Neural Network Kalman Filtering for 3-D Object Tracking From Linear Array Ultrasound Data

Arttu Arjas, Erwin J. Alles, Efthymios Maneas, Simon Arridge, Adrien Desjardins, Mikko J. Sillanpää, and Andreas Hauptmann, *Member, IEEE*

***Abstract*—Many interventional surgical procedures rely on medical imaging to visualize and track instruments. Such imaging methods not only need to be real time capable but also provide accurate and robust positional information. In ultrasound (US) applications, typically, only 2-D data from a linear array are available, and as such, obtaining accurate positional estimation in three dimensions is nontrivial. In this work, we first train a neural network, using realistic synthetic training data, to estimate the out-of-plane offset of an object with the associated axial aberration in the reconstructed US image. The obtained estimate is then combined with a Kalman filtering approach that utilizes positioning estimates obtained in previous time frames to improve localization robustness and reduce the impact of measurement noise. The accuracy of the proposed method is evaluated using simulations, and its practical applicability is demonstrated on experimental data obtained using a novel optical US imaging setup. Accurate and robust positional information is provided in real time. Axial and lateral coordinates for out-of-plane objects are estimated with a mean error of 0.1 mm for simulated data and a mean error of 0.2 mm for experimental data. The 3-D localization is most accurate for elevational distances larger than 1 mm, with a maximum distance of 6 mm considered for a 25-mm aperture.**

***Index Terms*—Kalman filtering, neural networks, object tracking, optical ultrasound (OpUS), out-of-plane artifacts.**

Manuscript received March 18, 2022; accepted March 18, 2022. Date of publication March 24, 2022; date of current version April 26, 2022. This work was supported in part by the Academy of Finland under Project 336796, Project 326291, and Project 338408; in part by the CMIC-EPSRC Platform under Grant EP/M020533/1; in part by the Wellcome Trust under Grant 203145Z/16/Z; in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant NS/A000050/1; and in part by the Rosetrees Trust under Grant PGS19-2/10006. *(Corresponding author: Arttu Arjas.)*

Arttu Arjas and Mikko J. Sillanpää are with the Research Unit of Mathematical Sciences, University of Oulu, 90014 Oulu, Finland (e-mail: arttu.arjas@oulu.fi).

Erwin J. Alles, Efthymios Maneas, and Adrien Desjardins are with the Wellcome/EPSRC Centre for Interventional and Surgical Sciences and the Department of Medical Physics and Biomedical Engineering, University College London, London WC1E 6BT, U.K.

Simon Arridge is with the Department of Computer Science, University College London, London WC1E 6BT, U.K.

Andreas Hauptmann is with the Research Unit of Mathematical Sciences, University of Oulu, 90014 Oulu, Finland, and also with the Department of Computer Science, University College London, London WC1E 6BT, U.K.

This article has supplementary downloadable material available at https://doi.org/10.1109/TUFFC.2022.3162097, provided by the authors.

Digital Object Identifier 10.1109/TUFFC.2022.3162097

I. Introduction

Tracking and localization of point-like objects are crucial to a large variety of medical applications in ultrasound (US) imaging, such as tracking of microbubbles for super-resolution US imaging [1], [2] or US-guided placement of fiducial markers for radiotherapy [3]. In addition, tracking of surgical tools (such as needles and catheters) is essential during minimally invasive procedures [4]–[6], as when placed inaccurately, these devices may cause trauma by damaging tissue or deliver ineffective treatment to the wrong location [6], [7]. As such, US is frequently used for guidance through imaging, but accurate localization in a 3-D target domain remains challenging. This is primarily caused by the nature of data acquisition using linear arrays, which assumes that all signals originate from within the image plane and, thus, only a 2-D B-mode image of the image plane is formed. We refer to this obtained 2-D image as the US image and assume that it consists of the reconstructed point, or point-like, source corresponding to the object we aim to track accurately. However, if this point source is located out-of-plane, it will primarily show as aberration in the reconstructed image domain. In addition, one may misinterpret features, such as a needle shaft as the tip [8]. Thus, the problem to provide an accurate positional estimate in 3-D from only 2-D US images is consequently a notoriously difficult task without any auxiliary information [9] and is a field of active research [10], [11]. Early approach used speckle information to estimate out-of-plane displacements [12], [13]. Another possibility for instrumented US tracking of needles was proposed by Xia *et al.* [14], [15], who designed a custom-made imaging probe consisting of a central array for conventional imaging and two side arrays for 3-D tracking [15]. Alternatively, one may approach the needle tracking problem in full 3-D to obtain accurate positional estimates [16].

In this work, we propose an alternative, real-time capable, method of performing 3-D tracking without the need for custom-made probes and using a single set of measurements per time step from a linear array. In the following, we assume a point source model for the tracked object. For the estimation of lateral and axial positions, we examine high-intensity pixels in the reconstructed 2-D US image, similar to [16], where tracking was performed in full 3-D. For the estimation of the elevational direction or out-of-plane distance of the point source to the image plane, we use a machine learning approach. In particular, significant markers are extracted from the measured time series and a neural network is trained using synthetic data modeled for a prototype optical ultrasound (OpUS) imaging setup [17] to predict out-of-plane distance and associated aberration in the axial position in the reconstructed 2-D US image. The markers used for the neural network are summary statistics extracted from the measurement data and correlated with the offset to establish nonlinear mapping between the two.

In addition, we assume regularity in the temporal evolution of the object location to improve robustness and reduce uncertainty in the estimation compared to independently analyzing subsequent images. The regularity assumption mimics conditions encountered in clinical practice, where the objects, such as needle tips and microbubbles, are expected to follow a smooth trajectory without rapid jumps or jitter during insertions into soft tissue. This can be incorporated, while retaining computational efficiency, by a Kalman filter [18], [19], which is a flexible method for estimating the state (position, velocity, and so on) of a dynamic system. It has been classically utilized in engineering applications such as target tracking and navigation but has also been extensively used in inverse problems and medical imaging [15], [19]–[23]. The underlying idea of Kalman filtering is to update the estimate of the state at time step $k + 1$ each time new data become available as opposed to smoothing, where the whole trajectory from time step 1 to k is updated as well. It has the appealing property, as opposed to estimating the full posterior of all states, that the problem does not become intractable as the number of data points increases.

We note that Kalman filtering has earlier been utilized for a needle tracking problem in 3-D [16] as well as for microbubble tracking in 2-D [24]. Takeshima *et al.* [25] tracked a wire tip using the Kalman filter and perform an elevational position estimation from geometric markers. In this work, we approach the problem in 3-D with only a single set of measurements (per time point) from a linear array and combine it with a neural network in order to obtain reliable estimates on the elevation to correct the axial position in the 2-D US image $\hat{x}$. We evaluate the proposed method by tracking a point source for simulated OpUS data and object trajectories with changing elevation. Robustness is evaluated with respect to increased noise in the measurement data and accuracy compared to positional estimation using only the pixel with maximum intensity (MI) in the OpUS image. Finally, we evaluate the method on experimental OpUS measurements.

II. Kalman Filter for Object Tracking and Out-of-Plane Correction

A. Image Formation and Object Tracking

In the following, we specifically consider a custom setup and simulation framework for a freehand optical US imaging system, that is, the US signal generation is modeled as pulse-echo imaging, where each source along the linear aperture emits a pressure wave, which reflects off the point scatterer and is detected by a single fiber-optic detector placed right next to the imaging aperture [17]. We note that the tracking framework here can be generalized to any linear US array, where every source element also acts as a detector. Moreover, we consider in the following the tracking problem in a 3-D space, that is, we aim to determine the 3-D coordinate x of the point source. Given the recorded radio frequency (RF) time series p, the reconstruction of the 2-D in-plane US image $\hat{x}$ is performed using a basic delay-and-sum algorithm (equivalent to dynamic focusing) [26].

When the location of the point source is in-plane, then the reconstructed image $\hat{x}$ can be directly used to estimate the location x reliably. On the other hand, if x is out-of-plane, then the 2-D reconstruction will lead to a distorted image, in the sense that the reconstructed axial position is located deeper in the target than the correct position [27]. The aberration occurs because the time of flight is larger for objects that are positioned out-of-plane due to the imaging geometry (see Fig. 1). Thus, this axial aberration needs to be detected and processed to simultaneously provide an estimate of the elevational distance between the point target and the image plane, and the corresponding coordinates projected onto the image plane.

B. Object Tracking for In-Plane Objects

Most tracking applications primarily assume that the object of interest is in-plane and features extracted from the reconstructed image $\hat{x}$ are good indicators of the actual position x. Thus, the majority of tracking algorithms are based on intensity values in the B-mode images for point marker tracking [28], [29] in combination with various image registration approaches [30]–[32]. More recent developments make use of deep learning techniques to estimate the coordinates of objects directly from the measured time series p [33]–[35]. Nevertheless, there is no clear gold standard to perform object tracking, as the particular approach depends heavily on the application and practical need [9].

In this study, we are concentrating on single object tracking and use the MI estimate for comparison and reference since it provides highly efficient and accurate estimates under ideal assumptions, i.e., high signal-to-noise ratio without any elevation. Consequently, we will also design our tracking model in the following by using the pixels in the US image with the highest intensity for the estimation of axial and lateral positions in the filtering process.

C. Kalman Filtering

Kalman filtering, a class of Bayesian filtering, is especially effective in situations where the data stream is over time and one must update the state given the new data and the history of the system; as such, it is ideally suited to robustly perform the object tracking considered in this study. Specifically, Kalman filtering [18] consists of closed-form update formulas for a linear Gaussian filtering problem, which will be discussed next. The estimation of axial and lateral coordinates is similar to the approaches suggested in [16] and [24], and we will then continue to extend our model to incorporate elevation and a correction of estimated axial coordinates.

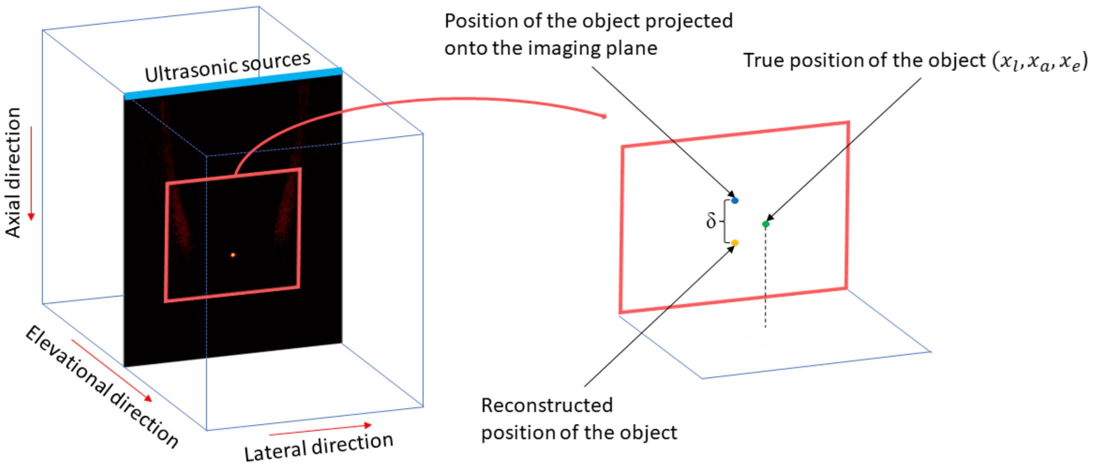

Fig. 1. Illustration of the measurement geometry. If the point source object is located out-of-plane (right side), then the object appears distorted in the reconstructed US image, i.e., too low by δ in the axial direction. This is why the axial position needs to be corrected.

1) Lateral and Axial Coordinates: The estimation of lateral and axial coordinates and corresponding velocities $\boldsymbol{x}_k = (x_{lk}\ x_{ak}\ v_{lk}\ v_{ak})^\mathsf{T}$ at time step k is based on the locations of the highest absolute intensity pixels in the image. We assume that these locations are spread around the location of the object. While the velocity of the object is not the main interest, it is introduced as an auxiliary variable to help in predicting the motion, as will be described in the following. We denote by $\boldsymbol{y}_{lk} \in \mathbb{R}^n$ the n lateral and by $\boldsymbol{y}_{ak} \in \mathbb{R}^n$ the n axial highest intensity locations and let $\boldsymbol{y}_k = (\boldsymbol{y}_{lk}^\mathsf{T}\ \boldsymbol{y}_{ak}^\mathsf{T})^\mathsf{T}$. We then build a model

$$\boldsymbol{y}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{r}_k \tag{1}$$

where the matrix

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \vdots & \vdots \\ 0 & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{2n\times 4} \tag{2}$$

associates given high-intensity locations in $\boldsymbol{y}_k$ with the actual coordinates of the object in $\boldsymbol{x}_k$.

Furthermore, we assume that the noise in the observed locations is normally distributed $\boldsymbol{r}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_k)$, where

$$\boldsymbol{R}_k = \begin{pmatrix} s_{lk}^2 \boldsymbol{I}_{n\times n} & \boldsymbol{0} \\ \boldsymbol{0} & s_{ak}^2 \boldsymbol{I}_{n\times n} \end{pmatrix} \tag{3}$$

with s_{lk}^2 being the sample variance of $\boldsymbol{y}_{lk}$ and s_{ak}^2 the sample variance of $\boldsymbol{y}_{ak}$. In this way, uncertainty is naturally incorporated into the model as how spread out the high-intensity locations are.

We model the motion of the object with a constant velocity model [36]

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{G}\boldsymbol{c}_k \tag{4}$$

where $\boldsymbol{c}_k \sim \mathcal{N}(\boldsymbol{0}, \operatorname{diag}(\sigma_l^2, \sigma_a^2))$ is assumed to be a random acceleration component and σ_l^2 and σ_a^2 are lateral and axial process noise variances, respectively. The matrices $\boldsymbol{A}$ and $\boldsymbol{G}$ are defined as [19]

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

$$\boldsymbol{G} = \begin{pmatrix} \frac{1}{2}\Delta t^2 & 0 \\ 0 & \frac{1}{2}\Delta t^2 \\ \Delta t & 0 \\ 0 & \Delta t \end{pmatrix} \tag{6}$$

where Δt is the time between subsequent observations.

The velocity and acceleration of the object at time step $k-1$ is used to give an accurate prediction of the position at time step k. Writing (4) explicitly for the positional variables only, we obtain

$$\begin{aligned} x_{lk} &= x_{l(k-1)} + \Delta t v_{l(k-1)} + \frac{1}{2}\Delta t^2 c_{lk} \\ x_{ak} &= x_{a(k-1)} + \Delta t v_{a(k-1)} + \frac{1}{2}\Delta t^2 c_{ak}. \end{aligned} \tag{7}$$

This means that we assume the position at time step k to be close to the position at time step $k-1$ plus the displacement given by the time between subsequent observations, velocity, and acceleration.

2) Out-of-Plane Offset and Axial Aberration: In case there is an offset between the imaging plane and the object, we observe aberration in the reconstructed axial coordinate due to the geometry of the imaging problem (see Fig. 1). In this case, location estimation based on only the high-intensity pixels would result in a biased estimate of the axial coordinate. Instead, we use information in the measurement data p to estimate the offset and axial aberration and use the knowledge to correct the estimate of the axial coordinate. To do this, we train a neural network with training data obtained from an OpUS simulator. Details on the neural network will be provided in Section II-D1 and on the simulator in Section III-B. We note that instead of a neural network, other sufficiently expressive nonlinear prediction models could be used.

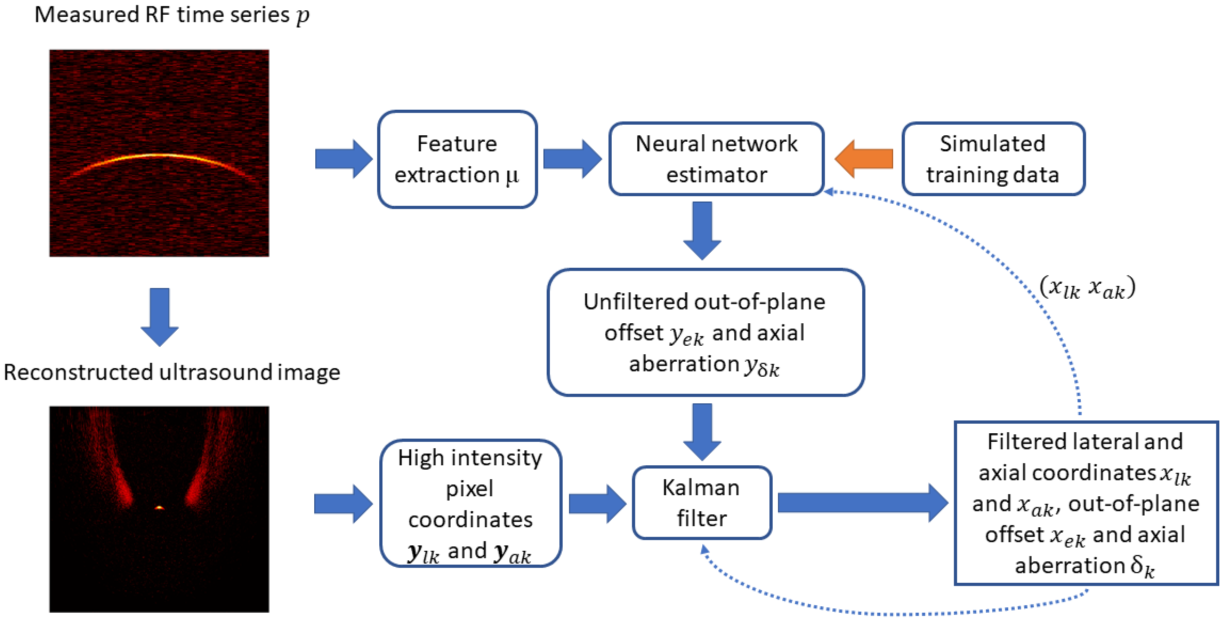
Measured RF time series p
Reconstructed ultrasound image
Feature extraction μ
Neural network estimator
Simulated training data
Unfiltered out-of-plane offset y_{ek} and axial aberration $y_{\delta k}$
$(x_{lk}\ x_{ak})$
High intensity pixel coordinates $\boldsymbol{y}_{lk}$ and $\boldsymbol{y}_{ak}$
Kalman filter
Filtered lateral and axial coordinates x_{lk} and x_{ak}, out-of-plane offset x_{ek} and axial aberration δ_k

Fig. 2. Flowchart for the NNK filtered tracking. We extract a mean value μ from the highest amplitude entries in the measured time series p and use the highest intensity pixels in the 2-D US image. A previously trained (orange arrow) neural network estimator then uses the last state estimates of lateral x_{lk} and axial x_{ak} coordinates together with μ to estimate offset and axial aberration. The next state is then updated via Kalman filtering to provide a robust positional estimation.

The filtering model is then extended to include the unfiltered out-of-plane offset and axial aberration, denoted as y_{ek} and $y_{\delta k}$, that are received as output from the neural network. We can then define $\boldsymbol{y}_k^* = (\boldsymbol{y}_k^\mathsf{T}\ y_{ek}\ y_{\delta k})^\mathsf{T}$. Filtered out-of-plane offset and axial aberration and their velocities, denoted as x_{ek}, δ_k, v_{ek}, and $v_{\delta k}$ are included in the state vector $\boldsymbol{x}_k^* = (\boldsymbol{x}_k^\mathsf{T}\ x_{ek}\ \delta_k\ v_{ek}\ v_{\delta k})^\mathsf{T}$, The extended model is

$$\begin{aligned}\boldsymbol{y}_k^* &= \boldsymbol{H}^*\boldsymbol{x}_k^* + \boldsymbol{r}_k^*\\ \boldsymbol{x}_k^* &= \boldsymbol{A}^*\boldsymbol{x}_{k-1}^* + \boldsymbol{G}^*\boldsymbol{c}_k^*\end{aligned} \tag{8}$$

where $\boldsymbol{r}_k^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_k^*)$, $\boldsymbol{c}_k^* \sim \mathcal{N}(\boldsymbol{0}, \text{diag}(\sigma_l^2, \sigma_a^2, \sigma_e^2, \sigma_\delta^2))$

$$\boldsymbol{H}^* = \begin{pmatrix}\boldsymbol{H} & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{H}}\end{pmatrix} \tag{9}$$

with

$$\tilde{\boldsymbol{H}} = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{pmatrix} \tag{10}$$

$$\boldsymbol{A}^* = \begin{pmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}\end{pmatrix} \tag{11}$$

$$\boldsymbol{G}^* = \begin{pmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}\end{pmatrix} \tag{12}$$

and σ_e^2 and σ_δ^2 are the process noise variances of the out-of-plane offset and axial aberration components, respectively. Finally, we let

$$\boldsymbol{R}_k^* = \begin{pmatrix}\boldsymbol{R}_k & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{R}}_k\end{pmatrix} \tag{13}$$

where

$$\tilde{\boldsymbol{R}}_k = \max\left(s_{lk}^2, s_{ak}^2\right)\boldsymbol{I}_{2\times 2}. \tag{14}$$

3) State Estimation:

As stated earlier, a major benefit arising from linearity and Gaussianity of the filtering models are the closed-form update formulas for mean $\boldsymbol{m}_k$ and covariance $\boldsymbol{P}_k$ of the state. At $k = 0$, we assume $\boldsymbol{x}_0^* \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{P}_0)$, where $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{P}_0 = 15\boldsymbol{I}$ to serve as an uninformative prior. Note that the mean $\boldsymbol{m}_k$ corresponds to the estimated coordinates for $\boldsymbol{x}_k^*$ and $\boldsymbol{P}_k$ is the corresponding covariance matrix. At every round, the prior predictions for mean and covariance are updated recursively as

$$\begin{aligned}\boldsymbol{m}_k^{\text{pr}} &= \boldsymbol{A}^*\boldsymbol{m}_{k-1}\\ \boldsymbol{P}_k^{\text{pr}} &= \boldsymbol{A}^*\boldsymbol{P}_{k-1}\boldsymbol{A}^{*\mathsf{T}} + \boldsymbol{Q}^*\end{aligned} \tag{15}$$

where $\boldsymbol{Q}^* = \boldsymbol{G}^*\text{diag}(\sigma_l^2, \sigma_a^2, \sigma_e^2, \sigma_\delta^2)\boldsymbol{G}^{*\mathsf{T}}$. We then evaluate the neural network as described in Section II-D to provide the estimates of offset and axial aberration in $\boldsymbol{y}_k^*$, and then, the Kalman update can be performed by

$$\begin{aligned}\boldsymbol{u}_k &= \boldsymbol{y}_k^* - \boldsymbol{H}^*\boldsymbol{m}_k^{\text{pr}}, && \text{(Prediction residual)}\\ \boldsymbol{S}_k &= \boldsymbol{H}^*\boldsymbol{P}_k^{\text{pr}}\boldsymbol{H}^{*\mathsf{T}} + \boldsymbol{R}_k^*, && \text{(Measurement covariance update)}\\ \boldsymbol{K}_k &= \boldsymbol{P}_k^{\text{pr}}\boldsymbol{H}^{*\mathsf{T}}\boldsymbol{S}_k^{-1}, && \text{(Gain update)}\\ \boldsymbol{m}_k &= \boldsymbol{m}_k^{\text{pr}} + \boldsymbol{K}_k\boldsymbol{u}_k, && \text{(State update)}\\ \boldsymbol{P}_k &= \boldsymbol{P}_k^{\text{pr}} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^\mathsf{T}, && \text{(State covariance update).}\end{aligned} \tag{16}$$

Estimates of all coordinates are then given by $\boldsymbol{m}_k$. Their variances can be found in the diagonal of $\boldsymbol{P}_k$ and could be used for uncertainty quantification of the Kalman updates. The estimated axial aberration is then subtracted from the estimated axial coordinate to yield an estimate for the actual axial coordinate as

$$m_{ak}^* = m_{ak} - m_{\delta k}. \tag{17}$$

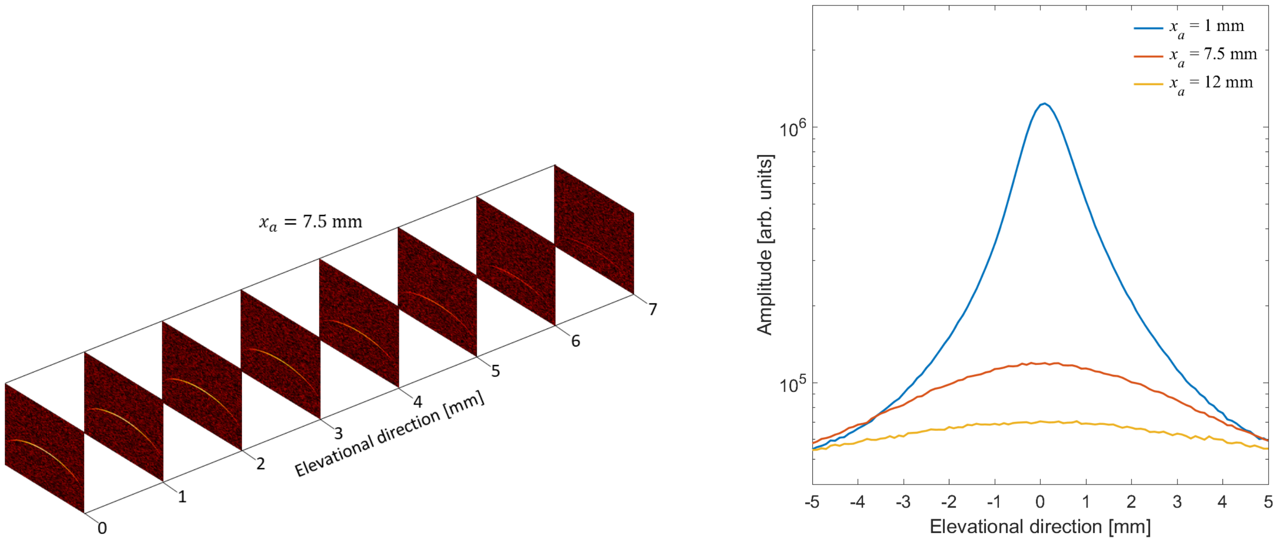

Fig. 3. RF time series (left) showing the decay in amplitude as the distance from imaging plane increases. The decay is also shown on the right for different axial depths. In general, the rate of decay decreases with increased depth. The lateral coordinate x_l was set to 0 in these tests.

D. Application to Object Tracking

We apply the Kalman filtering method to the task of object tracking modeled from OpUS image reconstructions and measurement data. After generating suitable training data and training the neural network, tracking can be performed as outlined in the following. A diagram of the workflow is shown in Fig. 2.

1) Neural Network (Training Data and Architecture): We use an OpUS simulator [37], [38] to generate training data for the neural network. First, we define a uniform 20 × 20 × 20 grid of coordinates. The grid has bounds ±12 mm in the lateral direction, 0.5–14.5 mm in the axial direction, and ±10 mm in the elevational direction. Two additional grid points were placed in-plane (zero in elevational direction). In each grid point, we simulate measurement data with coordinate values equal to the grid point.

To obtain a marker for the offset estimation, we note that US source elements typically emit near-omnidirectional pressure fields within the image plane but are usually designed to emit highly directional fields in the out-of-plane direction. This is achieved through a combination of eccentric element geometries and acoustic lenses [27]. As a result, the amplitude of pulse-echo signals from point objects depends strongly on the elevational (out-of-plane) position and generally reduces with increasing elevational offset. The shape of the decay also depends on the position of the object. In short, the out-of-plane amplitude decay decreases as the axial depth increases, as shown in Fig. 3. This is why both the lateral and axial coordinates are used as inputs for the neural network. Thus, the pulse-echo signal strength across the aperture can be used as an effective marker of the elevational position. To exploit this, we use the RF time series p to compute the mean absolute value μ of those time series samples belonging to either highest or lowest 1% of a Gaussian defined by the mean and variance of p. This way most of the purely noisy part of the data is ignored. We also compute the distance between the mean of apparent (reconstructed) axial coordinates of $n = 15$ highest intensity pixels and the real axial coordinate used to simulate the data. This distance reflects the axial aberration that needs to be corrected for.

A neural network Λ_θ with parameters θ is then trained to map lateral and axial coordinates, and mean absolute value of high amplitude entries in measured time series, denoted as $\boldsymbol{u} = (x_a \; x_l \; \mu)^\mathsf{T}$, to a prediction of unfiltered out-of-plane offset and axial aberration $\boldsymbol{w} = (y_e \; y_\delta)^\mathsf{T}$. Since the simulator output is almost symmetric with positive and negative offsets, we train the network with absolute offset values ≥ 0. This means that we can only estimate the magnitude of the offset, not the direction. The network chosen is a standard multilayer perceptron [39] with two hidden layers and 20 nodes in each layer. Each hidden layer has a sigmoid activation function, whereas for the output layer, the activation function is linear. The network is trained by finding a set of parameters θ^* such that the mean squared error between the neural network output and the ground truth is minimized, i.e.,

$$\theta^* = \operatorname*{argmin}_{\theta} \sum_{i=1}^{M} \|\Lambda_\theta(\boldsymbol{u}_i) - \boldsymbol{w}_i\|_2^2 \tag{18}$$

where M is the size and i is an index over the training data. We used the Levenberg–Marquardt algorithm, Marquardt [40] and Levenberg [41], to train the neural network. The dampening parameter was set to the default value of 10^{-3}. The optimization stopped when the validation performance did not improve in six epochs in a row or the relative norm of the gradient of the minimized function was smaller than 10^{-7}.

2) Tracking: We track the point source from a sequence of optical US image reconstructions. At $k = 0$, we set $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{P}_0 = 15\boldsymbol{I}$. We find the coordinates of $n = 15$ highest intensity pixels and use the neural network to estimate the unfiltered out-of-plane offset and axial aberration. Since the neural network input contains the lateral and axial positions of the point source, we use the estimate from the previous time step. If the motion of the object is somewhat regular, this does not have a big impact on the estimation accuracy. The coordinate estimates are then updated with Kalman filter update formulas [see (16)]. An estimate of the true axial coordinate of the object is then obtained by subtracting the axial aberration estimate from the apparent axial coordinate obtained directly from the 2-D US image. An illustration of the full tracking workflow is shown in Fig. 2 and summarized as pseudocode in Algorithm 1.

Algorithm 1 NNK Filtered Tracking

```
Initialisations: $\boldsymbol{m}_0 = \mathbf{0}$, $\boldsymbol{P}_0 = 15\boldsymbol{I}$
function NNK(Inputs: process noise variances $\sigma_l^2$, $\sigma_a^2$, $\sigma_e^2$ and $\sigma_\delta^2$)
    $k \leftarrow 1$
    while new data acquired do
        Update mean $\boldsymbol{m}_k^{\text{pr}}$ and covariance $\boldsymbol{P}_k^{\text{pr}}$ by Eq. (15)
        Compute marker $\mu_k$ and high intensity pixel locations $\boldsymbol{y}_{lk}$ and $\boldsymbol{y}_{ak}$
        $\boldsymbol{u} \leftarrow (m_{l(k-1)}, m^*_{a(k-1)}, \mu_k)$
        $(y_{ek}, y_{\delta k}) \leftarrow \Lambda_\theta(\boldsymbol{u})$
        Perform Kalman update with (16)
        Perform axial aberration correction with (17)
        Display image overlaid with coordinate estimates
        $k \leftarrow k + 1$
    end while
end function
```

III. Optical US and Experiments

A. Experimental Setup

The experimental validation of the method was performed using a custom OpUS imaging system comprising a handheld imaging probe. We have chosen the OpUS imaging system for this study due to three main advantages: it offers direct access to the RF data, it was previously accurately characterized in-house, and the system can be accurately and highly efficiently modeled numerically—thus making it an ideal fit for this study. This system, which was described in full in [17], uses scanning optics to couple excitation light sequentially into the proximal ends of 64 optical fibers arranged in a linear array. This light is delivered to an optically absorbing coating deposited at the distal ends, where it is converted into divergent US waves via the photoacoustic effect [42]. Thus, an OpUS source aperture is rapidly scanned to enable video-rate and real-time imaging in a 2-D imaging plane. Backscattered US waves are detected using a single fiber-optic US detector comprising an optically resonant plano-concave Fabry–Pérot cavity [43], with an lateral extent of 25 mm.

B. Simulated Data

A highly efficient and accurate simulator of the OpUS imaging setup, as previously described in [37] and [38] and based on the FOCUS US simulator [44], [45], was used to evaluate the performance of our method with synthetic data examples produced with the OpUS simulator. In total, four synthetic data sets were generated to test different properties of the tracking method. Noise amplitude was computed such that SNR for in-plane locations was 6.5 dB and decreasing SNR with elevational distance, due to decreasing signal strength. The first data set (Experiment 1) comprises 101 time points and a smooth, curved object trajectory with linear motion at constant velocity in the elevational direction to test the overall performance [see Fig. 4]. We remind that our proposed method extends the MI estimation with Kalman filtering and incorporation of elevational offset estimation and axial aberration correction. Thus, Experiment 1 shows the importance of the aberration correction. We then examine other factors, such as noise in the second data set (Experiment 2), which is the same as the first one, but with tenfold noise in every tenth measured time series to investigate the robustness of the method. The third data set (Experiment 3) follows also the same axial–lateral trajectory as Experiment 1, but the object is positioned in-plane for all frames. The fourth data set (Experiment 4) has stationary lateral and axial coordinates with a constant change in elevation and is meant to test the accuracy of offset estimation.

1) Reference Methods for Comparison: In addition to the proposed combination of neural network tracking with Kalman filtering [neural network Kalman (NNK)], we test two other reduced models: plain Gaussian random walk (NNK-RW) and independent subsequent states (NNK-I). Mathematically, they differ with respect to the dynamic model: NNK-RW assumes that $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \boldsymbol{c}_k$ and NNK-I that $\boldsymbol{x}_k = \boldsymbol{c}_k$ [compared to (4)]. We compared our method to MI tracking, which estimates the object location as the pixel with the highest intensity and thus only outputs a 2-D location. To evaluate the performance of all considered methods, we computed the mean 2-D Euclidean distance from the estimated axial and lateral coordinates to the ground truth using synthetic data. We additionally evaluate the accuracy of the 3-D positional estimate with Experiment 4. Finally, we examined the localization accuracy of NNK as a function of depth (axial coordinate) and out-of-plane offset with an axial line trajectory simulated with different values for out-of-plane offsets. This evaluation was done using the 3-D Euclidean distance.

C. Experimental Data

To test the out-of-plane tracking abilities of the method, we performed one physical experiment closely matching simulated Experiment 4 and one to test accuracy when moving further out-of-plane. In the first experiment, the tip of a metal pushpin (tip diameter: 50 μm) was used to emulate a point object and was submerged in water as a homogeneous background medium. This pin was placed centrally within the imaging aperture at an axial distance of 7.5 mm and was attached to a manual translation stage (PT1/M, Thorlabs, Bergkirchen, Germany) to allow for controlled motion orthogonal to the image plane (i.e., "out-of-plane") and provide ground-truth positions for quantitative evaluation. The tip of this pin was placed at out-of-plane positions ranging between -3 and $+5$ mm at a regular step size of 100 μm, and at each position, a 2-D OpUS image was acquired. For the second physical experiment, the out-of-plane and lateral positions were varied simultaneously to mimic a nonorthogonal drift of the object. The object was initially located centrally in the image at an axial depth of 7.5 mm and moved in increments of 100 μm (lateral) and 200 μm (out-of-plane) to a total out-

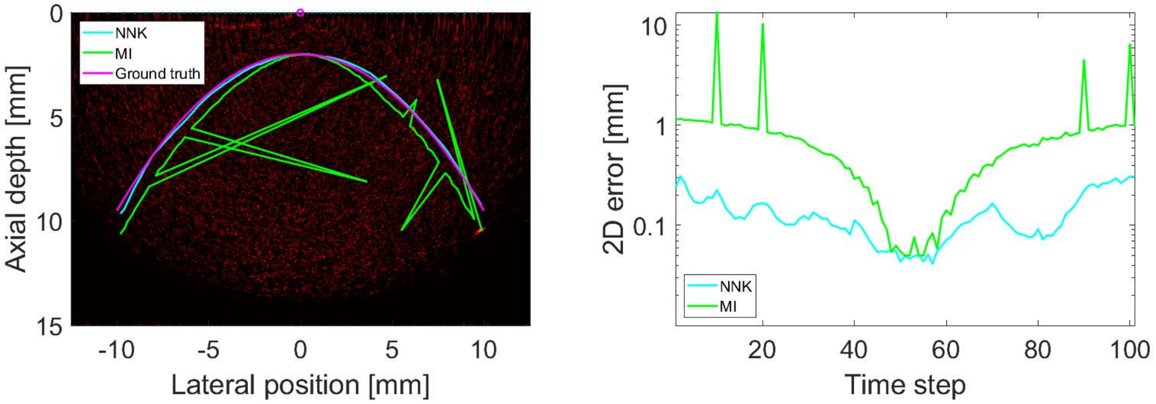

Fig. 4. Tracking for synthetic experiment 2 with MI and NNK. The axial coordinate is overestimated with MI due to the absence of axial aberration correction and the location is severely misestimated in some frames because of increased noise.

of-plane position of 10 mm. The SNR for the experimental data is estimated to be around 4 dB.

D. Tuning Parameter Selection

The tracking algorithm requires the selection of four process noise variance parameters that can be used to fine-tune the process. Values that are too low ($<(10^{-4}$ mm$)^2$) may cause the estimated trajectory to be too restricted in case of rapid changes in the position or velocity of the object. With ideal data (high SNR), values that are too high have little effect, but with noisier data robustness suffers. This transition starts to take place at around the value of (0.2 mm$)^2$. Thus, the parameters were chosen empirically as (0.005 mm$)^2$ to allow enough flexibility to recover from sudden changes in the position and velocity but at the same time provide robustness against noise.

IV. Results

A. Results on Simulated Data

Table I shows the errors for the four tracking experiments and different varying methods. The proposed NNK methods perform clearly better than MI with data where the tracked object is out-of-plane, due to the correction of the axial aberration caused by out-of-plane offset: the localization error for all NNK methods is 0.13 mm, while for MI, it is five-fold. For occasionally noisier data filtering-based NNK and NNK-RW that retain their performance and clearly outperform NNK-I and MI that do not assume dependence between subsequent positions, this indicates that a filtering approach is necessary to provide robustness. This benefit of filtering and axial aberration correction is clearly visible in Fig. 4, where MI overestimates the axial coordinate and, for some noisy images, the estimate jumps off the trajectory (green spikes). Nevertheless, if we consider no out-of-plane offset without additional noise, all methods perform comparably well with a localization error of around 0.11 mm. In terms of worst case performance, NNK performs the best with maximum error less than three times the mean error in every experiment. In Experiment 2 with increased noise, this ratio increases to almost five with NNK-RW and over 10 with NNK-I and MI.

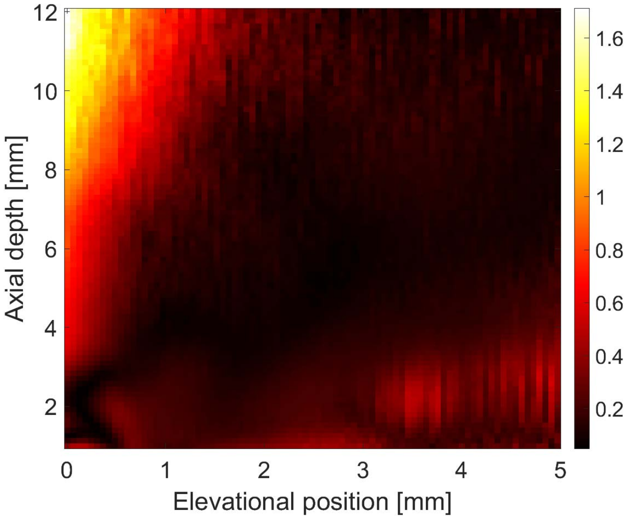

Fig. 5. 3-D localization error (mm) of NNK as a function of elevational position and axial depth for lateral position at 0 mm.

Videos of tracking results with NNK for Experiments 1 and 4 are presented in Supplementary Videos 1 and 2.

Fig. 5 shows how the 3-D error depends on depth and out-of-plane offset for lateral position at 0 mm. Interestingly, the error is largest (~1.6 mm) when the offset is small and the depth is large. For bigger offsets, the error gets smaller. This indicates that there are no strong enough markers in the data to reliably estimate all three coordinates when the out-of-plane offset is small. However, the reliability increases with higher out-of-plane offsets. The relatively poor performance at shallow depths and large elevational offset (bottom right of Fig. 5) is caused by the directivity of the US sources and a large propagation distance, which result in poor SNR and hence poor amplitude estimates. The same pattern was obtained for different lateral positions (data not shown), with minor fluctuations in the stable region and degradation for the extreme points close to the imaging boundaries. We also note that even though the out-of-plane offset estimation is not correct for all instances, the estimated axial aberration and filtering approach still provides accurate results, as can be seen in Fig. 6: lateral/axial trajectory is very close to the ground truth.

TABLE I

2-D Errors (Axial/Lateral) as Mean Distance (Standard Deviation, Maximum Distance) in Millimeters With Respect to Ground Truth of Different Tracking Schemes for the Synthetic Data Experiments: Proposed Method (NNK), the Two Reduced Models Using Gaussian Random Walk (NNK-RW) and Independent Subsequent States (NNK-I), and the Reference Method (MI)

Test data	NNK	NNK-RW	NNK-I	MI (reference)
Exp. 1: Regular	0.12 (0.076) [0.34]	0.12 (0.078) [0.37]	0.13 (0.078) [0.38]	0.62 (0.36) [1.17]
Exp. 2: Increased noise	0.13 (0.080) [0.42]	0.15 (0.12) [0.78]	0.51 (1.77) [13.09]	1.06 (2.19) [16.51]
Exp. 3: No offset	0.11 (0.053) [0.26]	0.11 (0.052) [0.26]	0.11 (0.052) [0.26]	0.10 (0.058) [0.28]
Exp. 4: Stationary axial & lateral	0.096 (0.040) [0.19]	0.094 (0.035) [0.18]	0.098 (0.039) [0.20]	0.33 (0.34) [1.30]

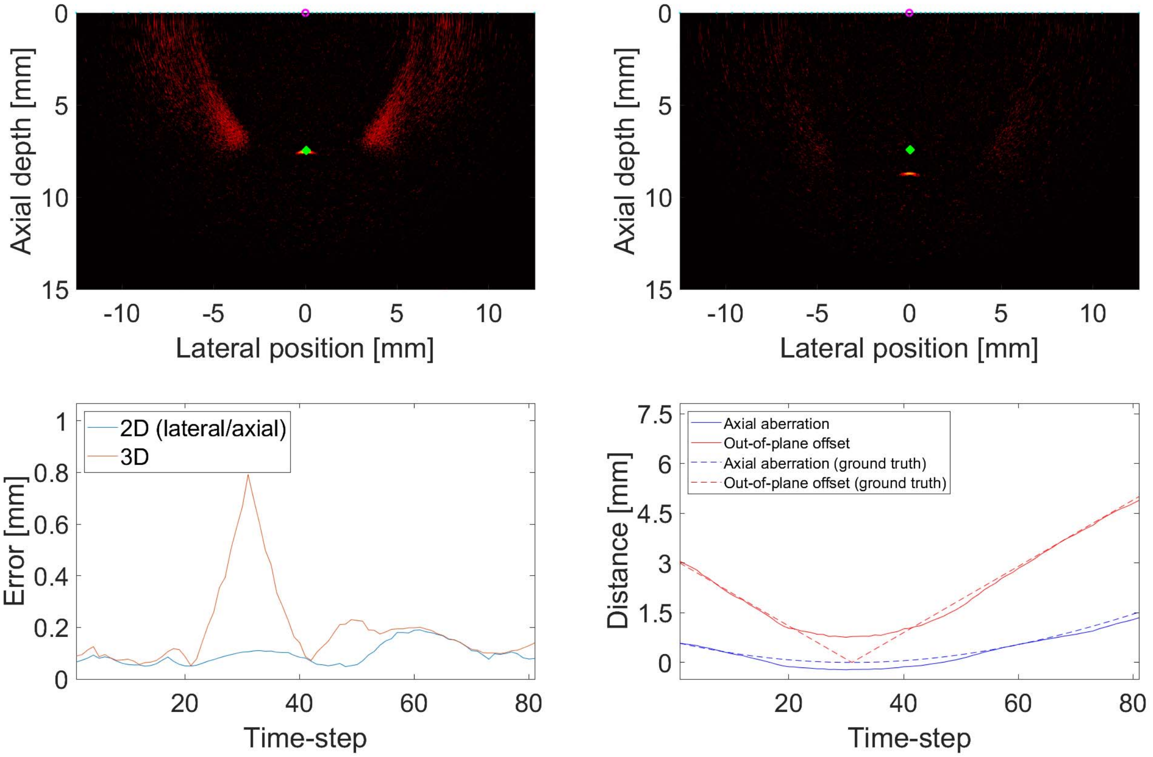

Fig. 6. Tracking for synthetic experiment 4. Top left: tracked location (green dot) at time step 40 (out-of-plane distance ~1 mm). Top right: tracked location at the last time step (out-of-plane distance 5 mm). Bottom left: 2-D and 3-D error (Euclidean distance) from the ground truth. Bottom right: out-of-plane offset (elevational distance) and axial aberration over time.

B. Results on Experimental Data

Before we could apply NNK for tracking, the experimental data required normalization to match the amplitude (in arbitrary units) of synthetic data. While the experimental data are clearly noisier than synthetic data, the tracking method performs reasonably well. The axial aberration correction works and the out-of-plane offset largely follows the expected trajectory (see Figs. 7 and 8). However, when going farther than 6 mm away from the imaging plane in the second experiment, the algorithm breaks down and the localization error increases rapidly. The axial/lateral localization error is mostly below 0.3 mm and with a mean of around 0.2 mm for the first experimental dataset. For the second dataset, the same holds for frames 1–35, after which the error starts to increase. In the first experimental dataset, most of the 3-D errors originate from the elevational component. This effect is not as pronounced in the second dataset. We note that in the first dataset, the estimation accuracy is worse for the first part with negative elevational distance. This indicates that the imaging probe suffers from a source of asymmetry (e.g., acoustic shadowing by an edge or unexpected source or receiver directivity) that has not been accurately accounted for in the numerical model. This effect can also be observed in the video of this tracking experiment in Supplementary Video 3.

C. Computation Times

Performing one iteration of tracking took on average 298 ms. The time was split as follows: reconstructing the image (269 ms), finding high-intensity pixels (21 ms), neural network prediction (6.6 ms), and Kalman filtering (0.43 ms), and displaying the image (83 ms). Hence, reconstructing the image is clearly the most time-consuming task and the NNK framework only adds a small computational overhead.

Preliminary tasks include generating training data and training the neural network. Training data with 8800 rows were generated in about 1 h. Training the neural network took on average only 20 s with a median of 17 s (over ten training attempts). Generating the training data and training the neural network has to be done only once, which means that tracking is essentially performed in real time. Computations were performed on a workstation with AMD Ryzen Threadripper

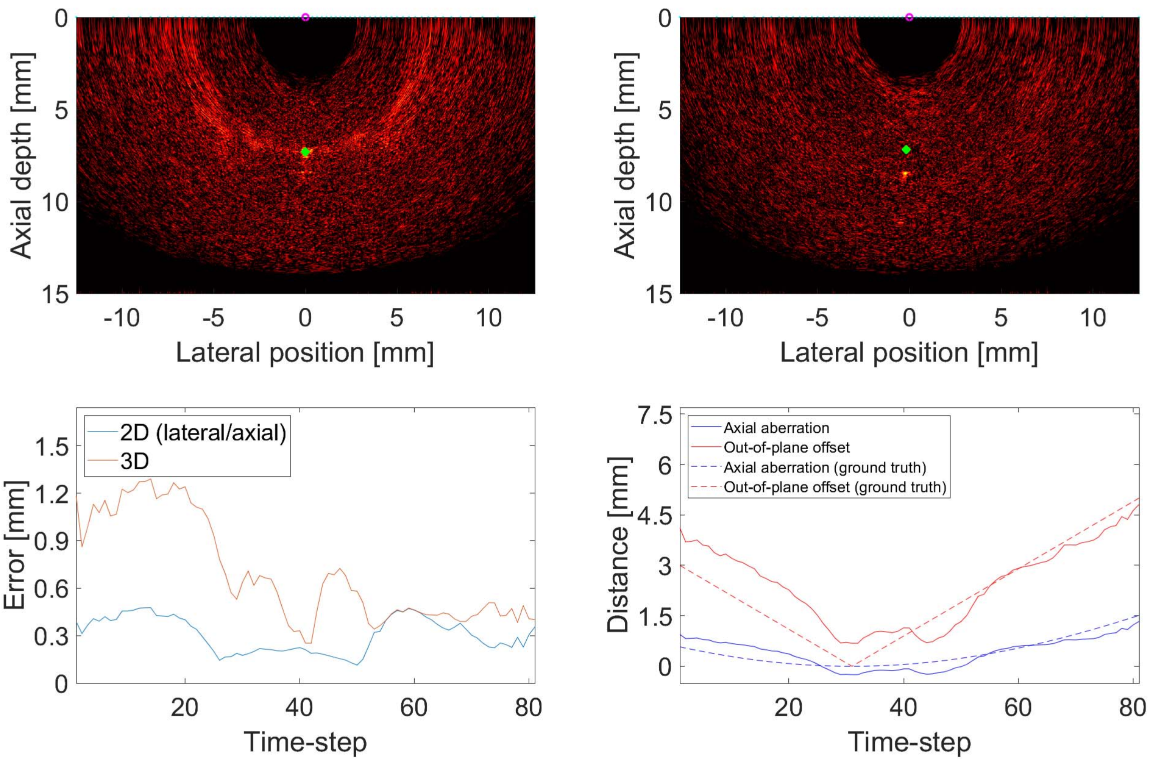

Fig. 7. Tracking for experimental dataset 1 matched to synthetic experiment 4. Top left: tracked location (green dot) at time step 40 (out-of-plane distance ~1 mm). Top right: tracked location at the last time step (out-of-plane distance 5 mm). Bottom left: 2-D and 3-D error (Euclidean distance) from the ground truth and (bottom right) out-of-plane offset (elevational distance) and axial image aberration over time.

2950X processor and 32-GB RAM. The codes for NNK are written in MATLAB, while the OpUS simulator uses routines compiled from C++ for CPU.

V. Discussion

A. Discussion of Markers

Our experiments show that the magnitude of the simulated measurement data coupled with axial and lateral position is correlated with the out-of-plane offset and an axial positional aberration, as shown in Fig. 3. This correlation can be exploited with machine learning to find a nonlinear relationship between these quantities. We also found that this correlation holds with experimental data after data normalization. Nevertheless, the tracking with experimental data is less stable and shows reduced accuracy. This can be partly attributed to reduced SNR in the measurement data as well as deviations from the ideal assumptions in the simulation, but the Kalman filtering offers a framework to partly mitigate these negative effects and is still able to provide a stable estimation of the axial/lateral coordinate.

In this work, we have used an OpUS simulator and computed all markers, offset, and axial aberration, from the simulated data. We note that under the assumption of a homogenous medium, we can alternatively calculate the axial aberration analytically using the point-spread function of the imaging system. Nevertheless, we have observed in conducted experiments that an analytic calculation can help for small distances in the simulated data but will lead to decreased accuracy for the experimental datasets. Thus, computing the axial aberration from the reconstructed US images for training seems to provide more generalizable markers for the estimation process. Furthermore, this fully simulated framework can be extended to heterogeneous media.

B. Offset Accuracy and Range

The quantitative analysis shows that tracking accuracy is worse for small offsets and larger depths. Most of this error seems to be caused by the out-of-plane offset estimation, while axial and lateral components are tracked well. This indicates that even though the offset might be incorrectly estimated, the proposed axial aberration correction still works. We attribute the difficulty of estimating small elevations to the shallow slope of the out-of-plane amplitude decay for larger axial depths, as shown in Fig. 3. This decrease in accuracy is also seen in the error matrix in Fig. 5 and worsens with increasing axial distance. Thus, it is important to provide both offset and axial aberration, to provide accurate tracking results within the Kalman filtering. For the simulated data, this effect shows symmetrically at roughly out-of-plane distance under 1 mm. For the experimental data, the threshold distance is similar, but an asymmetric behavior can be observed, where positive elevational distance is underestimated and negative overestimated, as shown in Fig. 7. This indicates that the purely simulated framework can in principle be transferred to the experimental case, but small asymmetries in the imaging probe would need to be investigated and accounted for to further improve the results.

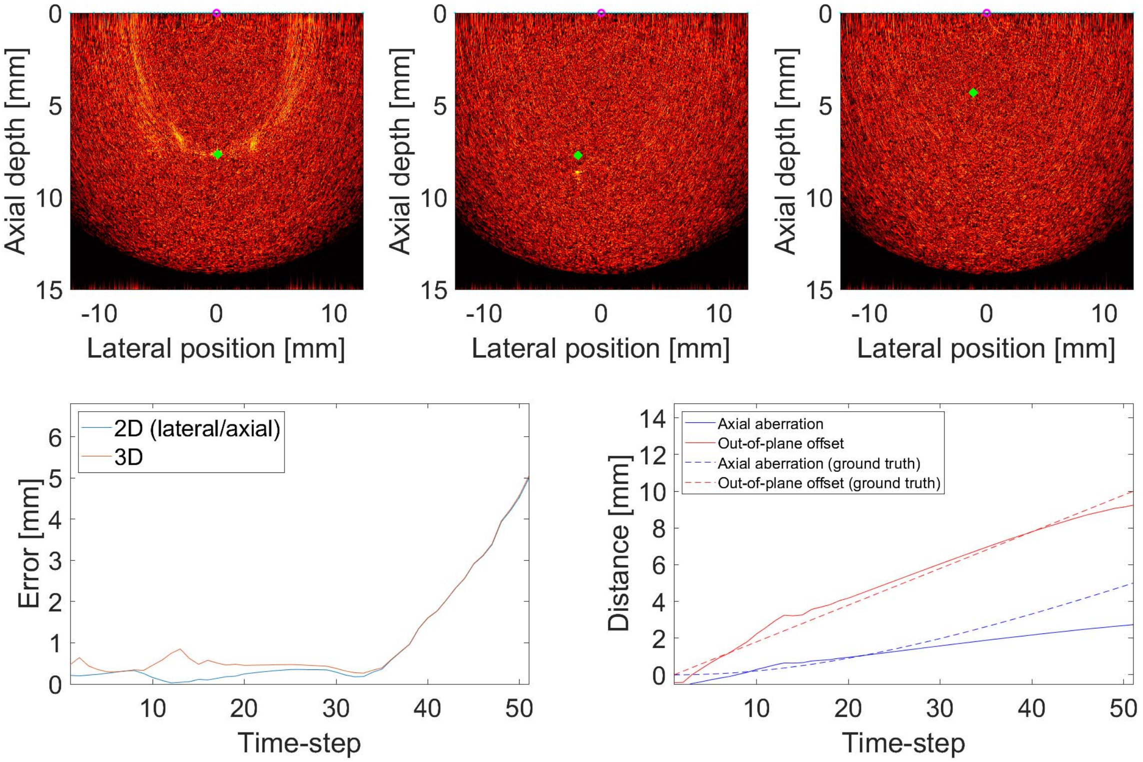

Fig. 8. Tracking for experimental dataset 2. Top left: tracked location (green dot) at time step 1 (out-of-plane distance 0 mm). Top middle: tracked location at time step 20 (out-of-plane distance ∼4 mm). Top right: tracked location at the last time step (out-of-plane distance 10 mm). Bottom left: 2-D and 3-D error (Euclidean distance) from the ground truth and (bottom right) out-of-plane offset (elevational distance) and axial image aberration over time.

The imaging system in this work uses unfocussed, weakly directional circular optical US sources that insonify a wide elevational range. Consequently, out-of-plane tracking can be performed over a large elevational range limited by the SNR of the B-scan: in this work, up to 6 mm for positive elevational distance. For larger out-of-plane offsets, the RF data SNR is insufficient to reliably detect the pulse-echo signal. However, for imaging probes comprising directional sources or in the presence of an acoustic lens, this range could be different.

C. Limitations and Clinical Applicability

In this study, we show that one can successfully use the correlation between out-of-plane amplitude decay and axial/lateral positions to estimate 3-D locations from linear array data. Nevertheless, this correlation was observed in a simplified simulated and experimental setting assuming homogeneous media, i.e., a water bath in the experimental setup. In order to move toward clinically realistic scenarios, we need to consider various deviations from the ideal case. In the following, we discuss limitations and extensions needed for clinical applicability.

1) Speckle: The tracking presented is based on MI pixels, and as such, speckle of low-to-moderate intensity (compared to the intensity of the image of the object, for instance in the case of a highly echogenic needle tip) is not expected to interfere with the estimation procedure. However, strong speckle could result in tracking errors if only the amplitude is used as marker μ. In this case, the NN would likely need to be adapted to not only extract amplitude information but also its variation across the imaging aperture—as this spatial variation for the actual object would differ from that of speckle signal.

2) Inhomogeneous Media: In this work, we have presented results for homogeneous media. For an application to inhomogeneous media with spatially varying speed of sound, the NN needs to be trained differently. Here, approximate synthetic training data could, for instance, be generated using ensemble-mean speed of sound maps observed over a group of patients and simulated with advanced methods such as the k-Wave toolbox [46]. In addition, acoustical attenuation would affect the extracted parameter μ and hence complicate accurate out-of-plane tracking. For applications to actual tissue, this attenuation should be included in the model used to train the NN.

3) Object Geometry: The results presented in this work were obtained for point-like objects, such as clinically encountered in the form of microbubbles, fiducial markers, brachytherapy seeds, and radio-opaque markers on surgical instruments. In order to extend the method to finite-sized objects, their US response needs to be accurately modeled. Conceptually, the method applies to finite-sized spherically symmetric objects, such as large needle tips or spherical implants. However, more complicated object geometries, such as long needles or asymmetric beads, are complicated due to nonlinearities arising from high echogenicity and ambiguities in differentiating between needle tips and shafts or different object orientations. Such objects would require further refinement in the NN markers and the underlying acoustical model to make accurate predictions.

4) Tracking Range and Accuracy: In the experimental results presented here, an out-of-plane tracking range of up to ±6 mm was demonstrated and was limited by SNR. This range could be further extended, provided that the imaging probe emits a sufficiently diverging field in the elevational direction and SNR is improved, for instance using coded excitation schemes. The optical US imaging system considered here does not apply acoustic focusing in the elevational direction and hence is ideally suited to tracking across a wide out-of-plane range. The achieved range of ±6 mm is clinically highly relevant, as correcting object placement over larger distances is typically not possible without removal and reentry of a surgical tool. For clinical imaging systems, which typically apply elevational focusing, geometrical distortion and signal amplitude decay resulting from out-of-plane offsets will still occur, and the proposed method can still be applied, provided that it is retrained. However, the out-of-plane tracking range and accuracy will depend on the tightness of the elevational focusing and hence will vary with both the F-number of the elevational focusing lens and the axial position of the object relative to the focal distance.

5) Experimental Setup: Here, we used a prototype optical US imaging setup to perform experimental validation measurements. While these were reasonably successful (cf. Figs. 7 and 8) due to the availability of a highly accurate and efficient numerical model, the developmental nature of this system limited its practicality. Slow fluctuations were observed in the efficiency of the optical US sources and the sensitivity of the detector, which resulted in unforeseen variations in the US amplitudes. As the NN estimation requires the amplitude to be accurately known, these fluctuations limited the range of object trajectories to those that could be traversed quickly. This also resulted in slight differences between simulated and experimental data. Nevertheless, the estimation network generalized well to the experimental data, and the filtering approach further stabilized the estimation process. However, in principle, any US imaging system that grants access to RF data could be used, even those generating focused transmissions—although the NN would need to be retrained for each considered setup and tracking accuracy and range will vary.

D. Extensions

The presented framework can be extended to tracking multiple point sources. In that case, a data association task would have to be solved [19], [47]. This means determining which pixels belong to which target. Furthermore, this would also require a more complicated setup for training data generation and extraction of markers from the measured time series.

Another interesting avenue to pursue would be the extension to needle tracking (as opposed to point object tracking) where the shaft can be mistaken for the tip, which would require shape detection instead of simple tracking. Alternatively, one can overcome the shaft problem by adjusting our framework to data obtained with an active listening needle [8] that relies on the reception of US pulses by a fiber-optic hydrophone (FOH) integrated into the needle.

Finally, due to the limitations mentioned in Section V-C, it might be promising to consider the full RF time series as input to a convolutional neural network that is also capable of extracting geometric markers from the data. This information can be still paired with manually extracted markers, such as amplitude, to improve the tracking accuracy and robustness for future applications.

VI. Conclusion

This work proposes a neural network and Kalman filtering approach to perform accurate and robust object tracking in 3-D from linear array data. The essential step is that a neural network estimates the third dimension and its impact on the 2-D US image, in form of aberration in the axial coordinate. Then Kalman filtering is performed for all coordinates to provide a robust estimate with respect to noise. We have shown that the framework can provide high accuracy in estimating axial and lateral coordinates for objects that are not in-plane as well as the corresponding elevational distance. If the point source is too close to the imaging plane, it remains difficult to provide an accurate estimate on the elevational distance, but the proposed NNK framework is still capable to provide a robust and accurate estimate on the lateral/axial coordinate.

Acknowledgment

The authors would like to thank the three anonymous reviewers for their constructive comments, which significantly helped them to improve the presentation of the work. Codes for neural network Kalman (NNK) can be found in https://github.com/aarjas/NNK.

References

[1] D. Ackermann, G. Schmitz, and S. Member, "Detection and tracking of multiple microbubbles in ultrasound B-mode images," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 63, no. 1, pp. 72–82, Jan. 2016.

[2] K. Christensen-Jeffries *et al.*, "Super-resolution ultrasound imaging," *Ultrasound Med. Biol.*, vol. 46, no. 4, pp. 865–891, 2020.

[3] G. Dhadham, S. Hoffe, C. Harris, and J. Klapman, "Endoscopic ultrasound-guided fiducial marker placement for image-guided radiation therapy without fluoroscopy: Safety and technical feasibility," *Endoscopy Int. Open*, vol. 4, no. 3, pp. E378–E382, Mar. 2016.

[4] K. J. Chin, A. Perlas, V. W. S. Chan, and R. Brull, "Needle visualization in ultrasound-guided regional anesthesia: Challenges and solutions," *Regional Anesthesia Pain Med.*, vol. 33, no. 6, pp. 532–544, 2008.

[5] T. H. Helbich, W. Matzek, and M. Fuchsjäger, "Stereotactic and ultrasound-guided breast biopsy," *Eur. Radiol.*, vol. 14, no. 3, pp. 383–393, Mar. 2004.

[6] F. Daffos, M. Capella-Pavlovsky, and F. Forestier, "Fetal blood, sampling during pregnancy with use of a needle guided by ultrasound: A study of 606 consecutive cases," *Amer. J. Obstetrics Gynecol.*, vol. 153, no. 6, pp. 655–660, Nov. 1985.

[7] D. Karakitsos *et al.*, "Real-time ultrasound-guided catheterisation of the internal jugular vein: A prospective comparison with the landmark technique in critical care patients," *Crit. Care*, vol. 10, no. 6, pp. 1–8, 2006.

[8] W. Xia *et al.*, "In-plane ultrasonic needle tracking using a fiber-optic hydrophone," *Med. Phys.*, vol. 42, pp. 5983–5991, Oct. 2015.

[9] P. Beigi, S. E. Salcudean, G. C. Ng, and R. Rohling, "Enhancement of needle visualization and localization in ultrasound," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, pp. 169–178, Sep. 2020.

[10] X. Guo, H.-J. Kang, R. Etienne-Cummings, and E. M. Boctor, "Active ultrasound pattern injection system (AUSPIS) for interventional tool guidance," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e104262.

[11] M. Graham *et al.*, "*In vivo* demonstration of photoacoustic image guidance and robotic visual servoing for cardiac catheter-based interventions," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1015–1029, Apr. 2020.

[12] A. Krupa, G. Fichtinger, and G. D. Hager, “Real-time tissue tracking with B-mode ultrasound using speckle and visual servoing,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2007, pp. 1–8.

[13] N. Afsham, M. Najafi, P. Abolmaesumi, and R. Rohling, “A generalized correlation-based model for out-of-plane motion estimation in freehand ultrasound,” *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 186–199, Jan. 2014.

[14] W. Xia *et al.*, “Looking beyond the imaging plane: 3D needle tracking with a linear array ultrasound probe,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, Dec. 2017.

[15] W. Xia *et al.*, “Three-dimensional ultrasonic needle tip tracking with a fiber-optic ultrasound receiver,” *J. Visualized Exp.*, vol. 138, Aug. 2018, Art. no. 57207.

[16] P. Chatelain, A. Krupa, and M. Marchal, “Real-time needle detection and tracking using a visually servoed 3D ultrasound probe,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 1676–1681.

[17] E. J. Alles, E. C. Mackle, S. Noimark, E. Z. Zhang, P. C. Beard, and A. E. Desjardins, “Freehand and video-rate all-optical ultrasound imaging,” *Ultrasonics*, vol. 116, Sep. 2021, Art. no. 106514.

[18] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Trans. ASME, D, J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[19] S. Särkkä, *Bayesian Filtering Smoothing*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[20] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, vol. 160. Cham, Switzerland: Springer, 2006.

[21] S. Prince, V. Kolehmainen, J. P. Kaipio, M. A. Franceschini, D. Boas, and S. R. Arridge, “Time-series estimation of biological factors in optical diffusion tomography,” *Phys. Med. Biol.*, vol. 48, no. 11, p. 1491, 2003.

[22] G. Liang, F. Dong, V. Kolehmainen, M. Vauhkonen, and S. Ren, “Non-stationary image reconstruction in ultrasonic transmission tomography using Kalman filter and dimension reduction,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[23] J. Hakkarainen, Z. Purisha, A. Solonen, and S. Siltanen, “Undersampled dynamic X-ray tomography with dimension reduction Kalman filter,” *IEEE Trans. Comput. Imag.*, vol. 5, no. 3, pp. 492–501, Sep. 2019.

[24] O. Solomon, R. J. G. van Sloun, H. Wijkstra, M. Mischi, and Y. C. Eldar, “Exploiting flow dynamics for superresolution in contrast-enhanced ultrasound,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 10, pp. 1573–1586, Oct. 2019.

[25] H. Takeshima, T. Tanaka, and R. Imai, “Position detection of guidewire tip emitting ultrasound by using a Kalman filter,” *Jpn. J. Appl. Phys.*, vol. 60, no. 8, Aug. 2021, Art. no. 087002.

[26] J. A. Jensen, S. I. Nikolov, K. L. Gammelmark, and M. H. Pedersen, “Synthetic aperture ultrasound imaging,” *Ultrasonics*, vol. 44, pp. e5–e15, Dec. 2006.

[27] R. S. Cobbold, *Foundations of Biomedical Ultrasound*. London, U.K.: Oxford Univ. Press, 2006.

[28] S. Ipsen, R. Bruder, R. O’Brien, P. J. Keall, A. Schweikard, and P. R. Poulsen, “Online 4D ultrasound guidance for real-time motion compensation by MLC tracking,” *Med. Phys.*, vol. 43, no. 10, pp. 5695–5704, Sep. 2016.

[29] V. De Luca *et al.*, “The 2014 liver ultrasound tracking benchmark,” *Phys. Med. Biol.*, vol. 60, p. 5571, Jul. 2015.

[30] E. Harris, N. R. Miller, J. C. Bamber, J. R. N. Symonds-Tayler, and P. M. Evans, “Speckle tracking in a phantom and feature-based tracking in liver in the presence of respiratory motion using 4D ultrasound,” *Phys. Med. Biol.*, vol. 55, no. 12, p. 3363, 2010.

[31] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[32] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Symmetric log-domain diffeomorphic registration: A demons-based approach,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2008, pp. 754–761.

[33] D. Allman, A. Reiter, and M. A. L. Bell, “Photoacoustic source detection and reflection artifact removal enabled by deep learning,” *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1464–1477, Jun. 2018.

[34] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, “A deep learning-based approach to identify *in vivo* catheter tips during photoacoustic-guided cardiac interventions,” *Proc. SPIE*, vol. 10878, Feb. 2019, Art. no. 108785E.

[35] A. Yazdani, S. Agrawal, K. Johnstonbaugh, S.-R. Kothapalli, and V. Monga, “Simultaneous denoising and localization network for photoacoustic target localization,” *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2367–2379, Sep. 2021.

[36] L. Hong, “Discrete constant-velocity-equivalent multirate models for target tracking,” *Math. Comput. Model.*, vol. 28, no. 11, pp. 7–18, 1998.

[37] E. J. Alles *et al.*, “Video-rate all-optical ultrasound imaging,” *Biomed. Opt. Exp.*, vol. 9, no. 8, pp. 3481–3494, 2018.

[38] E. J. Alles and A. E. Desjardins, “Source density apodization: Image artifact suppression through source pitch nonuniformity,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 3, pp. 497–504, Mar. 2020.

[39] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1, no. 2. Cambridge, MA, USA: MIT Press, 2016.

[40] D. W. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963. [Online]. Available: http://www.jstor.org/stable/2098941

[41] K. Levenberg, “A method for the solution of certain non-linear problems in least squares,” *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, 1944. [Online]. Available: http://www.jstor.org/stable/43633451

[42] P. Beard, “Biomedical photoacoustic imaging,” *Interface Focus*, vol. 1, pp. 602–631, Aug. 2011.

[43] J. A. Guggenheim *et al.*, “Ultrasensitive plano-concave optical microresonators for ultrasound sensing,” *Nature Photon.*, vol. 11, no. 11, pp. 714–719, 2017.

[44] R. J. McGough, “Rapid calculations of time-harmonic nearfield pressures produced by rectangular pistons,” *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 1934–1941, May 2004.

[45] J. F. Kelly and R. J. McGough, “A time-space decomposition method for calculating the nearfield pressure generated by a pulsed circular piston,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 53, no. 6, pp. 1150–1159, Jun. 2006.

[46] B. E. Treeby and B. T. Cox, “K-wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields,” *J. Biomed. Opt.*, vol. 15, no. 2, 2010, Art. no. 021314.

[47] D. B. Reid, “An algorithm for tracking multiple targets,” *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 843–854, Dec. 1979.