and J. Hornegger, "Discrete tomography by convex–concave regularization and DC programming," *Discrete Applied Mathematics*, vol. 151, no. 1, pp. 229–243, 2005.
[32] J. Pansiot and E. Boyer, "CT from motion: Volumetric capture of moving shapes with X-rays and videos," in *BMVC*, London, Sep. 2017.
[33] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *IJCV*, vol. 61, no. 3, pp. 211–231, 2005.
[34] J.-S. Franco and E. Boyer, "Exact polyhedral visual hulls," in *BMVC*, 2003, pp. 329–338.
[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.
[36] I. S. Duff, R. G. Grimes, and J. G. Lewis, "Sparse matrix test problems," *ACM Trans. Math. Softw.*, vol. 15, no. 1, pp. 1–14, Mar. 1989.
[37] Z. Zhang, "A flexible new technique for camera calibration," *TPAMI*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.
[38] S. Stegmaier, M. Strengert, T. Klein, and T. Ertl, "A simple and flexible volume rendering framework for graphics-hardware-based raycasting," in *Eurographics/IEEE VGTC conference on Volume Graphics*, 2005, pp. 187–195.