Learning without Labeling: Domain Adaptation for Ultrasound Transducer Localization

Tobias Heimann[1], Peter Mountney[2], Matthias John[3], and Razvan Ionasec[2]

[1] Siemens AG, Corporate Technology, Erlangen, Germany
[2] Siemens Corporation, Corporate Technology, Princeton, NJ, USA
[3] Siemens AG, Healthcare Sector, Forchheim, Germany

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 transform between both imaging systems, we employ a discriminative learning based approach to localize the TEE transducer in X-ray images. Instead of time-consuming manual labeling, we 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. An evaluation on more than 1900 images reveals that our approach reduces detection failures by 95% compared to cross validation on the test set 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 from the available imaging information. X-ray fluoroscopy is the dominant imaging modality for these interventions, increasingly supported by 3D trans-esophageal echography (TEE) [2]. Both modalities show complementary information, but in clinical practice they are controlled and displayed completely independently from each other.

Recently, image fusion was 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 tracking (EMT) [3], but this approach requires EMT hardware to be attached to the transducer and is sensitive to EM field distortions. Alternatively, the pose of the transducer can be estimated from its appearance in the X-ray images, either directly [2,6] or supported by fiducial markers attached to the probe head [4]. Since the former approach does not require additional hardware, it is advantageous for integration into the clinical workflow, albeit more challenging to

K. Mori et al. (Eds.): MICCAI 2013, Part III, LNCS 8151, pp. 49–56, 2013.

implement. While 2D-3D registration [2] yields accurate results, it has a limited capture range of < 10 mm, requiring a manual initialization every time a new fluoroscopy sequence is acquired. Discriminative learning (DL) [6] 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, which is based on *in silico* training data that can be generated automatically in great quantities with perfectly accurate labels. To adapt the system to *in vivo* fluoroscopy data, we employ unsupervised domain adaptation, a technique which is widely used in speech processing and has recently gained attention in the computer vision community [5,1]. 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 [11]. 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. 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.

For each synthetic image, the 3D position and three Euclidean angles of the transducer are randomized with the constraint that the probe is oriented in posterior direction. The flexible shaft of the probe 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. 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. As background, we used a number of 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. Figure 1 gives an impression of the look of the generated images compared to *in vivo* data.

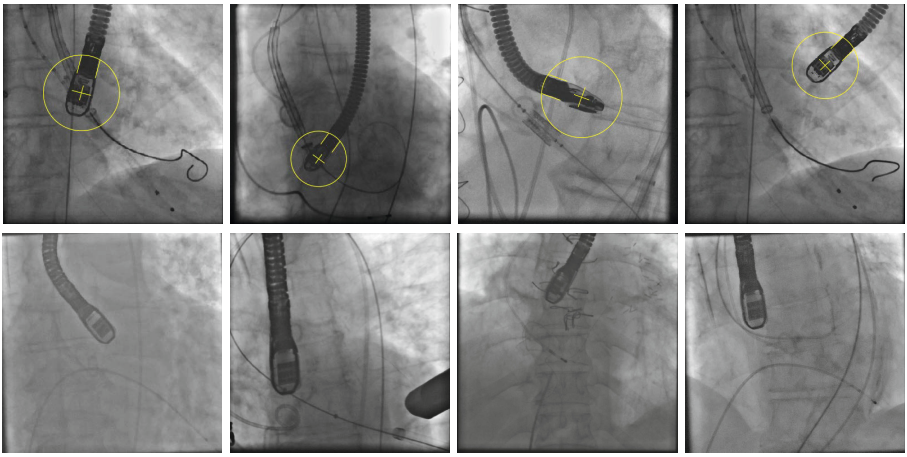

Fig. 1. A selection of generated *in silico* images with automatic labeling (top row) and *in vivo* fluoroscopy images (bottom row)

2.2 Transducer Localization by Discriminative Learning

Following the marginal space learning approach [10], transducer localization is performed in several stages by a pipeline of three discriminative classifiers. The first classifier Φ employs Haar-like features x_H 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 orientation detector Θ. Θ is based on steerable features x_S [10] calculated at 0.25 mm isotropic resolution. Possible angles of the transducer are discretized into 6° steps, and all correctly positioned samples deviating $< 4°$ from the annotated angle are labeled as Y^+. For test images, the 50 candidates with the highest $\hat{p}_\Theta(y = Y^+|x_S)$ are passed on to scale detector Ψ. Ψ is again based on steerable features x_S with 0.25 mm spacing and selects the most probable size of the transducer from a set of 9 hypotheses, corresponding to feature window sizes from 30–46 mm. Lastly, the 50 highest-ranked candidates are combined by weighted averaging according to their respective $\hat{p}_\Psi(y = Y^+|x_S)$ and produce the final output. All classifiers of the pipeline are implemented as probabilistic boosting trees (PBTs) [9], which combine high computational efficiency with competitive accuracy.

3 Domain Adaptation

A fundamental assumption in machine learning is that training and test data stem from the same distribution. In our approach, however, the training data originates from the *in silico* source domain S, while the test data comes from the *in vivo* target domain T. Consequently, the above assumption may not hold, in which case the classifiers would work along non-optimal decision boundaries.

Let x represent a feature vector for a sample and $y \in [Y^+, Y^-]$ its label, then the joint probability distribution $P(y, x)$ should be identical for source and target domain. In our case, we know that the marginalized label probabilities are equal, i.e. $P_S(y) = P_T(y)$, since images from both domains show exactly one transducer. Moreover, given a certain feature vector, the question if the corresponding image region shows a probe can also be decided without knowing its domain, which makes it reasonably safe to assume that $P_S(y|x) = P_T(y|x)$. However, the distribution of feature vectors in both domains is probably different, i.e. $P_S(x) \neq P_T(x)$, which leads to a situation called covariate shift [7].

3.1 Learning under Covariate Shift

As described by Shimodaira [7], a classifier can be adapted to different training and test distributions by minimizing its loss function. This is accomplished by assigning each training sample an instance weight according to the ratio of joint probabilities. Under covariate shift, this ratio simplifies to:

$$\frac{P_T(y,x)}{P_S(y,x)} = \frac{P_T(x)P_T(y|x)}{P_S(x)P_S(y|x)} = \frac{P_T(x)}{P_S(x)} \tag{1}$$

Conveniently, this formulation does not include any labels y, i.e. no annotations are required for the target domain in order to adapt the classifier.

There exist a number of approaches to estimate the required density ratio [8]. In this work, we employ the probabilistic classification approach, in which a classifier is trained to differentiate between samples $x_S \in S$ and $x_T \in T$. Among different types of classifiers, logistic regression is especially well suited for this task [8]. During training, all x_S are assigned to $y = 1$ and all x_T to $y = 0$. The density ratio can then be estimated using classifier output $\hat{p}$ by:

$$\frac{P_T(x)}{P_S(x)} = \frac{1}{\hat{p}(y=1|x)} - 1 \tag{2}$$

3.2 Instance Weighting for Object Localization

While instance weighting has already been employed for a number of different tasks [5], its application to object localization raises two important questions: Which samples should be used to train the logistic regression classifier, and should positive and negative samples be treated equally for weighting? Using all available samples would mean to extract feature vectors for every pixel in every available image multiple times (for different orientation and scale hypotheses). Not only would this result in the impractical amount of 10^{12} feature vectors, but it would also lead to highly unbalanced class labels Y^+ and Y^-. Moreover, as we use a relatively small number of background sequences to generate the *in silico* data, features for Y^- are repeating in the source domain. In summary, this would lead to background samples Y^- completely dominating the logistic regression, while it is the appearance of the transducer (labels Y^+) which should ideally drive the domain adaptation.

We propose a two step approach to solve this problem. In order to draw a subset of samples, we employ a DL pipeline trained on *in silico* data to localize the transducer in another set of synthetic images and unlabeled *in vivo* data. As even an average DL system will detect the transducer with reasonable accuracy on the majority of images, this step effectively reverses the class imbalance in favor of positive samples Y^+. Feature vectors for the drawn samples are normalized to zero mean and unit variance over the entire set and used to train the logistic regression. As the quality of the density ratio estimation may vary, we relax instance weights w as suggested by Shimodaira [7]:

$$w(x) = \left(\frac{P_T(x)}{P_S(x)}\right)^c \tag{3}$$

with $c \in [0..1]$ as regularization parameter. In this study, we set $c = 0.5$.

The domain adapted classifier is then trained on the *in silico* set used as test data in the first step. For each image of this set, the feature vector x of the drawn sample is used in Eq. 3 to estimate the instance weights for all positive samples. Negative samples remain unweighted.

4 Experiments and Results

4.1 Image Data

Image data originates from two clinical centers and was mostly acquired during standard TAVI procedures. Both centers used an Artis Zeego C-arm system (Siemens AG, Germany) for acquisition of fluoroscopy and an X7-2t 3D transducer (Philips, The Netherlands) for acquisition of TEE. In order to estimate the physical resolution of each fluoroscopy sequence, the pixel spacing of the fluoroscopic detector was divided by the radiologic magnification factor, which accounts for the projection geometry of the C-arm. In order to prevent problems with local feature calculation, we excluded approx. 25% of all frames in which the transducer was too close to the image boundaries. In prospective clinical application, the X-ray window could always be chosen to include the probe entirely, i.e. this data exclusion does not limit the applicability of the proposed approach. In the end, we used 68 sequences from 22 patients for our study, comprising 6280 frames in total. For 37 sequences comprising 1913 frames, the probe head was annotated manually by placing an oriented rectangle over it. We denote this set of annotated *in vivo* images as T_L, while the remaining unlabeled 4367 frames are denoted as T_U. Finally, using the method from Sec. 2.1, we generated two sets S_1, S_2 of 10,000 *in silico* images each. In a small annotation study, the point used for automatic labeling of these sets was selected to best match the center of the rectangle used for manual annotations.

4.2 Selecting the Stages for Domain Adaptation

The first set of experiments was conducted to analyze the effectiveness of domain adaptation (DA) for different stages of the detector pipeline. As baseline system,

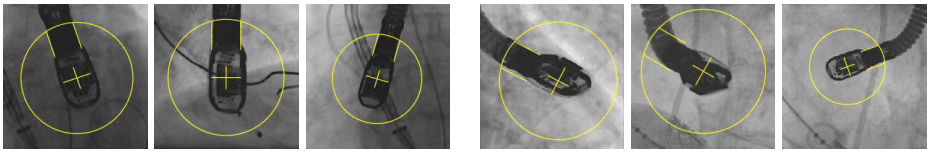

Fig. 2. A selection of *in silico* training samples that received high instance weights (left) and low instance weights (right) for the position detector

we first trained the pipeline presented in Sec. 2.2 on S_1 ($\Phi_0 \Rightarrow \Theta_0 \Rightarrow \Psi_0$). Subsequently, we trained another system on S_2 and used it to draw TEE probe samples from S_1 and T_U. The resulting samples were used to calculate three sets of instance weights for S_1, using the feature set selected by Φ_0, Θ_0, and Ψ_0, respectively. Some examples for samples that obtained very high and low weights are shown in Fig. 2. Training a position detector on the weighted data from S_1 yielded the first domain-adapted classifier Φ_A, which was integrated into pipeline "DA Pos" ($\Phi_A \Rightarrow \Theta_0 \Rightarrow \Psi_0$). Similarly, weighted orientation (Θ_A) and scale detectors (Ψ_A) were trained and included in pipelines "DA Pos+Ori" ($\Phi_A \Rightarrow \Theta_A \Rightarrow \Psi_0$) and "DA Pos+Ori+Scale" ($\Phi_A \Rightarrow \Theta_A \Rightarrow \Psi_A$).

All systems were evaluated on image set T_L. For a detailed analysis of each system, we looked at the detected candidates before the final averaging step and counted a true positive if one of the candidates had a position error < 1 mm, an orientation error $< 4°$, and a scale error < 3 mm. Plotting these counts against the average number of false positives results in the ROC-style curves shown in Fig. 3. The corresponding areas under the curve (AUCs) are given in Table 1.

As can be seen, domain adaptation on the position detector has the largest impact with an increase of 4.5% AUC relative to the baseline system. Domain adaptation on the orientation detector brings only slight additional improvements (+4.8% AUC relative to baseline), while trying to adapt the scale detector deteriorates the results again (only +1.6% AUC relative to baseline remain).

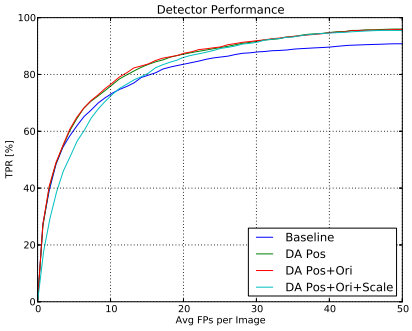

Fig. 3. True positive rate (TPR) vs. average number of false positives (FPs)

Table 1. Area-under-curve values

System	AUC
Baseline	78.3
DA Pos	81.8
DA Pos+Ori	82.0
DA Pos+Ori+Scale	79.5

Table 2. Mean errors with standard deviation for successful detections

	Failed Detections	Position Error	Orientation Error	Scale Error
in vivo Reference	7.34 %	1.5±2.5 mm	3.2±5.4°	3.8±3.0 %
in silico Baseline	2.35 %	0.9±1.1 mm	1.8±1.6°	6.0 (3.2±2.4) %
Domain Adaptation	0.37 %	0.8±0.6 mm	1.7±1.3°	5.7 (3.0±2.3) %
Self Training	1.41 %	0.8±0.8 mm	1.6±1.4°	6.5 (3.0±2.2) %

4.3 Evaluation of Robustness and Accuracy

For the main evaluation, the reference system was trained directly on T_L without any synthetic data (using three-fold cross-validation for evaluation). The baseline system from the previous section, trained exclusively on *in silico* images, came second, and the best-performing domain adaptation ("DA Pos+Ori") third. For the last system, we used the samples drawn from T_U (as described in Sec. 3.2) to enlarge our synthetic training set and generated another unweighted system from $S_1 \cup T_U$. This is a popular approach in semi-supervised learning and called self-training [11]. For each system, the final output of the pipeline (after candidates are merged) was compared to the reference labels. In case the output was located outside the annotated probe area (circles in Figs. 1 & 2), the localization was counted as failure. For successful detections, average position, orientation and scale errors were computed. The complete results are displayed in Table 2. As it turned out, the labels of *in silico* images had a systematic bias of 5% regarding the scale of the transducer; the scale errors in parenthesis show the bias-corrected results. The complete detection pipeline runs in <40 ms per frame, enabling a real-time localization of the transducer in the operating room.

5 Discussion

Our results clearly demonstrate the dependency of DL systems on the available training data. The reference system in our experiments, although trained on the same domain as the test data, yields the worst overall results. The *in silico* system can compensate its different source domain by an eight times larger training set with perfectly placed labels and reduces the number of failed detections by a factor of three, while at the same time improving on all errors. Given these good results, we were surprised by the large impact of domain adaptation, which managed to reduce misdetections yet considerably further down to 5% of the reference system. Its success is based on up-weighting training samples that appear similarly in the target domain and down-weighting less common samples with e.g. very high contrast or large rotations (see Fig. 2). Obviously, generating *in silico* data with more realistic parameters from the start would have a similar effect, but – as for most applications – the true distribution of parameters in real-world data is not known. Since the largest differences between source and target domain appear in the feature set of the position detector (which has to

cope with different orientations and scales), this stage of the pipeline can benefit most from domain adaptation. In order to gain the complete 3D pose of the transducer, our TEE localization can be combined with 2D-3D registration [2] or template-matching [6] to deliver an automatic, robust, and accurate real-time fusion of TEE and fluoroscopy images.

We believe the combination of automatically generated data and unlabeled real-world images to be a highly promising approach for training DL systems. It resolves the need for thousands of annotated training samples, which is one of the main bottlenecks of machine learning in the medical domain. Moreover, the ability to create large quantities of training data for any X-ray imageable device (e.g. implants or new transducers) within hours offers unmatched flexibility.

References

1. Beijbom, O.: Domain adaptation for computer vision applications. Technical report, University of California, San Diego (June 2012)
2. Gao, G., Penney, G., Ma, Y., Gogin, N., Cathier, P., Arujuna, A., Morton, G., Caulfield, D., Gill, J., Rinaldi, C.A., Hancock, J., Redwood, S., Thomas, M., Razavi, R., Gijsbers, G., Rhode, K.: Registration of 3D trans-esophageal echocardiography to X-ray fluoroscopy using image-based probe tracking. Med. Image Anal. 16, 38–49 (2012)
3. Jain, A., Gutierrez, L., Stanton, D.: 3D TEE registration with X-ray fluoroscopy for interventional cardiac applications. In: Ayache, N., Delingette, H., Sermesant, M. (eds.) FIMH 2009. LNCS, vol. 5528, pp. 321–329. Springer, Heidelberg (2009)
4. Lang, P., Seslija, P., Chu, M.W.A., Bainbridge, D., Guiraudon, G.M., Jones, D.L., Peters, T.M.: US - fluoroscopy registration for transcatheter aortic valve implantation. IEEE Trans. Biomed. Eng. 59(5), 1444–1453 (2012)
5. Margolis, A.: A literature review of domain adaptation with unlabeled data. Technical report, University of Washington (2011)
6. Mountney, P., Ionasec, R., Kaiser, M., Mamaghani, S., Wu, W., Chen, T., John, M., Boese, J., Comaniciu, D.: Ultrasound and fluoroscopic images fusion by autonomous ultrasound probe detection. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 544–551. Springer, Heidelberg (2012)
7. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Statistical Planning and Inference 90, 227–244 (2000)
8. Sugiyama, M., Suzuki, T., Kanamori, T.: Density ratio estimation: A comprehensive review. In: Proc. Workshop on Statistical Experiment and Its Related Topics, Kyoto, Japan, pp. 10–31 (March 2010)
9. Tu, Z.: Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: Proc. ICCV, vol. 2, pp. 1589–1596 (October 2005)
10. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D.: Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans. Med. Imaging 27(11), 1668–1681 (2008)
11. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison (July 2008)