Medical Image Analysis 18 (2014) 1320–1328

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data

Tobias Heimann [a,*], Peter Mountney [b], Matthias John [c], Razvan Ionasec [b,1]

[a] Siemens AG, Corporate Technology, Imaging and Computer Vision, Erlangen, Germany
[b] Siemens Corporation, Corporate Technology, Imaging and Computer Vision, Princeton, NJ, USA
[c] Siemens AG, Healthcare Sector, Forchheim, Germany

ARTICLE INFO

Article history:
Received 7 January 2014
Received in revised form 17 March 2014
Accepted 11 April 2014
Available online 5 May 2014

Keywords:
Transfer learning
Domain adaptation
Object localization
Fluoroscopy
Ultrasound

ABSTRACT

The fusion of image data from trans-esophageal echography (TEE) and X-ray fluoroscopy is attracting increasing interest in minimally-invasive treatment of structural heart disease. In order to calculate the needed transformation between both imaging systems, we employ a discriminative learning (DL) based approach to localize the TEE transducer in X-ray images. The successful application of DL methods is strongly dependent on the available training data, which entails three challenges: (1) the transducer can move with six degrees of freedom meaning it requires a large number of images to represent its appearance, (2) manual labeling is time consuming, and (3) manual labeling has inherent errors.

This paper proposes to generate the required training data automatically from a single volumetric image of the transducer. In order to adapt this system to real X-ray data, we use unlabeled fluoroscopy images to estimate differences in feature space density and correct covariate shift by instance weighting. Two approaches for instance weighting, probabilistic classification and Kullback–Leibler importance estimation (KLIEP), are evaluated for different stages of the proposed DL pipeline. An analysis on more than 1900 images reveals that our approach reduces detection failures from 7.3% in cross validation on the test set to zero and improves the localization error from 1.5 to 0.8 mm. Due to the automatic generation of training data, the proposed system is highly flexible and can be adapted to any medical device with minimal efforts.

1. Introduction

Catheter-based procedures such as trans-aortic valve implantation (TAVI) or paravalvular leak closure are gaining increasing importance for the treatment of structural heart disease. The inherent challenge for the cardiac interventionalist is to infer the exact position of the catheter relative to the tissue from the available imaging information. X-ray fluoroscopy is the dominant imaging modality for these interventions, increasingly supported by 3D trans-esophageal echography (TEE) (Gao et al., 2012). Both modalities show complementary information, but in clinical practice they are controlled and displayed completely independently from each other.

Recently, image fusion has been proposed to combine both modalities and to provide the cardiac interventionalist with a better overview of the *in situ* conditions. The co-registration can be accomplished by means of electromagnetic (EM) tracking (Jain et al., 2009), but this approach requires EM tracking hardware to be attached to the transducer and is sensitive to EM field distortions. In Ma et al. (2010), authors present a feasibility study with a robotic arm for tracking a trans-thoracic echo probe. Apart from the difficulties of extending this system to TEE probes, the robotic hardware requirements severely limit the practical applicability of this approach. Alternatively, the pose of the transducer can be estimated from its appearance in the X-ray images, either directly (Gao et al., 2012; Mountney et al., 2012) or supported by fiducial markers attached to the probe head (Lang et al., 2012). Since the former approach does not require additional hardware, it is advantageous for integration into the clinical workflow, albeit more challenging to implement.

While 2D–3D registration (Gao et al., 2012) yields accurate results, it has a limited capture range of <10 mm, requiring a manual initialization every time a new fluoroscopy sequence is

* Corresponding author. Tel.: +49 9131 724607.
E-mail addresses: tobias.heimann@siemens.com (T. Heimann), peter.mountney@siemens.com (P. Mountney), matthias.mj.john@siemens.com (M. John), razvan.ionasec@siemens.com (R. Ionasec).
[1] Present address: Siemens AG, Healthcare Sector, Forchheim, Germany.

http://dx.doi.org/10.1016/j.media.2014.04.007
1361-8415/

acquired. Discriminative learning (DL) (Mountney et al., 2012) can locate the TEE probe everywhere in the image, but its performance is strongly dependent on quantity and quality of the available training data. In the medical domain, data is generally difficult to acquire, and the required manual labeling is an extremely tedious and time-consuming task. Moreover, trained operators cannot reproducibly annotate images with perfect accuracy, and every variation in ground truth will decrease the performance of the resulting DL system.

In this paper, we propose a novel approach for training a DL system based on *in silico* training data that can be generated automatically in large quantities with perfectly accurate labels. Since synthetic image generation cannot faithfully model all aspects of *in vivo* fluoroscopy data, the DL system must be adapted. For this purpose we employ unsupervised domain adaptation, a technique which has been widely used in speech processing and has recently gained attention in the computer vision community (Margolis, 2011; Beijbom, 2012). In particular, we show how unlabeled data from the target domain (i.e. *in vivo* images) can be used to improve the performance of object localization beyond what is achievable with semi-supervised learning (Zhu, 2008). We apply our approach to the estimation of in-plane parameters of a TEE probe in fluoroscopy images, i.e. 2D position, in-plane orientation, and scale.

This article is an extended version of Heimann et al. (2013); it explains the methodology in more detail and adds a number of new experiments to the domain adaptation. While based on the same image data, this new version uses updated, more accurate annotations for the *in silico* images, which leads to slightly different results in the evaluation. We start with presenting the basic learning method in the next section and explain our adaptation approach afterwards.

2. Learning from synthetic data

2.1. Generation of in silico images

The synthetic training data is based on digitally reconstructed radiographs, which approximate X-ray images from computed tomography (CT) volumes. The source is a high-resolution (0.18 mm/voxel) isotropic C-arm CT of the TEE transducer, which was aligned to the image axes and cropped to contain only the probe head. A binary mask of the transducer was prepared and multiplied with the original volume to remove streak artifacts in the surrounding air. Fig. 1 shows the final transducer volume in three-plane view and volume visualization.

For each synthetic image, we set up a virtual scene that represents a realistic C-arm geometry. The camera is located 120 cm away from the image plane and features a view angle between 6.5 and 11 degrees, simulating different zoom modes of the C-arm. The 3D position and three Euler angles of the virtual transducer are randomized with the constraints that (a) the probe is located at a distance between 33 and 47 cm away from the image plane, (b) the projected probe is completely inside the image frame, and (c) the probe head is oriented in inferior direction. The flexible tube to which the probe is attached is modeled by a 3D spline originating from a random position at the upper image boundary. Along this spline, a collection of rings is positioned in regular pattern. This is consistent with *in vivo* images captured during structural heart procedures.

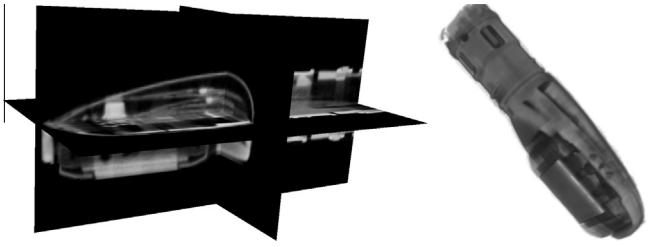

Fig. 1. 3D visualization of the processed C-arm CT volume of an X7-2t 3D TEE transducer (Philips, The Netherlands).

2D projections are generated using a composite ray-caster, i.e. every pixel is assigned the sum of all values along the respective ray through the volume. Key to generating realistic-looking images is the transfer function used to calculate the opacities along the ray. Based on the appearance of *in vivo* images, we chose an exponential transfer function with randomized parameters in order to generate sequences with slightly varying appearance and contrast. A gray value $x > 0$ in the TEE volume maps to opacity $\alpha(x)$ as follows:

$$\alpha(x) = c_0\left(\exp\left(\frac{x}{c_1}\right) - 1\right) \Big/ \left(\exp\left(\frac{7500}{c_1}\right) - 1\right) \quad (1)$$

with $c_0 \in [0.08, 0.12]$ setting the opacity for gray value 7500 and $c_1 \in [2200, 3800]$ setting the contrast as randomized parameters. Fig. 2 shows some example curves for different values of c_0, c_1.

As background, we used 12 cardiac fluoroscopy sequences without transducer and combined them with the generated ray-caster images by additive blending. Annotations were created automatically by storing the 2D position of a fixed point in the center of the transducer together with the respective Euler angles and the probe scale. Since the apparent size in the projected 2D image varies with the rotation angles, scale is measured as the width of upper-most, circular part of the transducer which connects to the flexible tube. Fig. 3 gives an impression of the look of the generated images compared to *in vivo* data.

2.2. Transducer localization by discriminative learning

Following the marginal space learning approach (Zheng et al., 2008), transducer localization is performed in several stages by a pipeline of three discriminative classifiers. The first classifier Φ employs Haar-like features x_H (Viola and Jones, 2004) to determine the 2D position of the probe in images rescaled to 1 mm isotropic pixel spacing. All pixels closer than 1 mm to the reference annotation are labeled as $y = Y^+$, all others as $y = Y^-$. During detection, the 50 candidates with the highest classifier output $\hat{p}_\Phi(y = Y^+|x_H)$ are passed on to the in-plane orientation detector Θ.

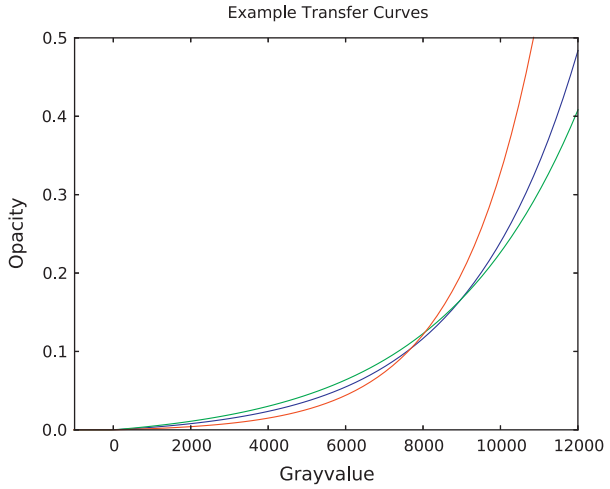

Fig. 2. Examples for different transfer curves used for generating *in silico* images.