Dynamic coronary roadmapping via catheter tip tracking in X-ray fluoroscopy with deep learning based Bayesian filtering

Ma, Hua; Smal, Ihor; Daemen, Joost; Walsum, Theo van

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Dynamic Coronary Roadmapping via Catheter Tip Tracking in X-ray Fluoroscopy with Deep Learning Based Bayesian Filtering[†,††]

Hua Ma[a,*], Ihor Smal[b], Joost Daemen[c], Theo van Walsum[a]

[a]*Biomedical Imaging Group Rotterdam, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands*
[b]*Department of Geoscience and Remote Sensing, Delft University of Technology, Delft, The Netherlands*
[c]*Department of Cardiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands*

Abstract

Percutaneous coronary intervention (PCI) is typically performed with image guidance using X-ray angiograms in which coronary arteries are opacified with X-ray opaque contrast agents. Interventional cardiologists typically navigate instruments using non-contrast-enhanced fluoroscopic images, since higher use of contrast agents increases the risk of kidney failure. When using fluoroscopic images, the interventional cardiologist needs to rely on a mental anatomical reconstruction. This paper reports on the development of a novel dynamic coronary roadmapping approach for improving visual feedback and reducing contrast use during PCI. The approach compensates cardiac and respiratory induced vessel motion by ECG alignment and catheter tip tracking in X-ray fluoroscopy, respectively. In particular, for accurate and robust tracking of the catheter tip, we proposed a new deep learning based Bayesian filtering method that integrates the detection outcome of a convolutional neural network and the motion estimation between frames using a particle filtering framework. The proposed roadmapping and tracking approaches were validated on clinical X-ray images, achieving accurate performance on both catheter tip tracking and dynamic coronary roadmapping experiments. In addition, our approach runs in real-time on a computer with a single GPU and has the potential to be integrated into the clinical workflow of PCI procedures, providing cardiologists with visual guidance during interventions without the need of extra use of contrast agent.

Keywords: dynamic coronary roadmapping, X-ray fluoroscopy, catheter tip tracking, deep learning, Bayesian filtering, particle filter.

1. Introduction

1.1. Clinical Background

Percutaneous coronary intervention (PCI) is a minimally invasive procedure for treating patients with coronary artery disease. During these procedures, medical instruments inserted through a guiding catheter are advanced to treat coronary stenoses.

A guiding catheter is firstly positioned into the ostium of the coronary artery. Through the guiding catheter, a balloon catheter carrying a stent is introduced over a guidewire to the stenosed location. The balloon is then inflated and the stent is deployed to prevent the vessel from collapsing and restenosing.

PCI is typically performed with image-guidance using X-ray angiography (XA). Coronary arteries are visualized with X-ray opaque contrast agent. During the procedure, interventional cardiologists may repeatedly inject contrast agent to visualize the vessels, as the opacification of coronary arteries only lasts for a short period. The amount of periprocedural contrast use has been correlated to operator experience,

---

procedural complexity, renal function and imaging setup (Piayda et al. (2018)). Furthermore, the risk for contrast induced nephropathy has been associated to contrast volume (Tehrani et al. (2013)). Manoeuvring guidewires and material, however, typically occurs without continuous contrast injections. In these situations, the navigation of devices is guided with "vessel-free" fluoroscopic images. Cardiologists have to mentally reconstruct the position of vessels and stenosis based on previous angiograms.

1.2. Dynamic Coronary Roadmapping

Dynamic coronary roadmapping (DCR) is a promising solution towards improving visual feedback and reducing usage of contrast medium during PCI (Elion (1989); Zhu et al. (2010); Manhart et al. (2011); Kim et al. (2018)). This approach dynamically superimposes images or models of coronary arteries onto live X-ray fluoroscopic sequences. The dynamic overlay serves as a roadmap that provides immediate feedback to cardiologists during the intervention, so as to assist in navigating a guidewire into the appropriate coronary branch and proper placement of a stent at the stenotic site with reduced application of contrast agent. Studies with a phantom setup using research software (Kim et al. (2018)) or on first cases of clinical interventions using commercially available systems (Dannenberg et al. (2016); Yabe et al. (2018); Takimura et al. (2018)) have investigated the usefulness of DCR in assisting PCI, reporting that DCR helps to reduce procedure time, radiation dose and contrast volume.

To develop a DCR system, it is important but yet a challenge to accurately overlay a roadmap of coronary arteries onto an X-ray fluoroscopic image, as limited information of vessels is present in the target fluoroscopic image for inferring the compensation of the vessel motion resulting from patient respiration and heartbeat. The methods that have been proposed on motion compensation for DCR can be generally grouped into two categories: direct roadmapping and model-based approaches.

Direct roadmapping methods use information from X-ray images and ECG signals to directly correct the motion caused by respiration and heartbeat. The first DCR system (Elion (1989)) used digital subtraction of a contrast sequence and a mask sequence to create a full cardiac cycle of coronary roadmaps. The roadmaps were stored and later synchronized with the live fluoroscopic sequence by aligning the R waves of their corresponding ECG signals. This system compensates the cardiac motion of vessels, yet does not correct the respiratory motion during interventions. Two later studies by Zhu et al. (2010) and Manhart et al. (2011) introduced image-based respiratory motion compensation methods for DCR. Their methods assumed an affine respiratory motion model in ECG-gated fluoroscopic frames and recovered the respiratory motion from soft tissues with special handling of static structures. The limitation of these approaches is that they require relevant tissue to be sufficiently visible in the field of view for reliable motion compensation which is not always the case. In addition, they require to be run on cardiac-gated frames. In a more recent work by Kim et al. (2018), binary vessel masks were created as the roadmaps from at least one cardiac cycle of angiographic images. Temporal alignment of the roadmaps and the fluoroscopic sequence, which compensated the cardiac motion of vessels, was performed by registering ECG signals using cross-correlation. Additionally, the respiratory motion was corrected by aligning the guidewire centerline in the fluoroscopy to the contour of vessels in the angiogram where the roadmaps were created. The system has been shown useful in a phantom-based study, nevertheless no accuracy evaluation of the spatiotemporal alignment was presented. Furthermore, the spatial registration relies on robust extraction of vessels and guidewires which is often challenging for X-ray images.

Unlike direct roadmapping, the model-based approaches build a model to predict motion in fluoroscopic frames. The motion models are often functions that relate the motion of roadmaps to surrogate signals derived from images or ECG, so that once the surrogates for fluoroscopic frames are obtained, the motion can be computed by the model. For cardiac interventions including PCI, the organ motion is mainly affected by respiratory and cardiac motion. Many previous works often built a motion model parameterized by a cardiac signal derived from ECG and a respiratory signal obtained from diaphragm tracking (Shechter et al. (2005); Timinger

2

et al. (2005); Faranesh et al. (2013)) or automatic PCA-based surrogate (Fischer et al. (2018)). Some other works model only the respiratory motion in cardiac-gated images (Schneider et al. (2010); King et al. (2009); Peressutti et al. (2013)). For a complete review on respiratory motion modeling, we refer readers to the survey article by McClelland et al. (2013). One limitation of the model-based approaches is that the motion models are often patient-specific, which requires training the model every time for a new subject. Additionally, once the surrogate values during inference are out of the surrogate range for building the model, e.g. for abnormal motion, extrapolation is needed, which may hamper accurate motion compensation.

1.3. Interventional / Surgical Tool Tracking

Tracking interventional tools is relevant for motion compensation (Schneider et al. (2010); Brost et al. (2010); Ma et al. (2012); Baka et al. (2015); Ambrosini et al. (2017b)). In particular for PCI, the guiding catheter tip typically remains within the coronary ostium which is in the field of view during the largest part of the intervention, making it a suitable landmark for tracking. Baka et al. (2015) have shown that catheter tip motion during PCI can be modeled as a combination of cardiac and respiratory motion. As using catheter tip displacement can only correct translational motion, Baka et al. (2015) further showed that, compared to a rigid motion model for the respiratory motion, modeling only the translational part of the respiratory motion deteriorated the accuracy marginally, which confirms the observations by Shechter et al. (2004) that the rotational part of respiratory motion is small. These findings motivate motion compensation for DCR through tracking the catheter tip in X-ray fluoroscopic sequences.

Many works have been proposed to address the problem of tracking interventional or surgical tools in medical images for various applications. The tracking methods from these works can be generally categorized into two kinds of approaches: tracking by detection, and temporal tracking.

The tracking by detection approaches treat tracking as a detection problem, which rely on features only from the current image without using information from previous frames. For example in electro-

physiology procedure, as the catheters present specific features in shape or intensity, *ad hoc* methods were proposed based on, e.g. blob detection, shape constrained searching and model-/template- based detection (Ma et al. (2012, 2013)). Chang et al. (2016) modeled the catheter tracking problem by optimizing the posterior in a Bayesian framework, in which the catheter was represented as a B-spline tube model and was tracked by fitting the B-spline to measurements based on gray intensity and vesselness image. Baur et al. (2016) proposed a convolutional neural network (CNN) to detect catheter electrodes in X-ray images, which treated catheter detection as a segmentation problem. The method used a weighted cross-entropy loss to cope with the class imbalancing problem due to the small size of the target. Laina et al. (2017) and Du et al. (2018) tracked surgical instruments using a deep network having an encoder-decoder architecture. Their approaches combined instrument segmentation and detection in a multi-task learning problem to make the tool detection in a cluttered background more robust.

Different from tracking by detection, which relies solely on the current image, temporal tracking also uses information from previous frames. The temporal information can reduce the search space for detection, or put additional constraints in the model, making the tracking more robust.

Temporal information has been used in various ways. Some methods mainly relied on a detection model, but incorporate temporal information in the preprocessing (Brost et al. (2010)) or post-processing (García-Peraza-Herrera et al. (2016)) step or in the input (Rieke et al. (2016); Ambrosini et al. (2017a)). Approaches based on background estimation have been used for catheter (Yatziv et al. (2012)) or guidewire (Petković and Lončarić (2010)) tracking. In these approaches, the background was recursively updated for every frame, and was used for enhancing the foreground containing instruments. Apart from those, many works adopted a Bayesian framework for tracking instruments via a *maximum a posteriori* (MAP) formulation. Representations based on key points (Wu et al. (2015)), B-splines (Wang et al. (2009); Pauly et al. (2010); Honnorat et al. (2011); Heibel et al. (2013)), or segment-like features (Vandini et al. (2017)) have been used to

model catheters or guidewires. Markov random field (MRF) was used to model the position or deformation of the control points in the B-spline (Pauly et al. (2010); Honnorat et al. (2011); Heibel et al. (2013); Wu et al. (2015)). In the work by Vandini et al. (2017), temporal information was incorporated in the prior term using Kalman filter. Particularly, learning-based approaches were used in several works to obtain the likelihood for a more robust measurement using probabilistic boosting tree (Wang et al. (2009); Wu et al. (2012)) or support vector regression (Pauly et al. (2010)). In addition, temporal tracking models based on Bayesian filtering were also a popular approach for instrument tracking. Ambrosini et al. (2017b) used a hidden Markov model (HMM) to track catheter tip in a 3D vessel tree, for which the likelihood was obtained based on the 3D-2D registration outcome. Speidel et al. (2006) used particle filters to track surgical tools in medical images. They used a likelihood based on the segmentation of instruments, and a dynamic model that incorporates samples from two previous time steps. In a later work, Speidel et al. (2014) used a multi-object particle filter to track multiple instrument regions simultaneously, in which a particle is the concatenation of the states of several objects.

Despite of many existing works on inverventional or surgical tool tracking in medical images, an automatic approach for tracking the tip of guiding catheter in X-ray fluoroscopy for PCI has not been investigated yet. The challenges of this task are: (1) different from the catheters for EP that can be viewed as blobs or a circle, the guiding catheter for PCI presents a dark tubular appearance which shows no prominent features; (2) the shape of the guiding catheter tip segment varies depending on the orientation of the C-arm, making feature-/model- based detection challenging; (3) the background may contain structures that have similar appearance to a catheter tip, such as vertebral structures or residual contrast agent, which makes robust tracking difficult.

1.4. Contributions

We propose and evaluate a novel approach for dynamic coronary roadmapping. The approach compensates changes in vessel shapes and cardiac motion by selecting the roadmap of the same cardiac phase through ECG alignment, and corrects the respiratory induced motion via tracking the tip of the guiding catheter. Our contributions are:

1. We develop a new way to perform dynamic coronary roadmapping on free breathing, non-cardiac-gated X-ray fluoroscopic sequences. Particularly, the respiratory-induced vessel motion is robustly compensated via the displacement of catheter tip.
2. We proposed a deep learning based method within a Bayesian filtering framework for online detection and tracking of guiding catheter tip in X-ray fluoroscopic images. The method models the likelihood term of Bayesian filtering with a convolutional neural network, and integrates it with particle filtering in a comprehensive manner, leading to more robust tracking.
3. We evaluate the proposed approach visually and quantitatively on clinical X-ray sequences, achieving low errors on both tracking and roadmapping tasks.
4. The proposed DCR method runs in real-time with a modern GPU, thus can potentially be used during PCI in real clinical settings.

2. Scenario Setup and Method Overview

The proposed method assumes that the scenario of performing dynamic coronary roadmapping to guide a PCI procedure consists of an offline phase and an online phase. An overview of the method is shown in Figure 1.

2.1. Offline Phase

This phase (Step 0 in Figure 1) is performed offline before the actual roadmapping is conducted. In this stage, roadmaps of coronary arteries containing multiple cardiac phases are created from an X-ray angiography sequence acquired with injection of contrast agent. A roadmap can be a vessel model in the form of centerlines, contours, masks, etc. It may also contain information of clinical interest, e.g. stenosis. Since the main focus of this paper is on accurate overlay of a roadmap, we do not investigate how to create the most suitable roadmaps, but use the images containing only vessels and catheters that
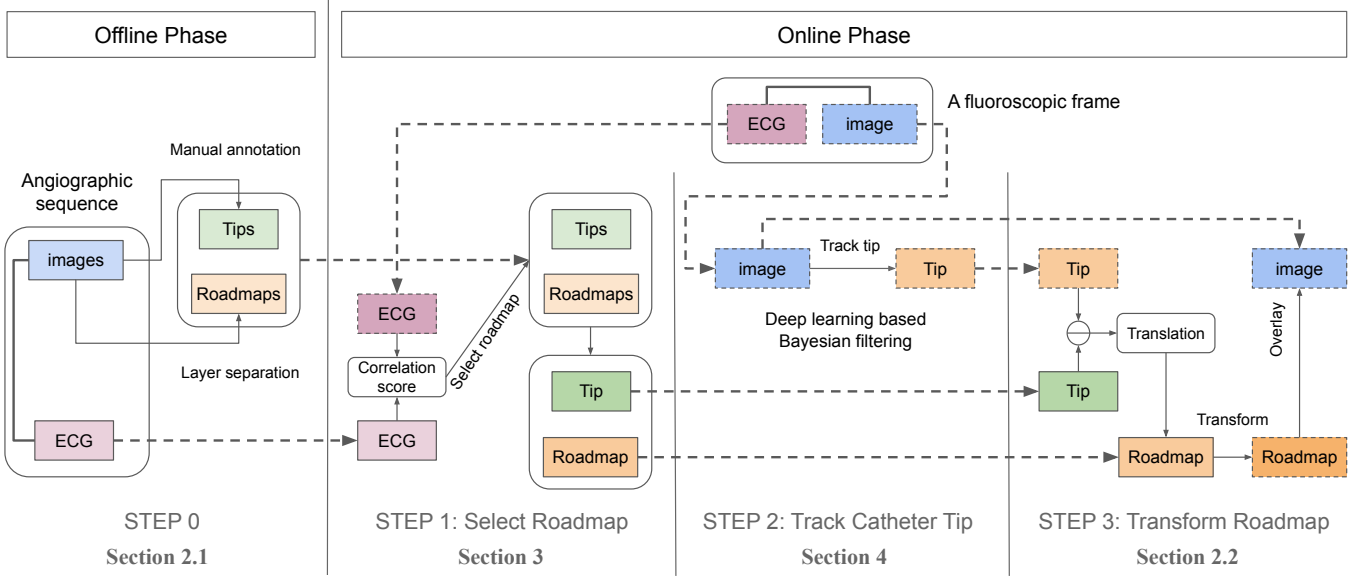
Figure 1: The overview of the proposed dynamic coronary roadmapping method. The colored blocks with a dash line border denote objects acquired in the online phase; the colored blocks with a solid line border are objects originated from the offline phase.

are created using the layer separation method by Ma et al. (2015) as the roadmaps to show the concept of dynamic coronary roadmapping. Along with the XA sequence, ECG signals are also acquired and stored for later selecting a roadmap that has similar cardiac phase to a given X-ray fluoroscopic frame in the online phase (see details in Section 3). Once the image sequence and ECG signals are acquired, the catheter tip location in every frame is obtained to serve as a reference point for roadmap transformation. In this work we manually annotated the catheter tip in the offline XA sequence. In real clinical scenarios, the annotation work can be done by the clinician or a person who assists the intervention, such as a technician or a nurse.

2.2. Online Phase

This is when the dynamic roadmapping is actually performed. In this phase, non-contrast X-ray fluoroscopic images with the same view angles as the roadmaps created during the offline phase are acquired sequentially. At the same time, ECG signals along with the roadmapping frames are also obtained and are compared with the stored ECG to select the most matched roadmap (Step 1 in Figure 1; see details in Section 3). This is to compensate the change of vessel shape and position between frames due to cardiac motion. Simultaneously, the catheter tip

location in the acquired X-ray fluoroscopic images is tracked online using the proposed deep learning based Bayesian filtering method in Section 4 (Step 2 in Figure 1). The displacement of catheter tip between the current image and the selected roadmap image is then obtained and are applied to transform the roadmap. Finally, the transformed roadmap is overlaid on the current non-contrast frame to guide the procedure (Step 3 in Figure 1).

3. ECG matching for Roadmap Selection

Roadmap selection in this work is achieved by comparing the ECG signal associated with the fluoroscopic image and the ECG of the angiographic sequence, such that the most suitable candidate roadmap is selected where the best match of the ECG signals is found. The selected roadmap has the same (or very similar) cardiac phase with the X-ray fluoroscopic image, which compensates the difference of vessel shape and pose induced by cardiac motoin. An approach similar to the ECG matching method by Kim et al. (2018) is used to accomplish this task.

To select roadmaps images based on ECG, a temporal mapping between X-ray images and ECG signal points needs to be built first. We assume that ECG signals and X-ray images are well synchronized during acqusition. In the offline phase, the begin-

5

ning and the end of the image sequence are aligned with the start and end ECG signal points; the XA frames in between are then evenly distributed on the timeline of ECG. This way, a mapping between the stored sequence images and its ECG signal can be set up: for each image, the closest ECG signal point to the location of the image on the timeline can be found; for each ECG point, an image that is closest to this point on the timeline can be similarly located. Once the mapping is available, all images with good vessel contrast filling and the ECG points that are associated to these images are selected from the XA sequence for the pool of roadmaps. In this process, at least one heartbeat of frames should be acquired, which is generally the case in our data. In the online phase, similar to the approach of Kim et al. (2018), for acquisition of each image, a block of N_{ECG} latest ECG signal points is constantly stored and updated in the history buffer. This is considered as the ECG signal corresponding to the fluoroscopic frame.

To compare the ECG signals associated with the angiographic sequence and the online fluoroscopic image, a temporal registration of the two signals using cross-correlation is applied (Kim et al. (2018)). The two ECG signals are first cross-correlated for every possible position on the signals, resulting in a 1D vector of correlation scores. The candidate frame for dynamic overlay is then selected as the one associated with the point on the ECG of the angiographic sequence that is corresponding to the highest correlation score.

4. Bayesian Filtering for Catheter Tip Tracking

Bayesian filtering is a state-space approach aiming at estimating the true state of a system that changes over time from a sequence of noisy measurement made on the system (Arulampalam et al. (2002)). One popular application area of this approach is tracking objects in a series of images.

4.1. Theory of Bayesian Filtering

Bayesian filtering typically includes the following components: hidden system states, a state transition model, observations and a observation model. Let $\mathbf{x}_k \in \mathbb{R}^2$ ($k = \{0, 1, 2, ...\}$) denote the state, the location of guiding catheter tip in the k-th frame, a 2D

vector representing the coordinates in the X-ray image space. The transition of the system from one state to the next state is given by the state transition model $\mathbf{x}_k = f_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1})$, where $\mathbf{v}_{k-1} \in \mathbb{R}^2$ is an independent and identically distributed (i.i.d.) process noise, $f_k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ is a possibly nonlinear function that maps the previous state $\mathbf{x}_{k-1}$ to the current state $\mathbf{x}_k$ with noise $\mathbf{v}_{k-1}$. The observation $\mathbf{z}_k$ in this work is defined as the k-th X-ray image of a sequence, so that $\mathbf{z}_k \in \mathbb{R}^{w \times h}$, where w and h are the width and height of an X-ray image. We further define the observation model as $\mathbf{z}_k = h_k(\mathbf{x}_k, \mathbf{n}_k)$, where $\mathbf{n}_k \in \mathbb{R}^{n_k}$ is an i.i.d measurement noise (n_k is the dimension of $\mathbf{n}_k$), $h_k : \mathbb{R}^2 \times \mathbb{R}^{n_k} \to \mathbb{R}^{w \times h}$ is a highly nonlinear function that generates the observation $\mathbf{z}_k$ from the state $\mathbf{x}_k$ with noise $\mathbf{n}_k$. The state transition model f_k and the observation model h_k, respectively, can also be equivalently represented using probabilistic forms, i.e. the state transition prior $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ from which $\mathbf{x}_k$ and $\mathbf{z}_k$ can be obtained by sampling.

With these definitions and $p(\mathbf{x}_0)$, the inital belief of $\mathbf{x}_0$, Bayesian filtering seeks an estimation of $\mathbf{x}_k$ ($k \geq 1$) based on the set of all available observations $\mathbf{z}_{0:k} = \{\mathbf{z}_i, i = 0, ..., k\}$ up to time k via recursively computing the posterior probability $p(\mathbf{x}_k|\mathbf{z}_{0:k})$ as Eq.(1) (Arulampalam et al. (2002)):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k}) \propto p(\mathbf{z}_k|\mathbf{x}_k) \underbrace{\int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{0:k-1})\mathrm{d}\mathbf{x}_{k-1}}_{p(\mathbf{x}_k|\mathbf{z}_{0:k-1})} .$$

$$(1)$$

Assuming the initial probability $p(\mathbf{x}_0|\mathbf{z}_0) = p(\mathbf{x}_0)$ is known, based on Eq.(1), Bayesian filtering runs in cycles of two steps: prediction and update. In the prediction step, the prior probability $p(\mathbf{x}_k|\mathbf{z}_{0:k-1})$, the initial belief of $\mathbf{x}_k$ given previous observations, is estimated by computing the integral in Eq.(1). In the update step, the prior probability is corrected by the current likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ to obtain the posterior $p(\mathbf{x}_k|\mathbf{z}_{0:k})$.

In Section 4.2, we will firstly introduce how to model the likelihood. Then in Section 4.3, a way to represent and efficiently approximate the posterior will be discussed. Finally in Section 4.4, a summary of the complete catheter tip tracking method will be given.

4.2. A Deep Learning based Likelihood

Directly modeling the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ is challenging due to (1) the complexity of the generation process h_k and (2) the computational complexity of $p(\mathbf{z}_k|\mathbf{x}_k)$ for every value $\mathbf{x}_k \in \mathbb{R}^2$. In this work, we simplify the problem by only computing the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ in the image pixel space, i.e. the integer pixel coordinate. For a subpixel $\mathbf{x}_k$, the value of $p(\mathbf{z}_k|\mathbf{x}_k)$ can possibly be approximated by interpolation. To this end, we propose to use a deep neural network $\mathcal{D}$ to approximate $p(\mathbf{z}_k|\mathbf{x}_k)$ for integer pixel locations. The network takes an image $\mathbf{z}_k$ as input and outputs a probability of observing the input $\mathbf{z}_k$ for every pixel location $\mathbf{x}_k$. Therefore, the approximated likelihood is a function of $\mathbf{x}_k$, denoted as $\mathcal{D}_{\mathbf{z}_k}(\mathbf{x}_k)$. Since $\mathbf{x}_k$ is defined within the scope of the image pixel space, $\mathcal{D}_{\mathbf{z}_k}(\mathbf{x}_k)$ is essentially a probability map having the same dimension and size with the input image $\mathbf{z}_k$, in which the entry at each location $\mathbf{x}_k^j$ ($j = 1, 2, \ldots, wh$) in the map represents the probability of observing $\mathbf{z}_k$ given $\mathbf{x}_k^j$. It is worth mentioning that the deep neural network is used for approximation of $p(\mathbf{z}_k|\mathbf{x}_k)$, which should be clearly distinguished from the generation model h_k that maps an $\mathbf{x}_k$ to $\mathbf{z}_k$. The existence of h_k is merely for the convenience of definition, its explicit form, however, is not required in the context of this work.

To obtain the training labels, we assume that there exists a mapping h_k, such that the training label can be defined as a distance-based probability map, i.e. the farther away $\mathbf{x}_k$ is from the ground truth tip location in the image $\mathbf{z}_k$, the less possible it is to observe $\mathbf{z}_k$ given $\mathbf{x}_k$ through the process h_k. This definition matches the intuition that from a location $\mathbf{x}_k$ that is far from the ground truth tip location, the probability of observing a $\mathbf{z}_k$ with the catheter tip being located at the ground truth position should be low. For simplicity, a 2D Gaussian probability density function (PDF) $\mathcal{N}(\mathbf{x}_k; \mathbf{x}_k', \sigma^2 I)$ centered at the ground truth tip location $\mathbf{x}_k'$ with variance $\sigma^2 I$ in the image space is used as the label to train the network (Figure 2c). Note that this training label makes the estimation of $p(\mathbf{z}_k|\mathbf{x}_k)$ equivalent to a catheter tip detection problem such that the deep neural network learns features of catheter tip and outputs high probability at locations where the features are present. Due to this reason, we also call $p(\mathbf{z}_k|\mathbf{x}_k)$ "detection output" or "detec-
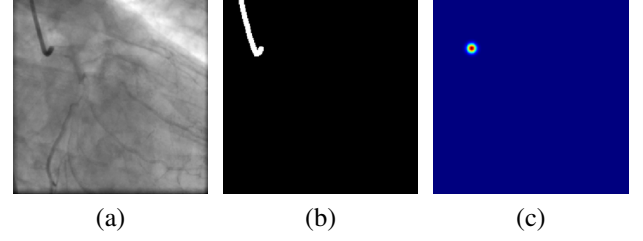

Figure 2: Input and ground truth labels for the deep neural network: (a) an input X-ray fluoroscopic image, (b) the binary catheter mask of (a) for catheter segmentation, (c) a 2D Gaussian PDF ($\sigma = 4$ px) for likelihood estimation for (a).

tion probability" and call the estimation of $p(\mathbf{z}_k|\mathbf{x}_k)$ "catheter tip detection" in the context of this paper.

The network that we use follows a encoder-decoder architecture with skip connections similar to U-net (Ronneberger et al. (2015)). Additionally, similar to the work by Milletari et al. (2016), residual blocks (He et al. (2016)) are adopted at each resolution level in the encoder and decoder to ease gradient propagation in a deep network. The encoder consists of 4 *down* blocks in which a residual block followed by a stride-2 convolution is used for extraction and down-scaling of feature maps. The number of feature maps is doubled in each downsampling step. The decoder has 4 *up* blocks where a transposed convolution of stride-2 is used for upsampling of the input feature maps. Dropout is used in the residual unit of the *up* block for regularization of the network. Between the encoder and the decoder, another residual block is used to process the feature maps extracted by the encoder. The detailed network architecture is shown in Figure 3.

Due to similar appearance between a guiding catheter tip and corners of a background structure, such as vertebral bones, lung tissue, stitches or guidewires, ambiguity may exist when the network is expected to output only one blob in the probability map. To alleviate the issue, we adopt a similar strategy as Laina et al. (2017), using a catheter mask (Figure 2b) as an additional label to jointly train the network to output both the catheter segmentation heatmap and the likelihood probability map. The segmentation heatmap is obtained by applying a 1×1 convolution with ReLU activation on the feature maps of the last *up* block. To compute the likelihood
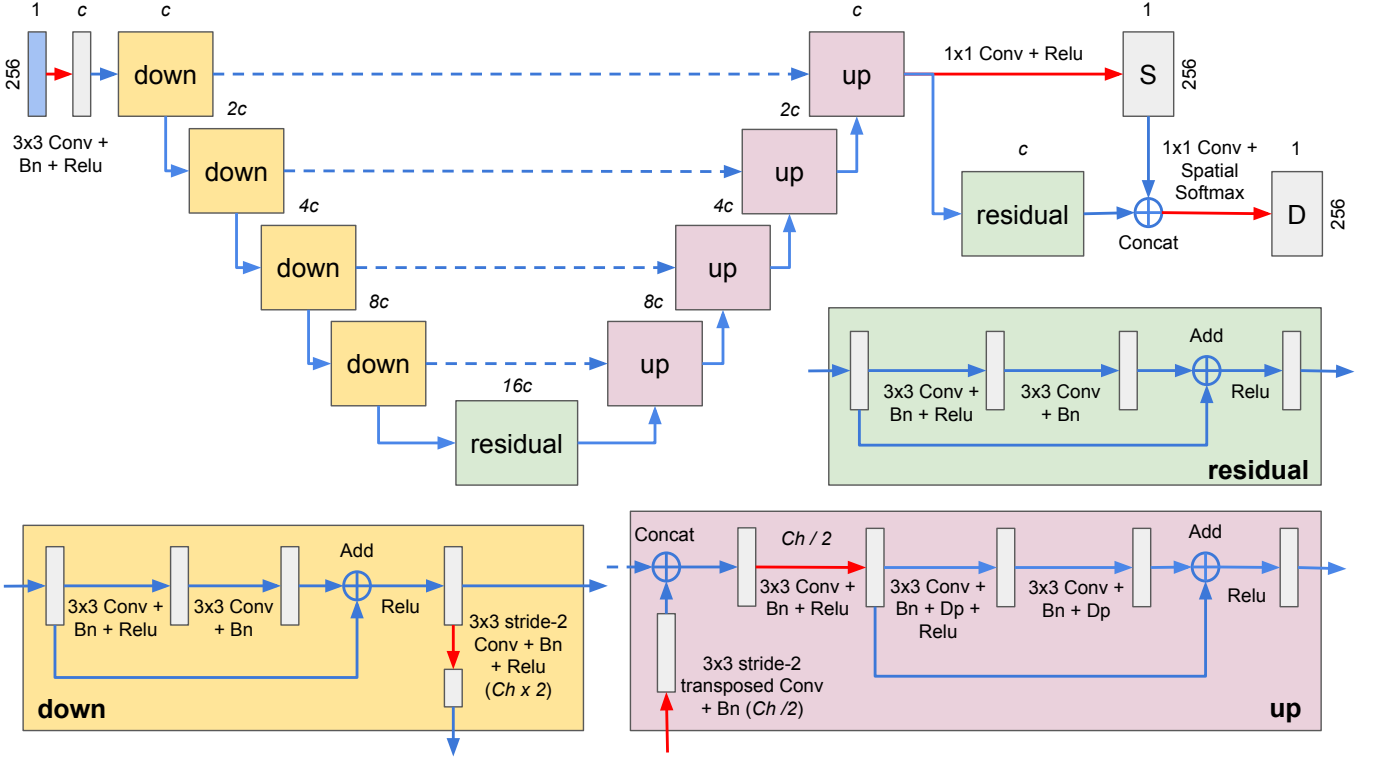
Figure 3: A joint segmentation and detection network for catheter tip detection. This figure shows an example network with 4 levels of depth (the number of down or up blocks). Meaning of abbreviations: *Conv*, 2D convolution; *Bn*, batch normalization; *Relu*, ReLU activation; *Dp*, dropout; *Concat*, concatenation; *Ch*, number of channels; *S*, segmentation output; *D*, detection output. The number above an image or feature maps indicates the number of channels; the number of channels in the residual network in a block is shown above the block; *c* is the basic number of channels, the channel number in the first down block. The number next to a rectangle denotes the size of the image or feature maps. Red arrows indicate a change of number of channels.

probability map, a residual block is firstly applied on the feature maps of the last *up* block. The output feature maps are then concatenated with the segmentation heatmap as one additional channel, followed by a 1×1 convolution. Finally, to ensure the network detection output fits the definition of a probability map on image locations, following the 1×1 convolution, a spatial softmax layer is computed as Eq.(2):

$$D_{k,l} = \frac{e^{A_{k,l}}}{\sum_{i,j} e^{A_{i,j}}}, \qquad (2)$$

where A is the output feature map of the 1×1 convolution, $A_{i,j}$ denotes the value of A at location (i, j), D is the final output of the detection network, a 2D probability map representing $p(\mathbf{z}_k|\mathbf{x}_k)$. The details are shown in Figure 3.

The training loss is defined as a combination of the segmentation loss and the detection loss. The segmentation loss L_s in this work is a Dice loss defined by Eq.(3):

$$L_s = 1 - \frac{2 \sum_{i,j} M_{i,j} S_{i,j}}{\sum_{i,j} M_{i,j}^2 + \sum_{i,j} S_{i,j}^2} \qquad (3)$$

where M denotes the ground truth binary catheter masks, S is the segmentation heatmap. The loss function for detection L_d is mean square error (MSE) given by Eq.(4):

$$L_d = \frac{1}{w \times h} \sum_{i \leq w, j \leq h} |T_{i,j} - D_{i,j}|^2 \qquad (4)$$

where T denotes the ground truth PDF, w and h are the width and height of an image. The total training loss L is defined as Eq.(5):

$$L = L_s + \lambda L_d \qquad (5)$$

where λ is a weight to balance L_s and L_d.

4.3. Approximation of the Posterior with Particle Filter

Once the deep neural network in Section 4.2 is trained, its weights are fixed during inference for computing the posterior $p(\mathbf{x}_k|\mathbf{z}_{0:k})$ for new data. Ideally, the network detection output $p(\mathbf{z}_k|\mathbf{x}_k)$ should be a Gaussian PDF during inference, as it is trained with labels of Gaussian PDFs. However, due to similar appearance of background structures or contrast residual, the detection output is unlikely to be a perfect Gaussian (possibly non-Gaussian or having multiple modes), which prevents the posterior $p(\mathbf{x}_k|\mathbf{z}_{0:k})$ in Eq.(1) being solved with an analytical method. In practice, the posterior can be approximated using a particle filter method (Arulampalam et al. (2002)).

Particle filter methods approximate the posterior PDF by a set of N_s random samples with associated weights $\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}$ (Arulampalam et al. (2002)). As N_s becomes very large, this discrete representation approaches the true posterior. According to Arulampalam et al. (2002), the approximation of the posterior $p(\mathbf{x}_k|\mathbf{z}_{0:k})$ is given by Eq.(6):

$$p(\mathbf{x}_k|\mathbf{z}_{0:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{6}$$

where $\delta(\cdot)$ is the Dirac delta function. The weight w_k^i can be computed in a recursive manner as Eq.(7) once w_{k-1}^i is known (Arulampalam et al. (2002)):

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)} \tag{7}$$

where $q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ is an importance density from which it should be possible to sample $\mathbf{x}_k^i$ easily. For simplicity, a good and convenient choice of the importance density is the prior $p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$ (Arulampalam et al. (2002)), so that the weight update rule (7) becomes $w_k^i \propto w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i)$.

A sample can be drawn from $p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$ in the following way. First, a process noise sample $\mathbf{v}_{k-1}^i$ is sampled from $p_v(\mathbf{v}_{k-1})$, the PDF of $\mathbf{v}_{k-1}$; then $\mathbf{x}_k^i$ is generated from $\mathbf{x}_{k-1}^i$ via the state transition model $\mathbf{x}_k^i = f_k(\mathbf{x}_{k-1}^i, \mathbf{v}_{k-1}^i)$. In this work, $p_v(\mathbf{v}_{k-1})$ is set to be a Gaussian $\mathcal{N}(\mathbf{0}, \sigma_v^2 I)$. The choice of motion model for f_k is important for an accurate representation of the true state transition prior $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. A random motion cannot characterize well the motion of catheter

tip in XA frames. In this work, we estimated the motion from adjacent frames using an optical flow method, as this approach 1) takes into account of the observation $\mathbf{z}_k$, which results in a better guess of the catheter tip motion, and 2) enables estimation of a dense motion field where the motion of a sample $\mathbf{x}_k^i$ can be efficiently obtained. Therefore, f_k is defined as Eq.(8):

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_{k-1}(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1} \tag{8}$$

where $\mathbf{u}_{k-1}(\cdot)$ is the motion from frame $k-1$ to frame k estimated with optical flow using the method by Farnebäck (2003), $\mathbf{u}_{k-1}(\mathbf{x}_{k-1})$ is the motion from state $\mathbf{x}_{k-1}$.

Once samples are drawn and their weights are updated, the so-called "resampling" of the samples should be performed to prevent the degeneracy problem, where all but one sample will have negligible weight after a few iterations (Arulampalam et al. (2002)). The resampling step resamples the existing samples according to their updated weights and then resets all sample weights to be $1/N_s$, so the number of effective samples which have actual contribution to approximate $p(\mathbf{x}_k|\mathbf{z}_{0:k})$ is maximized (Arulampalam et al. (2002)). If the resampling is applied at every time step, the particle filter becomes a sampling importance resampling (SIR) filter, and the weight update rule follows Eq.(9).

$$w_k^i \propto p(\mathbf{z}_k|\mathbf{x}_k^i) \tag{9}$$

The final decision on catheter tip location in frame k can then be computed as the expectation of $\mathbf{x}_k$, $\hat{\mathbf{x}}_k = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{z}_{0:k}) d\mathbf{x}_k$, which is in this case, the weighted sum of all samples:

$$\hat{\mathbf{x}}_k = \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i. \tag{10}$$

4.4. Summary

The overall catheter tip tracking using a deep learning based Bayesian filtering method is summarized in Algorithm 1.

Algorithm 1 Deep learning based Bayesian filtering for online tracking of catheter tip in X-ray fluoroscopy
---
Require: $\{\mathbf{z}_0, \ldots, \mathbf{z}_T\}$ (sequentially observed frames), $\mathcal{D}$ (A trained network from Section 4.2), $p(\mathbf{x}_0)$ (the initial PDF), σ_v^2 (the variance of $\mathbf{v}_{k-1}$, $k = 1, \ldots, T$), T (number of frames for tracking), N_s (number of samples)

1: Draw $\mathbf{x}_0^i \sim p(\mathbf{x}_0)$, set $w_0^i = 1/N_s$, $\forall\, i = 1, \ldots, N_s$
2: **for** $k = 1$ **to** T **do**
3:     Compute $\mathbf{u}_{k-1}$ from $\mathbf{z}_{k-1}$ to $\mathbf{z}_k$ using the optical flow method of Farnebäck (2003)
4:     **for** $i = 1$ **to** N_s **do**
5:         Draw $\mathbf{v}_{k-1}^i \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 I)$
6:         Compute the motion of $\mathbf{x}_{k-1}^i$: $\mathbf{u}_{k-1}^i = \mathbf{u}_{k-1}(\mathbf{x}_{k-1}^i)$
7:         Draw $\mathbf{x}_k^i \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$: $\mathbf{x}_k^i = \mathbf{x}_{k-1}^i + \mathbf{u}_{k-1}^i + \mathbf{v}_{k-1}^i$
8:         Update weight $w_k^i = p(\mathbf{z}_k|\mathbf{x}_k^i) = \mathcal{D}_{\mathbf{z}_k}(\mathbf{x}_k^i)$
9:     **end for**
10:    Normalize $w_k^i \leftarrow w_k^i / \sum_{i=1}^{N_s} w_k^i$, $\forall\, i = 1, \ldots, N_s$
11:    Prediciton in frame k: $\hat{\mathbf{x}}_k = \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i$
12:    Resample $\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}$ using the method of Arulampalam et al. (2002) (so all w_k^i are set to $1/N_s$ again)
13: **end for**
---

5. Experimental Setup

5.1. Data

Anonymized clinical imaging data were used for our experiments. The data were acquired with standard clinical protocol using Siemens AXIOM-Artis system, and are from 55 patients who underwent a PCI procedure at the Department of Cardiology at Erasmus MC in Rotterdam, Netherlands. Out of these data, we selected data from 37 patients which were acquired since the year 2014 to develop our method, and used the data from the other 18 patients acquired before the year 2013 for evaluation. The detailed information about the data is listed in Table 1.

In order to evaluate the proposed roadmapping method, for which an off-line angiographic sequence is required for roadmap preparation and an online fluoroscopic sequence taken from the same C-arm position is needed for performing the actual roadmapping (see Section 2), we selected the contrast frames from a real clinical sequence to simulate the off-line sequence, and chose the non-contrast frames from the same clinical sequence to simulate the online sequence. The selected contrast sequence were ensured sufficiently long to cover at least one complete cardiac cycle.

Table 1: Basic information of the acquired X-ray image data for our experiments. The number in the parenthesis next to the pixel size indicates the possible image size.

Data	Development	Evaluation
No. patients	37	18
No. sequences	354	34
Frame rate (fps)	15	15
Image size (px)	512×512	512×512
	600×600	600×600
	776×776	776×776
	960×960	1024×1024
	1024×1024	
Pixel size (mm)	0.108 (1024)	0.139 (1024)
	0.139 (1024)	0.184 (600)
	0.184 (600)	0.184 (776)
	0.184 (776)	0.184 (1024)
	0.184 (960)	0.216 (512)
	0.184 (1024)	0.279 (512)
	0.216 (512)	

5.2. Data Split for Catheter Tip Detection and Tracking

To develop the catheter tip tracking method, 1086 X-ray fluoroscopic images selected from 260 non-contrast sequences of 25 patients from the development set were used for training the network from Figure 3; 404 images from 94 non-contrast sequences of another 12 patients from the development set were used as validation set for the network model and hyperparameter selection. In the training and validation sets, 4-5 frames were randomly selected from each sequence, which are not necessarily continuous. To tune the parameters for tracking, 1583 images from 88 sequences out of the 94 from the same 12 patients of the validation set were used (6 sequences were not selected for this task due to very short sequence length not more than 5 frames). Finally, to evaluate catheter tip tracking accuracy, 1355 images from 34 non-contrast sequences of 18 patients from the evaluation set were used for testing. The frames selected for tracking from each sequence must be continuous; the number of selected frames for tracking might vary, depending on the number of the non-contrast frames in the sequences. Details of the datasets for training, validation and test are listed in Table 2.

5.3. Experimental Settings for Training the Deep Network

This section describes the basic experimental settings for training the deep neural network. Details of the training setup can be found in Appendix A.

5.3.1. Preprocessing

As the image data have different size ranging from 512×512 to 1024×1024, all images were resampled to a grid of 256×256 before being processed by the neural network. In addition, the image intensities were rescaled to the range from 0 to 1.

5.3.2. Training label

The standard deviation σ of the Gaussian PDF for the training label of the detection network was set to 4 pixels in the resampled image space (256×256). This choice corresponds to the estimation of the maximal possible catheter tip radius. An example of the Gaussian PDF is shown in Figure 2c.

5.3.3. Evaluation Metric

To select hyperparameters and model weights in training, an evaluation metric is required. As the deep network is essentially a catheter tip detector, accurate detection of the tip location is desired. Therefore, we chose the location with the highest value in the detection output, and computed the Euclidean distance between the chosen location and the ground truth tip coordinate as the evaluation metric to tune the deep network.

5.4. Setup for Evaluating Dynamic Coronary Roadmapping

It is in general a challenge to evaluate the roadmapping accuracy, as the structure of interest, e.g. coronary arteries in our case, is not directly visible in the target image. One possible choice introduced by Zhu et al. (2010) is to use the guidewire as a surrogate of the target vessel centerline in non-contrast images, as guidewire is always inside vessels and commonly present in image sequences during interventions. In this work, we follow a similar strategy to evaluate the accuracy of dynamic coronary roadmapping.

The first step is to select frames for roadmapping evaluation. From each non-contrast sequence in the test set for tracking in Section 5.2, we uniformly select 8-20 frames to annotate guidewire. The number of the selected frames from each sequence depends on the sequence length, the frame interval size and guidewire visibility. For some rare cases in our data where no guidewire is present in the image, we discarded that non-contrast frame, and chose those frames with little vessel contrast from the same sequence and annotated the vessel centerline. The selection results in 409 frames from 34 sequences in total. Once the target non-contrast frames for evaluating roadmapping are chosen, their corresponding angiographic frames were found using the ECG matching method in Section 3. We then annotated the centerline of the vessel corresponding to the guidewire in the non-contrast frames.

The next step is performing the transformation of the labelled vessel centerline from the angiographic frame to its corresponding target non-contrast frame via displacement of catheter tip in the two frames.

Table 2: Dataset of training, validation and test for detection and tracking of catheter tip in X-ray fluoroscopic frames.

	Training (detection)	Validation (detection)	Validation (tracking)	Test (tracking)
No. patients	25	12	12	18
No. sequences	260	94	88	34
No. frames	1086	404	1583	1355
Continous frames?	No	No	Yes	Yes

This step simulates the roadmapping transformation in the last step in Figure 1.

Finally, the distance between the guidewire annotation in the target frame and the transformed vessel centerline is reported as the roadmapping accuracy. In order to compute the distance between two point sets of annotations (e.g. Figure 4a), point-point correspondence between the two sets is required (Figure 4b). The point sets were firstly resampled with the point interval being 1 mm. We then followed the approach of van Walsum et al. (2008) to find such correspondences which minimizes the sum of the Euclidean distance of all valid point-point correspondence paths. This way guarantees no cross-over connection and each point in one set is connected to at least one point in the other set. As the annotated point sets may have different size, the point correspondences to endpoints are excluded such that we only focused on the distance between corresponding sections, not the entire centerlines (Figure 4c). Once the point-point correspondence is available, the distance between the two points in a pair can be used for evaluating the accuracy of DCR.
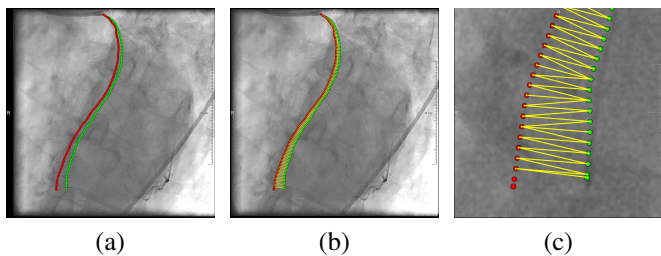

(a)　　　　(b)　　　　(c)

Figure 4: Correspondence between the labelled guidewire (green) and the transformed vessel centerline (red). The yellow lines connecting the two point sets illustrate the correspondence between red and green points.

5.5. Implementation

The proposed method was developed in Python. The framework used for developing the deep learning approach for likelihood approximation is PyTorch. The major experiments of dynamic coronary roadmapping were performed on a computer with an Intel Xeon E5-2620 v3 2.40 GHz CPU and 16 GB RAM running Ubuntu 16.04. The deep neural network and the optical flow method were running on an NVIDIA GeForce GTX 1080 GPU. The approach for evaluating dynamic coronary roadmapping was developed and running in MeVisLab on a computer with an Intel Core i7-4800MQ 2.70 GHz CPU and 16 GB RAM running Windows 7.

6. Experiments and Results

The following experiments are performed to assess the methods. First, In Section 6.1, the training of the deep neural network is described. Then in Section 6.2, the accuracy of catheter tip tracking using the optimized trained network and the tuned particle filter is presented. Section 6.3 describes the accuracy evaluation of dynamic coronary roadmapping via the proposed catheter tip tracking method. Finally, in Section 6.4, we measure the processing time of the proposed DCR approach.

6.1. Training the Deep Neural Network

The purpose of this experiment is to train the deep neural network to output reasonable likelihood probability map. The network hyperparameters were tuned to optimize the detection performance.

The training and validation data for detection mentioned in Section 5.2 were used for training the deep neural network. The evaluation metric mentioned in Section 5.3, the mean Euclidean distance between

the ground truth and the predicted tip location averaged over all validation frames, was used as the validation criteria for selecting the optimal training epoch and the network hyperparameters. When we evaluated hyperparameter settings, we firstly selected the training epoch with the lowest mean validation error for each setting, then the settings were compared using the model weights (trainable network parameters) of their chosen epochs.

The network hyperparameters we investigated in the experiments include (1) the basic channel number, i.e. the number of channels or feature maps in the first down block, (2) the network depth level, the number of down or up blocks, and (3) the dropout probability.

The validation errors for different hyperparameter settings using the experimental settings in Section 5.3 are shown in Table 3. The table shows that the hyperparameter setting with the lowest mean error, which has 4 level in depth and 64 channels in the first down block, achieves a validation error of about 2 mm. The table also shows other good choices of network architecture that have a small validation error (shown in red in Table 3): 32 channels in the first down block with 4 or 5 levels in depth, or 64 channels with 3 or 4 depth levels. The dropout regularization improves the accuracy of the model, compared to the ones without dropout.

The learning curves of the training process with the chosen hyperparameter setting are illustrated in Figure 5. The curves indicate that both segmentation and detection reach convergence after training 100 epochs.

We did not investigate a model with more than 64 channels or 5 depth levels, because (1) it will further increase the processing time which makes online applications less feasible; (2) the results in Table 3 show that such a setting (64 channels, 5 depth levels) starts increasing the validation error compared to those less complex models.

The subsequent experiments will be based on the network trained with the chosen hyperparameter setting indicated in Table 3 (64 channels, 4 depth levels, dropout 0.2, also see Table 4).
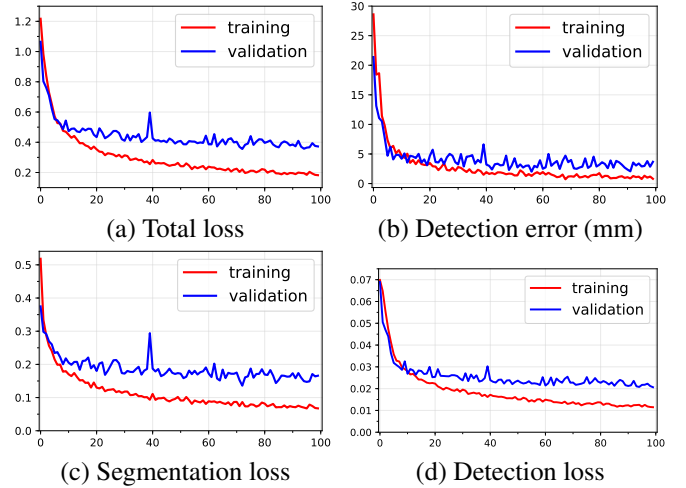

Figure 5: Learning curves for the chosen hyperparameter setting.

6.2. Catheter Tip Tracking

The purpose of this experiment is to assess the accuracy of catheter tip tracking with the proposed method in Section 4. Guiding catheter tip is tracked in X-ray fluoroscopy using Algorithm 1 based on a trained network with the optimal hyperparameter setting from Section 6.1. First, the parameters of the optical flow method used in Algorithm 1 and particle filter were tuned on the validation data for tracking in Section 5.2 (see Appendix B for details). Then in Section 6.2.1, we evaluated the tracking accuracy with the tuned optimal parameter setting (see Table 4) on the test dataset, and compared the proposed tracking method with alternative approaches using only the detection network in Section 4.2 or using only optical flow. Finally, in Section 6.2.2, we investigated tracking accuracy with different ways of tip initialization in the first frame.

6.2.1. Tracking Methods Evaluation

In this experiment, the proposed tracking method in Algorithm 1 uses the ground truth tip probability map of the first frame as the initial PDF $p(\mathbf{x}_0)$ to draw samples. This method is referred to as "Tracking". In addition, we compared the proposed method with three alternatives. The first one tracks catheter tip using only the detection network in Section 4.2 with the chosen network architecture and trained parameters in Section 6.1, therefore, no temporal information is used. This method is referred to as "De-

13

Table 3: Validation errors (mm) for different hyperparameter settings. Red cells show the settings with the 10 smallest validation errors. *bold* number indicates the setting with the lowest error.

Basic Number of Channels	Depth Level	Dropout none	0.1	0.2	0.3	0.4	0.5
8	3	5.43	4.99	5.02	5.37	4.38	4.24
	4	4.17	4.45	4.25	5.04	4.75	4.36
	5	3	4.14	3.53	4.28	3.95	4.11
16	3	3.74	4.29	3.57	4.11	3.74	3.4
	4	3.36	3.11	3.63	3.33	3.36	3.78
	5	3.38	2.89	3.16	2.52	2.71	2.74
32	3	2.99	3.02	3.26	2.82	3.26	2.56
	4	2.87	2.34	2.46	2.6	2.65	2.27
	5	3.04	2.51	2.21	2.29	2.3	2.25
64	3	2.19	2.54	2.34	2.27	2.26	2.49
	4	2.55	2.31	**2.04**	2.44	2.22	2.27
	5	2.42	2.29	2.73	2.77	2.61	2.85

Table 4: The chosen (hyper-)parameters for different building blocks of the catheter tip tracking algorithm. The parameters of the optical flow method can be found in Appendix B.1.

Building block	(hyper-)parameters	value
Deep learning	Basic channel number	64
	Depth	4
	Dropout	0.2
Particle filter	σ_v (px)	5
	N_s	1000

tection (Net)". The other two methods in this experiment use only optical flow to track catheter tip starting from the ground truth tip position in the first frame. The motion field towards the current frame, estimated by the two methods, was based on the deformation from the previous frame or the first frame in the sequence, respectively. The same implementation setting as in Appendix B.1 was used for these two methods. They are called "Optical Flow (previous)" and "Optical Flow (first)", or in short form, "OF (pre)" and "OF (1st)". Additionally, we refer the interested readers to Appendix C.1 where the influence of catheter segmentation on the detection and tracking approaches is reported.

The tracking accuracies of all methods reported in this section were obtained on the test set from Table 2. The mean, the median and the maximal tracking error between the predicted and the ground truth tip position of all test images are reported in Table 5. In addition, as the sequences in the test set have different lengths, we also computed the mean and the median error per sequence, and report the the average of the sequence mean and median errors, so that each sequence contributes equally in these metrics. Table 5 shows that the results from the detection network have large average errors which are caused by some completely failed cases. The proposed tracking method has median errors of about 1 mm and mean errors of about 1.3 mm. It achieves the lowest errors compared to the other 3 methods on all listed evaluation criteria.

Figure 6 illustrates the boxplots of tracking errors made by the 4 methods on all test images. It shows that the proposed tracking approach outperforms the detection method by avoiding making extremely large errors (Figure 6a); meanwhile, it maintains as accurate as the detection method for cases with small errors, and is more accurate than the methods based solely on optical flow (Figure 6b).

Figure 7 shows longitudinal views of tracking errors of the 4 methods on 4 example sequences. Although the optical flow methods show high accuracy when the target is on the track (row 4), they present periodic error patterns in two sequences due

Table 5: Catheter tip tracking errors (mm) of the 4 methods on the test (tracking) dataset. † indicates that the difference between that method and the "Tracking" method are statistically highly significant with the two-sided Wilcoxon signed-rank test ($p < 0.001$).

Evaluation Metrics	Optical Flow† (previous)	Optical Flow† (first)	Detection Net† (Section 4.2)	Tracking
Maximal error of all images	29.16	20.83	108.20	**17.72**
Median error of all images	1.78	1.22	**0.96**	**0.96**
Mean error of all images	3.74 ± 4.93	3.05 ± 4.05	5.62 ± 15.91	**1.29 ± 1.76**
Average of sequence median error	2.35 ± 2.52	2.64 ± 3.52	6.26 ± 17.11	**1.03 ± 0.49**
Average of sequence mean error	2.59 ± 2.69	3.31 ± 2.81	6.83 ± 13.88	**1.29 ± 0.94**
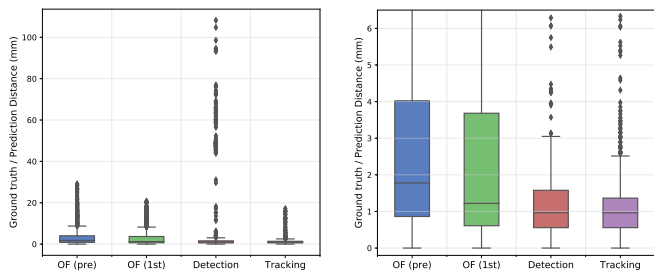
(a) Overall view of tracking errors

(b) A zoom-in view of (a)

Figure 6: Tracking errors for the 4 methods on all test images.

to large cardiac motion. The detection method shows peaks of large errors, this is because temporal relation between frames is not modeled by the approach, thus the detection on different frames is independent of each other. The proposed tracking method overcomes the problems that other methods have and presents accurate detection on these 4 sequences. The tracking results of the 4 methods on example frames from the 4 sequences are illustrated in Figure 8.

Figure 9 illustrates how the proposed tracking method works on the same 4 frames in Figure 8. It shows that the prior hypotheses (samples) assists to focus on the correct target location and results in reliable posterior estimation, especially when the detection produces ambiguity in cases of multiple catheters or contrast residual presented in images.

6.2.2. Catheter Tip Initialization

In this experiment, the initial PDF $p(\mathbf{x}_0)$ from which samples are drawn in the proposed tracking is investigated (Algorithm 1). In particular, we ex-

plored and evaluated the tracking accuracy with an automatic initialization using the probability map obtained from the trained detection network in Section 4.2 with the chosen setting in Section 6.1.

Figure 10 shows the boxplot of tracking errors on all test images with automatic initialization (Auto) and manual initialization (Manual) for which the ground truth tip probability map of the first frame was used. The tracking with automatic initialization presents an accuracy similar to the one with manual initialization for small tracking errors, but has more large tracking errors which influence the mean error over all test images (Table 6). We, therefore, defined the tracking errors on the right side of the gap in the boxplot (> 40 mm) as outliers, and explored the statistics without those outliers.

Table 6 indicates that, the mean and median error of the tracking with automatic initialization excluding the outliers are only slightly higher than the tracking with manual initialization and the detection method. While the tracking with automatic initialization has 100 outliers in total from 6 sequences, the detection method that has 10 sequences containing 106 outliers.

Unlike the detection method for which the outliers are mainly presented as the peaks in the longitudinal views (Figure 7), the outliers for the tracking with automatic initialization are more consistent over time. Figure 11 shows the temporal change of tracking errors for the 6 sequences with outliers using the tracking with automatic initialization. For the 3 sequences on the top row, the tracking with automatic initialization makes large errors at the beginning, but becomes accurate very fast in a few frames; for the 3 sequences on the bottom row, however, the tracking
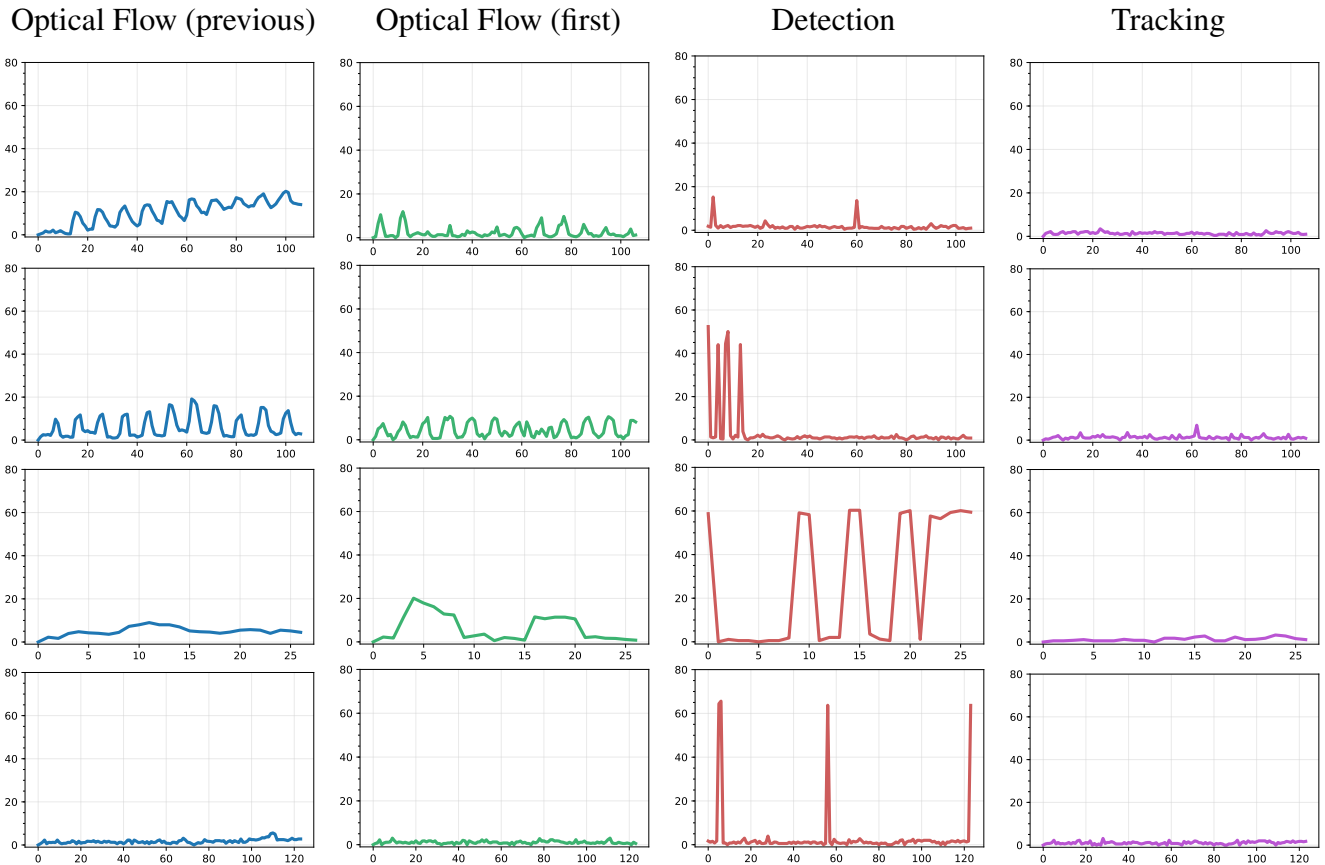
15

Figure 7: Longitudinal view of tracking errors made by the 4 methods on 4 test sequences (one sequence per row). The x-axis denotes the time steps of a sequence, the y-axis is the tracking error (mm).

errors remain large till the end of the sequences.

Figure 12 shows example frames to give an insight of the tracking with automatic initialization on the 6 sequences in Figure 11. For the 3 sequences on the top row (Figure 12a), although the initialization on the first frame (frame 0) is overall not correct, the true tip positions are still covered by some samples; once the detection in subsequent frames is correct, the tracker can still converge to the right target. For the 3 sequence on the bottom row (Figure 12b), the initializations of samples are ambiguous in frame 0; the detection in subsequent frames focuses on a wrong area also given by the initial samples due to residual of contrast agent or multiple catheters, the tracker then tends to find the wrong target.

6.3. Dynamic Coronary Roadmapping

In this experiment, the accuracy of dynamic coronary roadmapping using the proposed method with manual tip initialization was evaluated. For roadmap selection with ECG matching (Section 3), the number of online ECG signal points N_{ECG} was manually determined so that the ECG signal stored in the buffer corresponding to 12 X-ray frames (0.8 second in acquisition time). Following the setup in Section 5.4, we used the distance between the two points in each point pair as the evaluation metric for DCR (the length of a yellow line segment in Figure 4). As each frame may have different numbers of point pairs, depending on the length of the target guidewire, the average point pair distance per frame was also computed for evaluation. These distances were evaluated on 409 selected frames with manual annotation of guidewires and vessel centerlines (Section 5.4).

In the experiment, we compared the DCR with the proposed tracking method to those with manual tip tracking and without tracking. All three approaches were based on the same ECG matching method (Section 3) for selecting roadmaps. The accuracy of the DCR without tracking in Table 7 shows that the mean
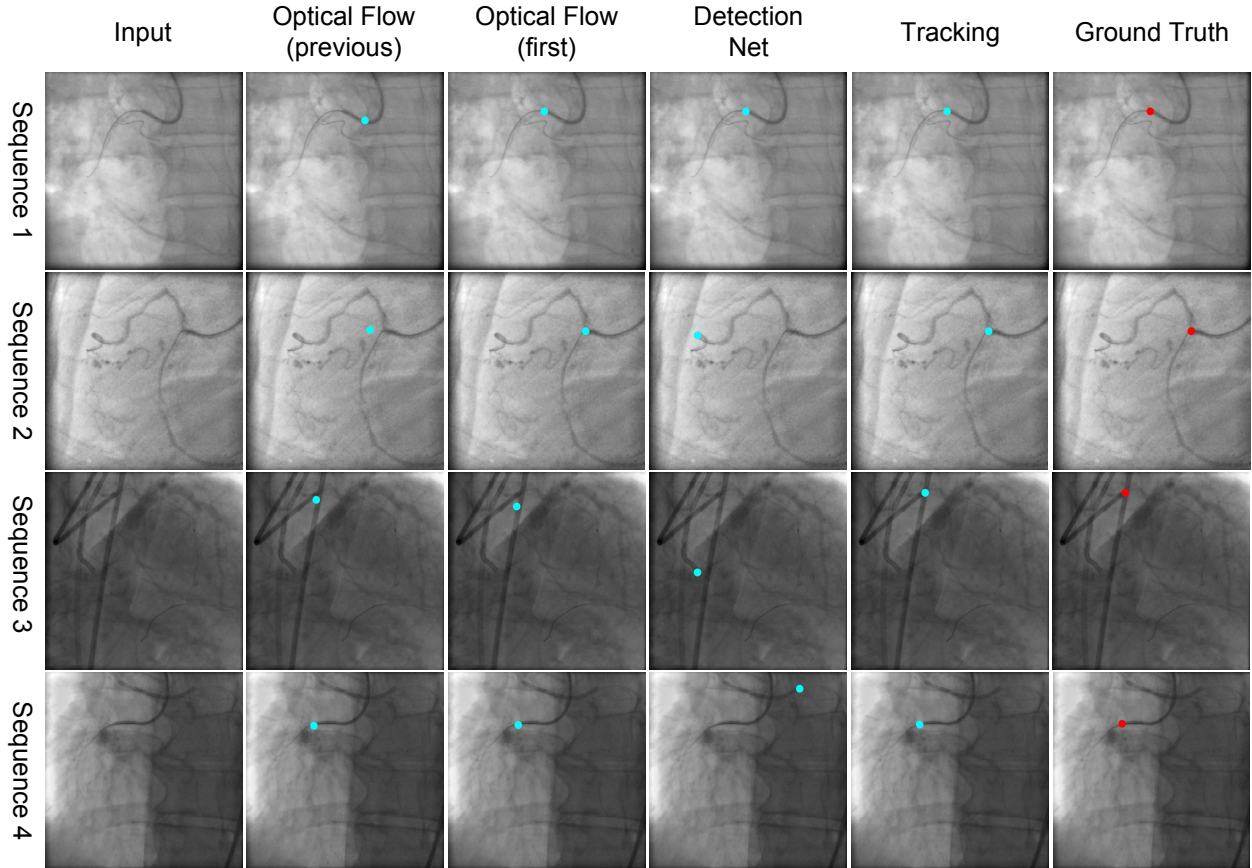
Figure 8: Tracking results on example frames from the same 4 sequences in Figure 7. The blue point indicates the predicted catheter tip location; the red point shows the ground truth location. (Best viewed in color)

distances are reduced to less than 3 mm by compensating only cardiac motion via roadmap selection with ECG matching. Table 7 also shows that the DCR with the proposed method achieves median distances of about 1.4 mm and mean distances of about 2 mm. The boxplots of the distances of all point pairs and the frame mean point distances of all 409 evaluation frames are illustrated in Figure 13. The comparison of the three DCR approaches from Table 7 and Figure 13 indicates that the accuracy of the proposed DCR method has shown improvement over the DCR without tracking, and is only slightly less than the DCR with manual tip tracking (although the difference is statistically significant). Additionally, interested readers are referred to Appendix C.2 where the influence of catheter segmentation on the accuracy of DCR is investigated.

Table 8 shows how the frame mean point distances of the 409 evaluation frames are distributed. The DCR with the proposed method has similar error distribution as the one with manual tip tracking: they both have about 1/3 of the distances less than 1 mm and 1/3 of the distances between 1 and 2 mm. The proposed method has slightly more distances larger than 5 mm than manual tip tracking. Both methods are more accurate than the DCR without tracking on intervals of small errors (< 2 mm).

Figure 14 shows overlays of selected roadmaps on example frames of 4 sequences with the three DCR approaches. The DCR without tracking presents mismatch of catheters, guidewires or residual of contrast agent in the images, whereas the other methods improve the alignment and show good match between the structures in the original X-ray image and the roadmaps. Compared to the DCR with manual tip tracking, the proposed method show similar visual alignment of the roadmaps to the original X-ray images. For a dynamic view of a roadmapping example, we refer readers to the supplemental material.
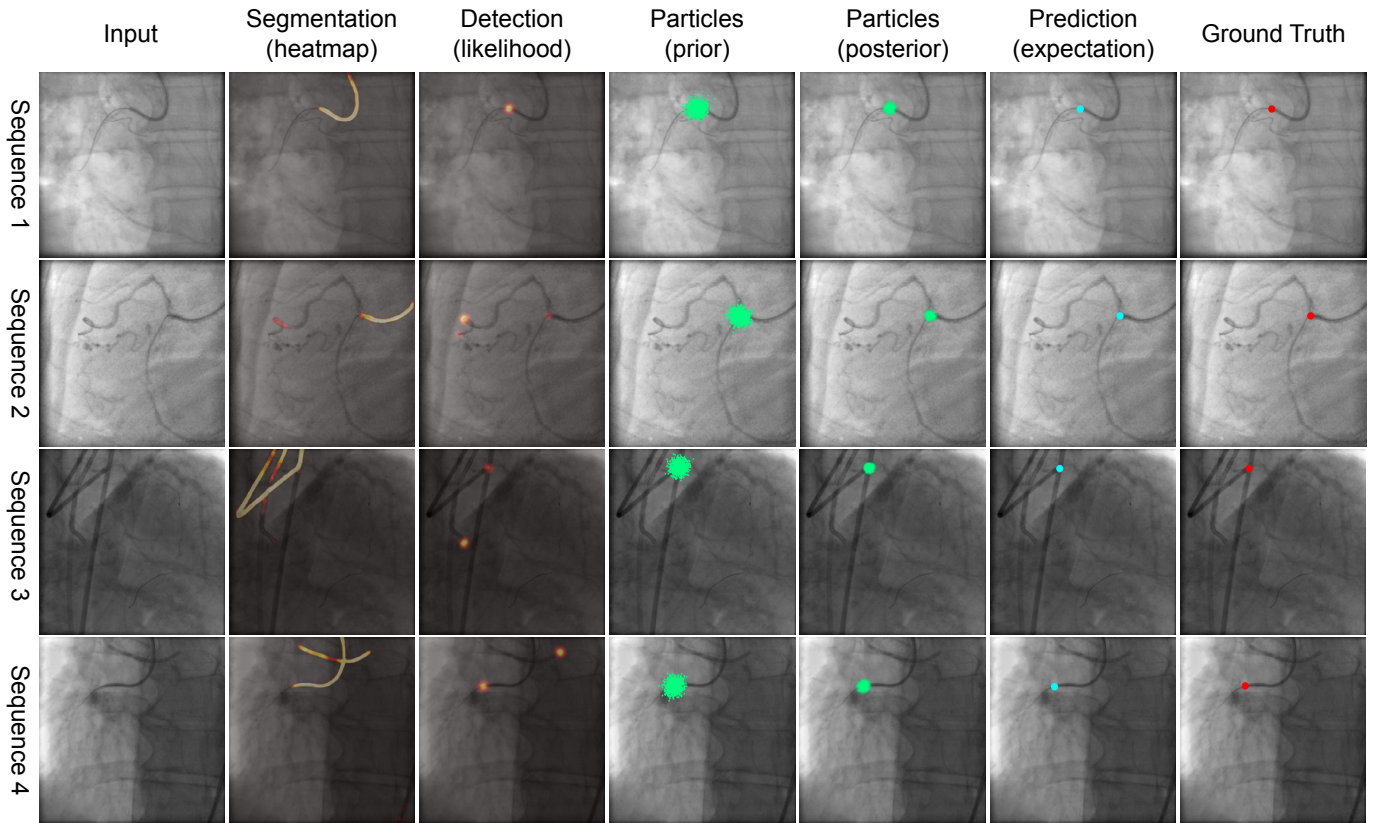
Figure 9: Workflow of the proposed tracking method on the same 4 frames in Figure 8. The high probability is shown with bright color in the detection map. Samples or particles are presented as green dots. The blue point indicates the predicted catheter tip location; the red point shows the ground truth location. (Best viewed in color)
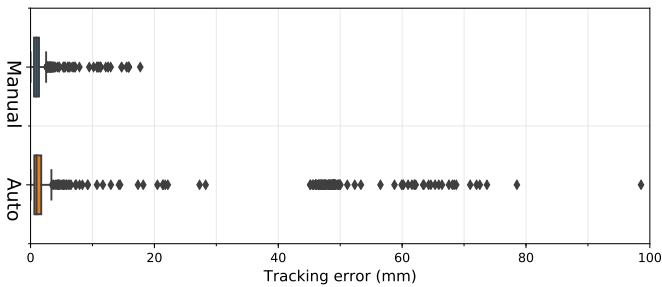

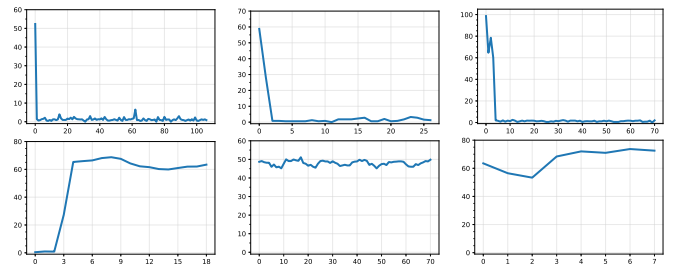Figure 10: Catheter tip tracking errors (mm) with manual and automatic initialization.

Figure 11: Longitudinal views of tracking errors (mm) for the 6 sequences with outliers using automatic initialization.

6.4. Processing Time

The processing time of all steps in the proposed DCR method was measured with the hardware and software setup in Section 5.5. The ECG matching method for roadmap selection was running in Python on the CPU of the Linux machine; the deep neural network and the optical flow component of the tracking method were running on the GPU.

In the experiments, the runtimes for roadmap se-

lection (step 1) and roadmap transformation (step 3) in Figure 1 were negligible (< 1 ms / frame). The runtime of the proposed catheter tip tracking method is shown in Table 9 and Figure 15. The average time to compute the likelihood with the deep learning setup (DL) is 31.5 ms / frame. The particle filtering (PF) step, which consists of the optical flow estimation, sample propagation, sample weight update and normalization, prediction and resampling, takes on

18

Table 6: Catheter tip tracking errors (mm) of detection and tracking with manual and automatic initialization

	Detection	Tracking	
		Manual init.	Automatic init.
Maximal error	108.20	17.23	98.58
Median error	0.96	0.96	0.96
Mean error	5.62 ± 15.91	1.29 ± 1.76	5.16 ± 13.91
No. of outliers (> 40 mm)	106	0	100
No. of sequences with outliers	10	0	6
Maximal error of inliers	31.06	17.23	28.28
Median error of inliers	0.96	0.96	0.96
Mean error of inliers	1.17 ± 1.78	1.29 ± 1.76	1.34 ± 2.15

Table 7: The statistics of DCR accuracy (mm) with three different tracking scenarios. With the two-sided Wilcoxon signed-rank test: † denotes that the difference between the DCR without tracking and that with the proposed tracking method is statistically highly significant ($p < 0.001$); * indicates a statistically significantly difference between the DCR using manual tip tracking and that with the proposed tracking approach ($p < 0.05$).

	Without Tracking†	Proposed Tracking Method	Manual Tip Tracking*
All point pairs			
Maximal distance	27.19	20.24	13.12
Median distance	1.97	1.43	1.35
Mean distance	2.94 ± 2.83	2.07 ± 2.08	1.85 ± 1.72
Frame mean distance			
Median distance	2.11	1.42	1.38
Average distance	2.76 ± 2.08	1.91 ± 1.52	1.75 ± 1.30

average 23 ms / frame. Therefore, the average tracking time in total is 54.5 ms / frame. The total average time of the proposed DCR including roadmap selection, catheter tip tracking and roadmap transformation is still less than the acquisition time of our data (66.7 ms / frame, 15 fps), indicating that the proposed DCR method would run in real-time with our setup.

7. Discussion

We have presented a new approach to perform on-line dynamic coronary roadmapping on X-ray fluoroscopic sequences for PCI procedures. The approach compensates the cardiac-induced vessel motion via selecting offline-stored roadmaps with appropriate cardiac phase using ECG matching, and corrects the respiratory motion of vessels by online tracking of guiding catheter tip in X-ray fluoroscopy using a proposed deep learning based Bayesian filtering. The proposed tracking method represents and tracks the posterior of catheter tip via a particle filter, for which a likelihood probability map is computed for updating the particle weights using a convolutional neural network. In the experiments, the proposed DCR approach has been trained and evaluated on clinical X-ray sequences for both tracking and roadmapping tasks.

One prerequisite of accurate tracking with the proposed approach is to obtain a reasonably good likelihood estimation, which requires to train the deep neural network to detect catheter tip well. In this work, we have investigated the influence of three network hyperparameters on the performance of the detection network (Section 6.1): the basic channel

Table 8: Distribution of frame mean point distances of the 409 evaluation frames.

Tracking Methods of DCR	Error Intervals (mm)					
	< 1	1-2	2-3	3-4	4-5	≥ 5
Without tracking	81	115	69	47	31	66
Proposed tracking method	131	145	61	32	17	23
Manual tip tracking	139	144	61	35	20	10

Table 9: Statistics of the runtime of catheter tip tracking (ms / frame) on the test (tracking) dataset.

	Deep Learning	Particle Filtering	Total Tracking Time
Mean	31.5 ± 10.3	23.0 ± 8.7	54.5 ± 12.3
Median	35.1	22.8	57.7

number and network depth level are model capacity parameters, the dropout adds regularization to the model. The experiment showed that the detection accuracy improves when the basic channel number and the network depth level increase (Table 3). This observation matches the expectation that a more complex model has higher capacity to model the variation in the data, hence results in better accuracy. When the complexity reaches a certain level, e.g. 64 basic channels and 5 level of depth, the network performance does not increase much compared to those with simpler settings, implying that the model starts overfitting on our dataset.
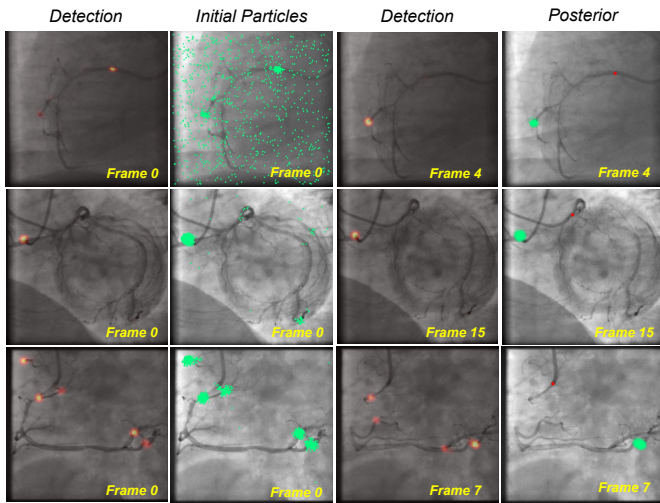
In addition to the deep neural network, the other important component of the proposed tracking approach is the sampling in the particle filter that yields the samples for representing the prior and the posterior of catheter tip position. First, a sufficient number of samples in the whole sample space are required to well characterize the probability distributions (see Appendix B.2). Second, the sample dynamics plays an important role in tracking, in particular, as indicated by Eq.(8), the process noise and the sample motion. The process noise influences the tracking accuracy, according to Table 10 in Appendix B.2. Additionally, sample motion is another key aspect of sample dynamics. Motion estimation has previouly been incorporated in a motion-based particle filter, such as adaptive block matching (Bouaynaya and Schonfeld (2009)). In our work, optical flow was chosen for motion estimation, as its non-parametric

nature allows to characterize the complexity of motion in X-ray fluoroscopy well. In addition, the advantage of such approach from a theoretical point of view is that it takes into account of the current observation, leading to a more optimal importance density (Arulampalam et al. (2002)) compared to random motion.

The tracking results in Section 6.2.1 show that the proposed tracking approach is able to track the catheter tip in X-ray fluoroscopy accurately with an average tracking error of about 1.3 mm. It also shows advantages over methods based only on optical flow or the detection network. The OF (pre) method relies heavily on tracking in the previous frame, hence the error could accumulate. The OF (first) method may suffer from large motion from the first frame to the current frame. The detection method uses information only from the current frame, no temporal relation between frames is utilized; therefore, it results in spikes in the longitudinal view, as shown in Figure 7. The proposed tracking method has a CNN to provide an accurate observation on the current frame which improves the accuracy of optical flow tracking within the framework of Bayesian filtering. In the meantime, the optical flow based particle filter maintains and propagates the prior knowledge from the initial tip position to provide a constraint on searching for the potentially correct positions, which is useful especially when the CNN detector fails to find the correct target area. The association of knowledge from two sources together improves the tracking accuracy

20
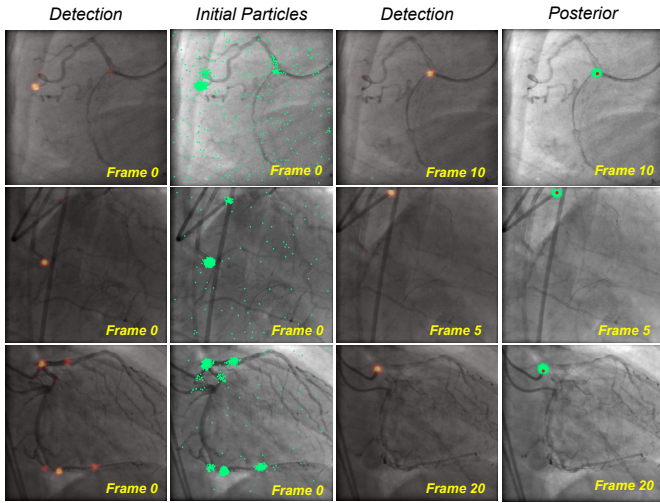
(a) Sequence 1-3 on the top row in Figure 11

Figure 13: Accuracy (mm) of DCR with three different tracking scenarios.

(b) Sequence 4-6 on the bottom row in Figure 11

Figure 12: Examples frames from the 6 sequences in Figure 11. The high probability in the detection heatmap is highlighted as bright color. Particles are presented as green dots. The red dots in the last column indicate the ground truth tip location. (Best viewed in color)
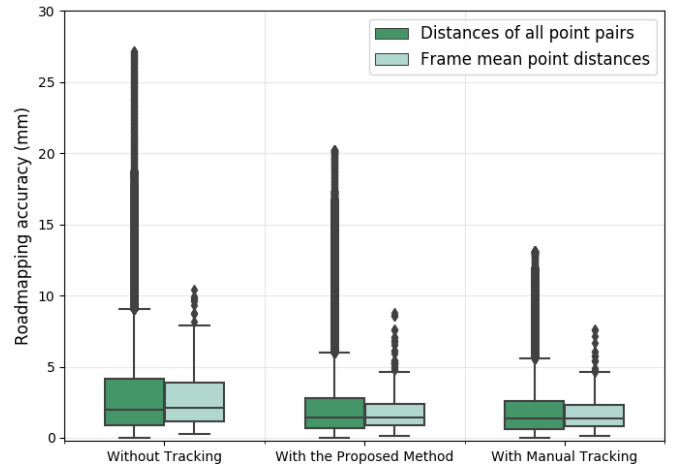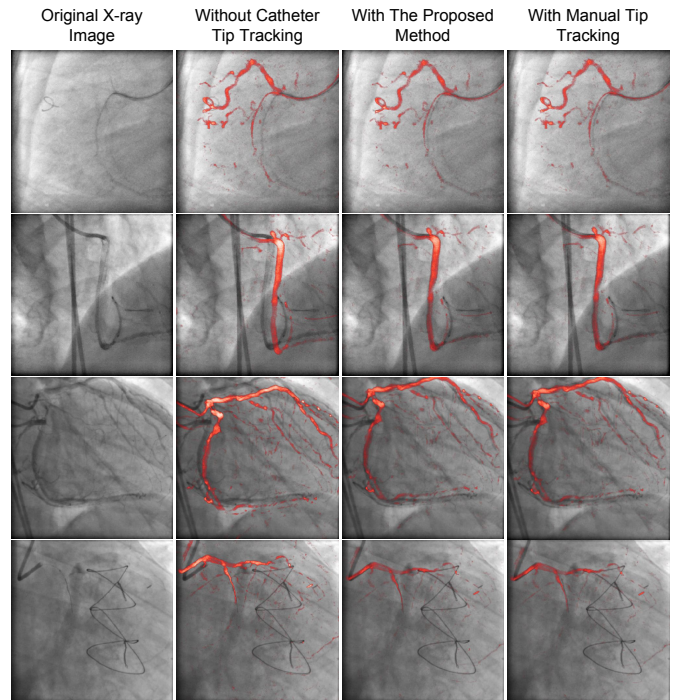

Figure 14: Examples of superimposition of selected roadmaps (red) on X-ray fluoroscopic frames. (Best viewed in color)
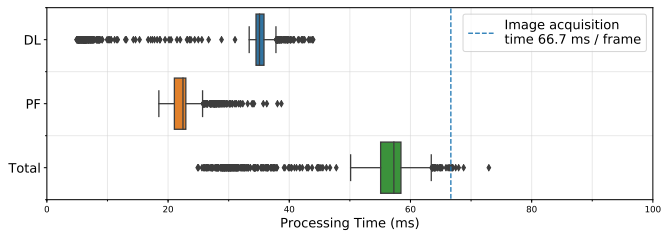
Figure 15: Runtime of catheter tip tracking (ms / frame) on all test frames.

compared to each single source.

The initial state is a also key component of tracking approaches. In the context of Bayesian filtering, the initial state provides the prior knowledge of the tracking target. Most tracking algorithms assume a known initial state from which the target is tracked, e.g. our proposed method with manual initialization in Section 6.2.2. In this case, the prior knowledge is provided by human. In Section 6.2.2, we also investigated a scenario where the initial state is given by the detection network, so that the complete tracking process is fully automated. The results indicate that, the proposed tracking method with automatic initialization works reasonably well on most sequences even when the initialization is sometimes incorrect (Figure 12a). This is because (1) the true position is covered by a few samples, and (2) the correct detection in later frames corrects the initial mistake in the first frame. The automatic initialization fails when (1) a wrong position is covered by a few samples and (2) the wrong detection in subsequent frames confirms the mistake in the initial frame (Figure 12b). This happens when there is contrast agent remaining in the image or there are multiple catheters, which are the major sources causing ambiguity in detection. In practice, the automatic initialization would work well when contrast agent is washed out and only one catheter is present in the field of view, otherwise manual initialization would be needed which requires only one click to initiate tracking.

Dynamic coronary roadmapping is the direct application of the catheter tip tracking results. In our experiments, the DCR was performed with manual tip initialization to show the potential of the proposed tracking method, and was compared with the DCR without tracking and with manual tracking. The results indicate that using catheter tip tracking can

improve DCR accuracy, as the respiratory-induced vessel motion is corrected by the displacement of catheter tip in addtion to cardiac motion correction. The results also show that the proposed DCR reaches a good accuracy (mean error is about 2 mm) and performs only slightly worse than its best case, the DCR with manual tip tracking which is not applicable for intraoperative use. Additionally, according to a previous study by Dodge et al. (1992), the average lumen diameters of human coronary arteries are between 1.9 mm (distal left anterior descending artery) and 4.5 mm (left main artery). This means that the accuracy achieved with the proposed approach is comparable with the size of coronary arteries.

Apart from catheter tip tracking, several other possible factors in different steps of the experiments may influence the final DCR accuracy. First, in the offline phase, the signal of contrast agent may become too strong and completely cover the catheter tip, complicating the tip visibility in some cases. In this situation, the uncertainty in the manual tip annotation may result in errors in roadmap transformation. Second, in the roadmap selection step, the offline-stored roadmaps are only discrete samples of complete cardiac cycles which might not fully characterize every possible change in the cardiac motion. This problem could possibly be addressed in the future by interpolating frames between the existing frames in the data. Additionally, variation exists between different cardiac cycles (McClelland et al. (2013)), therefore, choosing a roadmap from another cycle may cause inaccuracy for cardiac motion compensation. Finally, the way of DCR evaluation in Section 5.4 might also introduce inaccuracies in the error measurement. Since guidewires often attach to the inner curves of a vessel to take the shortest path, the small difference between the annotation of guidewire and vessel centerlines was ignored in the evaluation.

In addition to accuracy, processing speed is also critical for intraoperative applications. The results in Section 6.4 indicate that the total processing time of the proposed DCR approach is less than the image acquisition time, meaning that it runs in real-time on our setup. To build a real-time system for PCI in practice, the overall latency of the complete system needs to be considered. It is also worth noticing that the DL and PF steps of the proposed tracking method

are independent from each other. In practice, in case more than one GPU are available, the proposed DCR approach can be further accelerated by paralleling the DL and PF steps, making them running on different GPUs.

Compared to the previous works on DCR, the proposed approach in this paper shows advancement in several aspects. First, our systems works on non-cardiac-gated sequences which does not require additional setups for cardiac motion gating that were needed for some methods (Zhu et al. (2010); Manhart et al. (2011)). Second, our approach compensates both respiratory- and cardiac-induced vessel motion, which is more accurate than systems that correct only cardiac motion (Elion (1989)). In addition, the proposed DCR approach follows a data-driven paradigm that learns target feature from sequences acquired from different patients and various view angles, making it more robust than the method that relies on traditional vesselness filtering (Kim et al. (2018)) or methods that require specific tissue being present (Zhu et al. (2010); Manhart et al. (2011)). These are the major advantages of the proposed DCR over the existing direct roadmapping systems. Compared to model-based motion compensation, our approach does not require the extraction of motion surrogate signals and train a motion model for each new patient, but can be directly run with a trained model.

The proposed deep learning based Bayesian filtering method has several advantages over the existing instrument tracking methods. First, the deep learning component enables a more general framework to detect instruments in medical images than methods tailored for specific tools (Ma et al. (2012, 2013)). Compared to the existing detection methods based on deep learning (Baur et al. (2016); Laina et al. (2017); Du et al. (2018)), our approach takes into account of the information between frames; the Bayesian filtering framework allows interaction between temporal information and the detection of a convolutional neural network, making the tracking more robust. Bayesian frameworks have been used in many previous temporal instrument tracking methods. Particularly, the likelihood term in some works was designed based on registration or segmentation outcomes (Ambrosini et al. (2017b); Speidel et al. (2006)) or traditional machine learning approaches

with handcrafted features (Wang et al. (2009); Wu et al. (2012); Pauly et al. (2010)). In our method, we approximated the likelihood with a deep neural network learned from the clinical data which exempts the need of feature engineering but yet possesses more discriminative power; the network directly outputs the probability map, making it more straightforward to use. Finally, compared to the existing instrument tracking approaches based on Bayesian filtering (Ambrosini et al. (2017b); Speidel et al. (2006, 2014)), the state transition in our method was based on the motion estimated from two adjacent frames, which is more reliable than totally random motion or artificially-designed state transition models.

From a practical point of view, the proposed DCR approach could potentially fit into the clinical workflow of PCI. The offline phase of the method can be done efficiently by a person who assists the procedures: selecting and creating roadmaps from an angiography acquisition, annotating the catheter tip (one point) in the images. This phase is typically done before a fluoroscopy acquisition during which the guidewire advancement and stent placement are performed. In the online phase, when a fluoroscopic image is acquired, the proposed system selects the most suitable roadmap, tracks the catheter tip and transforms the roadmap to prospectively show a vessel overlay on the fluorosocpic image. The online updated coronary roadmap can provide real-time visual guidance to cardiologists to manipulate interventional tools during the procedure without the need of administering extra contrast agent.

In the future, it may be worth investigating the following directions related to this work. As the data used in this study was acquired from one hospital using a machine from a single vendor, it would be interesting to evaluate the proposed approach on multi-center data acquired with machines from different vendors. Next, since the ECG signals of our data appear to be regular, it may be necessary in a future study to acquire data with irregular ECG that could be obtained in practice, and validate the proposed approach on those data. Besides, it would be also interesting to validate our approach during PCI procedures in an environment simulating the real clinical settings. Additionally, from a methodological point of view, although the proposed tracking

method is invariant under different view angles, the whole DCR approach works only when the offline and online phase have the same view angle, i.e. it is a 2D roadmapping system. Therefore, one future direction would be to develop a 3D DCR system that would work with various view angles in the online phase.

8. Conclusion

We have developed and validated a novel approach to perform dynamic coronary roadmapping for PCI image guidance. The approach compensates cardiac motion through ECG alignment and respiratory motion by guiding catheter tip tracking during fluoroscopy with a deep learning based Bayesian filtering method. The proposed tracking and roadmapping approaches were trained and evaluated on clinical X-ray image datasets and were proved to perform accurately on both catheter tip tracking and dynamic coronary roadmapping tasks. Our approach also runs in real-time on a setup with a modern GPU and thus has the potential to be integrated into routine PCI procedures, assisting the operator with real-time visual image guidance.

References

Ambrosini, P., Ruijters, D., Niessen, W.J., Moelker, A., van Walsum, T., 2017a. Fully automatic and real-time catheter segmentation in X-ray fluoroscopy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 577–585.

Ambrosini, P., Smal, I., Ruijters, D., Niessen, W.J., Moelker, A., Van Walsum, T., 2017b. A hidden markov model for 3D catheter tip tracking with 2D X-ray catheterization sequence and 3D rotational angiography. IEEE Trans. Med. Imaging 36, 757–768.

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on signal processing 50, 174–188.

Baka, N., Lelieveldt, B., Schultz, C., Niessen, W., van Walsum, T., 2015. Respiratory motion estimation in X-ray angiography for improved guidance during coronary interventions. Physics in Medicine & Biology 60, 3617.

Baur, C., Albarqouni, S., Demirci, S., Navab, N., Fallavollita, P., 2016. Cathnets: detection and single-view depth prediction of catheter electrodes, in: International Conference on Medical Imaging and Virtual Reality, Springer. pp. 38–49.

Bouaynaya, N., Schonfeld, D., 2009. On the optimality of motion-based particle filtering. IEEE transactions on circuits and systems for video technology 19, 1068–1072.

Brost, A., Liao, R., Strobel, N., Hornegger, J., 2010. Respiratory motion compensation by model-based catheter tracking during EP procedures. Medical Image Analysis 14, 695–706.

Chang, P.L., Rolls, A., De Praetere, H., Vander Poorten, E., Riga, C.V., Bicknell, C.D., Stoyanov, D., 2016. Robust catheter and guidewire tracking using B-spline tube model and pixel-wise posteriors. IEEE Robotics and Automation Letters 1, 303–308.

Dannenberg, L., Polzin, A., Bullens, R., Kelm, M., Zeus, T., 2016. On the road: First-in-man bifurcation percutaneous coronary intervention with the use of a dynamic coronary road map and stentboost live imaging system. International journal of cardiology 215, 7–8.

Dodge, J.T., Brown, B.G., Bolson, E.L., Dodge, H.T., 1992. Lumen diameter of normal human coronary arteries. influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation. Circulation 86, 232–246.

Du, X., Kurmann, T., Chang, P.L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J.D., Stoyanov, D., 2018. Articulated multi-instrument 2D pose estimation using fully convolutional networks. IEEE transactions on medical imaging .

Elion, J.L., 1989. Dynamic coronary roadmapping. US Patent 4,878,115.

Faranesh, A.Z., Kellman, P., Ratnayaka, K., Lederman, R.J., 2013. Integration of cardiac and respiratory motion into MRI roadmaps fused with X-ray. Medical physics 40.

Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion, in: Scandinavian conference on Image analysis, Springer. pp. 363–370.

Fischer, P., Faranesh, A., Pohl, T., Maier, A., Rogers, T., Ratnayaka, K., Lederman, R., Hornegger, J., 2018. An MR-based model for cardio-respiratory motion compensation of overlays in X-ray fluoroscopy. IEEE transactions on medical imaging 37, 47–60.

García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking, in: International Workshop on Computer-Assisted and Robotic Endoscopy, Springer. pp. 84–95.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Heibel, H., Glocker, B., Groher, M., Pfister, M., Navab, N., 2013. Interventional tool tracking using discrete optimization. IEEE transactions on medical imaging 32, 544–555.

Honnorat, N., Vaillant, R., Paragios, N., 2011. Graph-based geometric-iconic guide-wire tracking, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 9–16.

Kim, D., Park, S., Jeong, M.H., Ryu, J., 2018. Registration of angiographic image on real-time fluoroscopic image for image-guided percutaneous coronary intervention. International journal of computer assisted radiology and surgery 13, 203–213.

King, A.P., Boubertakh, R., Rhode, K.S., Ma, Y., Chinchapatnam, P., Gao, G., Tangcharoen, T., Ginks, M., Cooklin, M., Gill, J.S., et al., 2009. A subject-specific technique for respiratory motion correction in image-guided cardiac catheterisation procedures. Medical Image Analysis 13, 419–431.

Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J.P., Eslami, A., Tombari, F., Navab, N., 2017. Concurrent segmentation and localization for tracking of surgical instruments, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 664–672.

Ma, H., Dibildox, G., Banerjee, J., Niessen, W., Schultz, C., Regar, E., van Walsum, T., 2015. Layer separation for vessel enhancement in interventional X-ray angiograms using morphological filtering and robust PCA, in: Workshop on Augmented Environments for Computer-Assisted Interventions, Springer. pp. 104–113.

Ma, Y., Gogin, N., Cathier, P., Housden, R.J., Gijsbers, G., Cooklin, M., O'Neill, M., Gill, J., Rinaldi, C.A., Razavi, R., et al., 2013. Real-time X-ray fluoroscopy-based catheter detection and tracking for cardiac electrophysiology interventions. Medical physics 40.

Ma, Y., King, A.P., Gogin, N., Gijsbers, G., Rinaldi, C.A., Gill, J., Razavi, R., Rhode, K.S., 2012. Clinical evaluation of respiratory motion compensation for anatomical roadmap guided cardiac electrophysiology procedures. IEEE Transactions on Biomedical Engineering 59, 122–131.

Manhart, M., Zhu, Y., Vitanovski, D., 2011. Self-assessing image-based respiratory motion compensation for fluoroscopic coronary roadmapping, in: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE. pp. 1065–1069.

McClelland, J.R., Hawkes, D.J., Schaeffter, T., King, A.P., 2013. Respiratory motion models: a review. Medical image analysis 17, 19–42.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE. pp. 565–571.

Pauly, O., Heibel, H., Navab, N., 2010. A machine learning approach for deformable guide-wire tracking in fluoroscopic sequences, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 343–350.

Peressutti, D., Penney, G.P., Housden, R.J., Kolbitsch, C., Gomez, A., Rijkhorst, E.J., Barratt, D.C., Rhode, K.S., King, A.P., 2013. A novel Bayesian respiratory motion model to estimate and resolve uncertainty in image-guided cardiac interventions. Medical image analysis 17, 488–502.

Petković, T., Lončarić, S., 2010. Guidewire tracking with projected thickness estimation, in: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, IEEE. pp. 1253–1256.

Piayda, K., Kleinebrecht, L., Afzal, S., Bullens, R., ter Horst, I., Polzin, A., Veulemans, V., Dannenberg, L., Wimmer, A.C., Jung, C., et al., 2018. Dynamic coronary roadmapping during percutaneous coronary intervention: a feasibility study. European journal of medical research 23, 36.

Rieke, N., Tan, D.J., di San Filippo, C.A., Tombari, F., Alsheakhali, M., Belagiannis, V., Eslami, A., Navab, N., 2016. Real-time localization of articulated surgical instruments in retinal microsurgery. Medical image analysis 34, 82–100.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Schneider, M., Sundar, H., Liao, R., Hornegger, J., Xu, C., 2010. Model-based respiratory motion compensation for image-guided cardiac interventions, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2948–2954.

Shechter, G., Ozturk, C., Resar, J.R., McVeigh, E.R., 2004. Respiratory motion of the heart from free breathing coronary angiograms. IEEE transactions on medical imaging 23, 1046.

Shechter, G., Shechter, B., Resar, J.R., Beyar, R., 2005. Prospective motion correction of X-ray images for coronary interventions. IEEE Transactions on Medical Imaging 24, 441–450.

Speidel, S., Delles, M., Gutt, C., Dillmann, R., 2006. Tracking of instruments in minimally invasive surgery for surgical skill analysis, in: International Workshop on Medical Imaging and Virtual Reality, Springer. pp. 148–155.

Speidel, S., Kuhn, E., Bodenstedt, S., Röhl, S., Kenngott, H., Müller-Stich, B., Dillmann, R., 2014. Visual tracking of da vinci instruments for laparoscopic surgery, in: Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling, International Society for Optics and Photonics. p. 903608.

Takimura, H., Muramatsu, T., Tsukahara, R., Nakano, M., Nishio, S., Takimura, Y., Yabe, T., Kawano, M., Hada, T., 2018. Usefulness of novel roadmap guided percutaneous coronary intervention for bifurcation lesions (the world's first cases). Journal of the American College of Cardiology 71, A1365.

Tehrani, S., Laing, C., Yellon, D.M., Hausenloy, D.J., 2013. Contrast-induced acute kidney injury following pci. Euro-

25

pean journal of clinical investigation 43, 483–490.

Timinger, H., Krueger, S., Dietmayer, K., Borgert, J., 2005. Motion compensated coronary interventional navigation by means of diaphragm tracking and elastic motion models. Physics in Medicine & Biology 50, 491.

Vandini, A., Glocker, B., Hamady, M., Yang, G.Z., 2017. Robust guidewire tracking under large deformations combining segment-like features (seglets). Medical image analysis 38, 150–164.

van Walsum, T., Schaap, M., Metz, C.T., van der Giessen, A.G., Niessen, W.J., 2008. Averaging centerlines: mean shift on paths, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 900–907.

Wang, P., Chen, T., Zhu, Y., Zhang, W., Zhou, S.K., Comaniciu, D., 2009. Robust guidewire tracking in fluoroscopy, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 691–698.

Wu, W., Chen, T., Strobel, N., Comaniciu, D., 2012. Fast tracking of catheters in 2D fluoroscopic images using an integrated CPU-GPU framework, in: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, IEEE. pp. 1184–1187.

Wu, X., Housden, J., Ma, Y., Razavi, B., Rhode, K., Rueckert, D., 2015. Fast catheter segmentation from echocardiographic sequences based on segmentation from corresponding X-ray fluoroscopy for cardiac catheterization interventions. IEEE transactions on medical imaging 34, 861–876.

Yabe, T., Muramatsu, T., Tsukahara, R., Nakano, M., Takimura, Y., Takimura, H., Kawano, M., Hada, T., 2018. The impact of percutaneous coronary intervention using the novel dynamic coronary roadmap system. Journal of the American College of Cardiology 71, A1103.

Yatziv, L., Chartouni, M., Datta, S., Sapiro, G., 2012. Toward multiple catheters detection in fluoroscopic image guided interventions. IEEE Transactions on Information Technology in Biomedicine 16, 770–781.

Zhu, Y., Tsin, Y., Sundar, H., Sauer, F., 2010. Image-based respiratory motion compensation for fluoroscopic coronary roadmapping, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 287–294.

Appendix

A. Details of the Training Setup

A.1. Data Augmentation

To increase the number of training samples and their diversity, data augmentation was used. The augmentation includes geometric transformation such as flipping (left-right, up-down), rotation of multiple of 90 degrees, random affine transformation (translation -10 to 10 px, scaling 0.9 to 1.1, rotation -5 to 5 degrees, shear -5 to 5 px), random elastic deformation (deformation range -4 to 4 px, grid size of control points 64 px). A training sample has 0.5 probability of being processed with one of the transformations. The probability for applying each transformation is: flipping left-right (1/24), flipping up-down (1/24), rotation of multiple of 90 degrees (1/12), affine transformation (1/6), elastic deformation (1/6), no transformation (1/2). To make the trained model robust to noise, in addition to the geometric transformations, we also augmented data by adding Gaussian noise to the pixel value with a zero mean and a standard deviation between 0.01 and 0.03. The probability of adding the noise is 0.5.

A.2. Training Settings

The λ value in the training loss Eq. (5) was set to 10 to make the scale of the two terms similar. Adam optimizer was used to minimize the loss function with a learning rate 0.0001. The number of training samples in a batch is 4. The network was trained with 100 epochs to ensure convergence.

B. Parameter Tuning for Catheter Tip Tracking

This section describes the details of tuning the parameters of optical flow and particle filter for catheter tip tracking.

B.1. Tuning Optical Flow Parameters

The approach of Farnebäck (2003) was used as the optical flow implementation in Algorithm 1. A grid search to find the optimal parameter setting was done on the following parameters of the method: (1) the image scale to build the pyramids, (2) the number pyramid levels, (3) the averaging window size, (4) the number of iterations, (5) the size of the pixel neighborhood used to find polynomial expansion in each pixel, and finally (6) the standard deviation of the Gaussian that is used to smooth derivatives used as a basis for the polynomial expansion.

The above parameters were tuned independently of the deep neural network, as optical flow directly estimates the catheter tip motion between two frames. To tune the parameters, we tracked the catheter tip in X-ray fluoroscopy starting from the ground truth tip position in the first frame using the motion field between two adjacent frames estimated with optical flow. The average and median distance

between the tracked tip position and the ground truth were used as the evaluation criteria for the tuning.

The method of Farnebäck (2003) was implemented by using the OpenCV function `calcOpticalFlowFarneback`. With consideration of the suggested parameter values from the documentation, the parameter setting chosen for optical flow from the grid search is `pyr_scale = 0.5`, `levels = 3`, `winsize = 10`, `iterations = 30`, `poly_n = 5`, `poly_sigma = 1.1`. Details of the parameters can be found on the function documentation page[1].
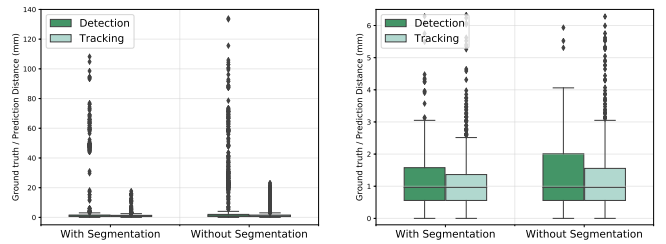
B.2. Tuning Particle Filter Parameters

The parameters to tune for the particle filter are the number of samples N_s and the variance of process noise σ_v^2. When tuning them, we fixed the parameters of the trained network and the optical flow method, and used their optimal parameter settings during this experiment. Following Algorithm 1, we tracked the catheter tip from the ground truth position (probability map) in the first frame, and used the mean and median distance between the tracked and the true position as the validation metric.

The tracking results on the validation (tracking) set are shown in Table 10. The table shows that 100 samples are suboptimal, while 1000 samples seem sufficient, as 10000 samples result in tracking accuracies similar to 1000 samples. It also shows that the optimal choices of the standard deviation of the process noise are 4 or 5 px for the downsampled images. One possible reason for such choices may be that they are similar to the size of guiding catheters. In general, good choices for N_s are 1000 and 10000, for σ_v are 4 and 5. By considering the mean, the standard deviation and the median of tracking errors, the parameter setting $\sigma_v = 5$, $N_s = 1000$ was chosen.

C. Detection, tracking and roadmapping without catheter segmentation

Training of the network in Figure 3 requires catheter labels for detection and segmentation. As the segmentation labels are often more expensive to acquire than the detection label in practice, we also

(a) Overall view of tracking errors

(b) A zoom-in view of (a)

Figure 16: Tracking errors on all test images with and without catheter segmentation.

investigated the performance of catheter tip detection, tracking and dynamic coronary roadmapping without segmenting the catheter. To this end, we used a similar encoder-decoder network architecture as Figure 3 which computes only the detection output directly after the last *up* block of the decoder with a 1×1 convolution followed by a spatial softmax layer, instead of having two outputs. We then followed the same way as the approach using the network with segmentation to search for (hyper-)parameters for the approach without segmentation. The following parameter setting was chosen for the experiments in this section: for deep learning, the basic channel number is 64, the depth is 5, the dropout rate is zero; for particle filtering, $\sigma_v = 3$, $N_s = 10000$. With this setup, we compared the performance of the approach without catheter segmentation to the proposed approach with segmentation on catheter tip detection and tracking (Appendix C.1) and dynamic coronary roadmapping (Appendix C.2) on the test set from Table 2.

C.1. Catheter tip detection and tracking

The same metrics in Table 5 are used to report the accuracy of catheter tip detection and tracking without catheter segmentation. Table 11 and Figure 16 both manifest that the segmentation sub-task improves the accuracy of catheter tip detection and tracking. Although the improvement on the tracking task is marginal and not statistically significant ($p = 0.06$), the segmentation helps to reduce the magnitude and amount of outliers (large errors).

---

[1] `https://docs.opencv.org/2.4/modules/video/doc/motion_analysis_and_object_tracking.html?`

Table 10: Catheter tip tracking errors (mm) on the validation (tracking) dataset of different parameter settings for particle filter. The tracked tip point was rounded to the pixel center. The error of all images (mean ± std / median) are presented. Red cells show the good choices of parameters; *bold* number indicates the chosen setting.

σ_v	N_s		
(px)	100	1000	10000
3	1.52 ± 2.19 / 0.79	1.49 ± 2.18 / 0.79	1.48 ± 2.18 / 0.79
4	1.50 ± 2.17 / 0.79	1.46 ± 2.17 / 0.79	1.47 ± 2.18 / 0.79
5	1.52 ± 2.21 / 0.79	**1.47 ± 2.17 / 0.74**	1.47 ± 2.19 / 0.74
6	1.53 ± 2.39 / 0.79	1.49 ± 2.33 / 0.79	1.48 ± 2.29 / 0.74
7	1.56 ± 2.42 / 0.79	1.50 ± 2.29 / 0.74	1.50 ± 2.39 / 0.74
8	1.58 ± 2.41 / 0.79	1.51 ± 2.40 / 0.74	1.51 ± 2.42 / 0.74
9	1.56 ± 2.22 / 0.79	1.53 ± 2.43 / 0.79	1.52 ± 2.45 / 0.61
10	2.25 ± 6.18 / 0.79	1.54 ± 2.46 / 0.79	1.53 ± 2.47 / 0.61

Table 11: Catheter tip tracking errors (mm) with and without catheter segmentation on the test (tracking) dataset. † indicates that the difference between the detection with and without segmentation is statistically highly significant with the two-sided Wilcoxon signed-rank test ($p < 0.001$). No statistically significantly difference is observed between the tracking with and without segmentation using the two-sided Wilcoxon signed-rank test ($p = 0.06$).

Evaluation Metrics	With Segmentation		Without Segmentation	
	Detection†	Tracking	Detection	Tracking
Maximal error of all images	108.20	**17.72**	133.94	23.2
Median error of all images	**0.96**	**0.96**	**0.96**	**0.96**
Mean error of all images	5.62 ± 15.91	**1.29 ± 1.76**	9.32 ± 19.73	1.75 ± 3.17
Average of sequence median error	6.26 ± 17.11	**1.03 ± 0.49**	9.34 ± 18.64	1.42 ± 2.14
Average of sequence mean error	6.83 ± 13.88	**1.29 ± 0.94**	10.41 ± 15.94	1.69 ± 1.97

C.2. Dynamic coronary roadmapping

In this experiment, the same setup in Section 6.3 was used to assess the accuracy of DCR using catheter tip tracking without segmenting the catheter. Table 12 indicate that tracking the catheter tip with catheter segmentation improves the DCR accuracy compared to that without catheter segmentation. Although the improvement is not statistically significant ($p = 0.43$), the approach with segmentation is more robust by making less large roadmapping errors (Figure 17).
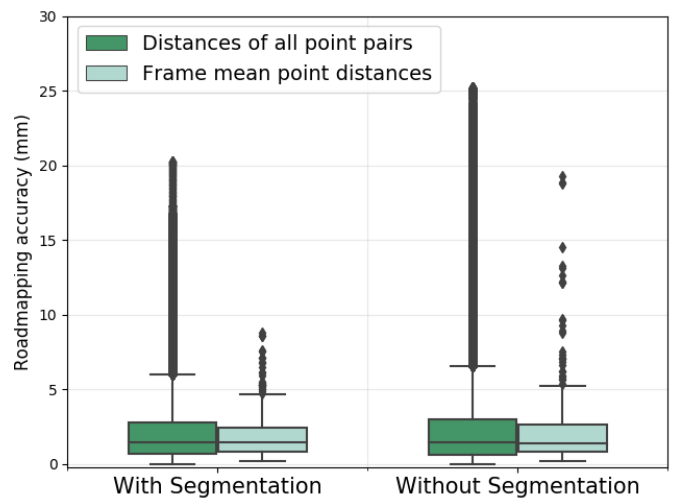

Figure 17: Accuracy (mm) of DCR via catheter tip tracking with and without catheter segmentation.

Table 12: The statistics of DCR accuracy (mm) via catheter tip tracking with and without catheter segmentation. With the two-sided Wilcoxon signed-rank test, no statistically significantly difference is observed between the DCR with and without segmentation ($p = 0.43$).

	With Segmentation	Without Segmentation
All point pairs		
Maximal distance	20.24	25.20
Median distance	1.43	1.43
Mean distance	2.07 ± 2.08	2.44 ± 3.15
Frame mean distance		
Median distance	1.42	1.40
Average distance	1.91 ± 1.52	2.23 ± 2.59