Multi-Center MRI Carotid Plaque Component Segmentation Using Feature Normalization and Transfer Learning

Arna van Engelen*, Anouk C. van Dijk, Martine T. B. Truijman, Ronald van't Klooster, Annegreet van Opbroek, Aad van der Lugt, Wiro J. Niessen, M. Eline Kooi, and Marleen de Bruijne

***Abstract*—Automated segmentation of plaque components in carotid artery magnetic resonance imaging (MRI) is important to enable large studies on plaque vulnerability, and for incorporating plaque composition as an imaging biomarker in clinical practice. Especially supervised classification techniques, which learn from labeled examples, have shown good performance. However, a disadvantage of supervised methods is their reduced performance on data different from the training data, for example on images acquired with different scanners. Reducing the amount of manual annotations required for each new dataset will facilitate widespread implementation of supervised methods. In this paper we segment carotid plaque components of clinical interest (fibrous tissue, lipid tissue, calcification and intraplaque hemorrhage) in a multi-center MRI study. We perform voxelwise tissue classification by traditional same-center training, and compare results with two approaches that use little or no annotated same-center data. These approaches additionally use an annotated set of different-center data. We evaluate 1) a nonlinear feature normalization approach, and 2) two transfer-learning algorithms that use same and different-center data with different weights. Results showed that the best results were obtained for a combination of feature normalization and transfer learning. While for the other approaches significant differences in voxelwise or mean volume errors were found compared with the reference same-center training, the proposed approach did not yield significant differences from that reference. We conclude that both extensive feature normalization and transfer learning can be valuable for the development of supervised methods that perform well on different types of datasets.**

***Index Terms*—Atherosclerosis, carotid, classification, magnetic resonance imaging (MRI), segmentation, transfer learning.**

Manuscript received November 01, 2014; revised December 17, 2014; accepted December 17, 2014. Date of publication December 19, 2014; date of current version May 29, 2015. This work was performed within the framework of CTMM, the Center for Translational Molecular Medicine (www.ctmm.nl), project PARISk (Grant 01C-202), and supported by the Dutch Heart Foundation. W. Niessen and M. de Bruijne were financially supported by the Netherlands Organisation for Scientific Research (NWO). *Asterisk indicates corresponding author.*

*A. van Engelen is with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: a.vanengelen@erasmusmc.nl).

A. C. van Dijk is with the Department of Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands, and also with the Department of Neurology, Erasmus MC, 3000 CA Rotterdam, The Netherlands.

M. T. B. Truijman is with the Department of Radiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University Medical Center, 6229 HX Maastricht, The Netherlands, and also with the Department of Clinical Neurophysiology, Maastricht University Medical Center, 6229 HX Maastricht, The Netherlands .

R. van't Klooster is with the Division of Image Processing, Department of Radiology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.

A. van Opbroek is with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands.

A. van der Lugt is with the Department of Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands.

W. J. Niessen is with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands, and also with the Department of Imaging Science and Technology, Faculty of Applied Sciences, Delft University of Technology, 2600 GA Delft, The Netherlands.

M. E. Kooi is with the Department of Radiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University Medical Center, 6229 HX Maastricht, The Netherlands.

M. de Bruijne is with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands, and also with the Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark.

Digital Object Identifier 10.1109/TMI.2014.2384733

I. Introduction

RUPTURE of atherosclerotic plaques in the carotid artery is one of the main causes of cerebrovascular ischemia [1], [2]. The general consensus is that rupture-prone vulnerable plaques are characterized by a thin or ruptured fibrous cap, a large lipid-rich necrotic core (LRNC), presence of intraplaque hemorrhage (IPH), active inflammation [3]–[5], and little calcification [6]. In current clinical practice the decision to perform surgical treatment is, however, still based on the degree of vessel narrowing as determined by noninvasive imaging [7], [8]. It has been hypothesized that plaque composition can help in assessing the risk of rupture and thereby will improve the selection of patients for intervention [7], [9], [10].

Due to its superior soft-tissue contrast, magnetic resonance imaging (MRI) is the preferred imaging modality to visualize the different tissues in the atherosclerotic vessel wall [11], [12]. The appearance of plaque tissues in different MR image sequences has been well established with respect to histology [13]–[15]. Moreover, plaque components as measured from MRI have been related to future cerebrovascular events [5], [16], [17].

Automated segmentation of plaque components would greatly facilitate possible implementation of carotid MR imaging in daily clinical practice. Several methods have been proposed for this segmentation [18]–[21]. These are all supervised classification methods that used a training set with class labels obtained either from registered histology, or from

manual annotations. All performed voxel classification using MRI intensities, intensity gradients and wall distances as features. Liu *et al.* [18] used Parzen window estimation in a naive-Bayesian network and Hofman *et al.* [19] compared different approaches of which a quadratic Bayesian classifier performed best, while we [20], [21] used a linear discriminant classifier. These methods obtained reasonable to good results, varying between components. However, a limitation of such supervised methods is that they specifically assume that training and target data follow the same distribution. This raises problems when training and target data are different, for example when the MR sequence protocol changes, a scanner is replaced, or in multi-center studies. In these situations images typically have different contrast characteristics. The purpose of this study is to develop methods that facilitate the application of supervised learning methods to new or unseen data, by acquiring no or only few annotations on the new dataset. Methods like this will facilitate widespread implementation of supervised methods in medical imaging.

Some approaches to overcome this problem have been investigated. Fischl *et al.* [22] incorporated physics of the MRI acquisition into a brain tissue segmentation algorithm. Theoretically, with knowledge of intrinsic tissue properties (T1 and T2 relaxation time, and tissue proton density), tissue appearance can be modelled given any MR settings. However, these intrinsic properties are often unknown. Another approach involves image normalization. Normalization by matching the mean and variance of the image intensities or by matching two percentiles from the intensity histogram is commonly used, however, mainly for different imaging sessions on the same scanner (among others in [21]). More elaborate normalization methods have been used to handle differences between scanners or protocols. For example by matching more percentiles from the MRI intensity histogram [23] resulted in better performance of brain tissue segmentation when training and test data came from different MRI contrasts [24]. Artan *et al.* [25] applied a classifier trained on data from one device to data from a different device using iterative classification and intensity rescaling of the target data. For chest radiographs and chest computed tomography (CT), normalization of scans acquired with different settings by splitting and weighting different frequency bands, has shown to improve segmentation performance [26], [27]. These methods all normalize the entire image. In our application we are only interested in a relatively small part of the image, the carotid vessel wall. In the normalization methods mentioned above other structures in the image may have a large effect on normalization. Therefore, instead of normalizing the images, we present a way of feature normalization that is able to handle nonlinear scaling of feature spaces from different sources.

Transfer learning [28] is an approach that is still relatively new to the field of medical image analysis. Transfer learning comprises machine-learning methods designed to better handle differences in distributions, labeling functions, and/or features between training and test data. These methods use training data with different properties (called source data), and in some cases a small set of labeled data that has the same properties as the data to analyze (called target data). For example, on nonmedical data Wu *et al.* [29] used a weighted support vector machine (SVM) and a weighted k-nearest-neighbour (kNN) classifier in which source and target samples are weighted differently. Ablavsky *et al.* [30] present an approach for the segmentation of microscopy images, where an SVM classifier trained on a small set of labeled target data was regularized using an SVM trained on a larger set of source data. For brain tissue segmentation, Van Opbroek *et al.* [31] proposed a reweighting SVM where iteratively a weighted SVM classifier was calculated and the weights of misclassified source samples were reduced. This was done in order to reduce the influence of source samples that contradict the rest of the data, while maintaining both target samples and informative source samples. These examples have shown the advantage of transfer-learning methodologies when little training data from the target data type is available.

In this study we aim to develop methods for plaque-component segmentation on multi-center MRI data that has been acquired on MRI scanners from different vendors, and with significant differences in MRI pulse sequence implementation. In contrast to traditional supervised learning as described in [20], [21], [32], we investigate strategies to improve the performance of supervised methods to segment data with different properties than the training data, i.e., from different centers. We evaluate 1) the performance of voxel classification when training and test data come from the same center as well as from different centers (the traditional reference methods), 2) the performance of transfer learning, where we train on a small number of labeled samples from the target data and a large set of annotated source data, and 3) the effect of extensive feature normalization to improve the performance of cross-center analysis.

II. Methods

An overview of the methods and experimental set up is shown in Fig. 1. In Section II-A we first discuss the general approach used for reference. The incremental changes to improve wider implementation of such methods on different data is discussed in Sections II-B and II-C.

A. General Segmentation Methodology

We used a voxel classification approach to segment tissue components. In such a supervised approach a number of characteristics (features) are computed for each voxel. A model is then built on a training set to assign each voxel to one of the classes. Applying this model to features computed for each voxel in a test dataset results in a segmentation in which each voxel receives one class label. In our experiments we selected all voxels for training and testing from the manually segmented vessel wall.

Before computing features a few preprocessing steps were applied to the images. The scans acquired in center 2 showed a considerable intensity bias field due to coil inhomogeneity (Fig. 2). This was corrected for in all five sequences by N4 inhomogeneity correction [33]. The images from center 1 did not show any coil inhomogeneity in the images, so N4 was not applied to the images from this center. Images from both centers were normalized in order to obtain similar intensity ranges between subjects. Here a region of interest (ROI) of 4×4 cm around the lumen center was identified in each image slice. The 5th% of the intensity histogram in the 3D ROI per image was

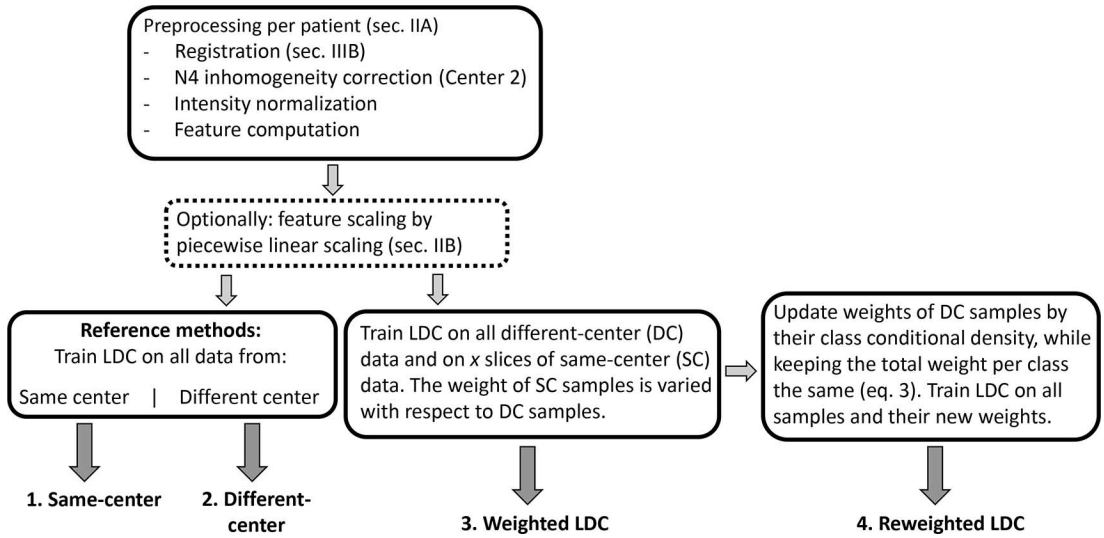

Fig. 1. Flowchart of the proposed methods. After preprocessing, optionally the proposed feature scaling approach is applied, and four approaches are presented: the reference methods same-center training and different-center training, and the two proposed transfer-learning methods.

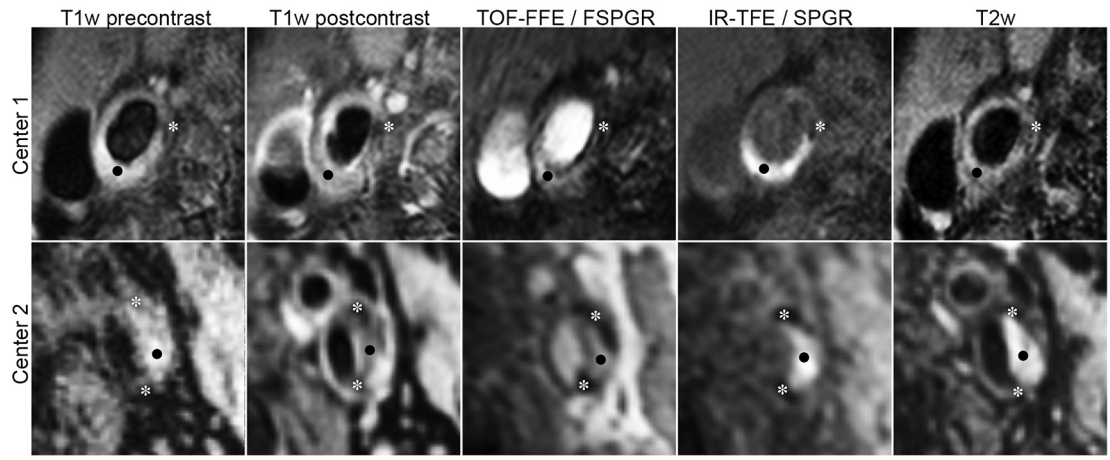

Fig. 2. Examples of registered MR images from both centers. Center 1: A calcium spot (*) appears hypointense on all sequences. A region of IPH (black dot) is hyperintense on the IR-TFE sequence, TOF-FFE, and T1w precontrast, and shows no signal enhancement on the postcontrast T1w image. Center 2: a hyperintense region on the SPGR and T1w precontrast scan indicates IPH (black dot), and is hypointense on the T1w postcontrast scan. Two hypointense regions of calcification (*) are visible in especially the FSPGR and SPGR scans.

set to 0 and the intensity of all voxels was linearly scaled such that the 95th% of the histogram was set to 1000, for each scan individually. We assume this ROI was large enough to exclude any influence from plaque composition to the 5% and 95% histogram values.

Similar to [21], the computed features consisted of 1) image intensities of all MRI sequences, since image intensity is the first main characteristic that differentiates tissue classes, 2) the images blurred with a Gaussian filter [$\sigma = 0.3$ mm ($= 1$ voxel in the data from center 1)] to reduce noise and increase spatial smoothness, 3) the gradient magnitude and Laplacian after blurring at that same scale, as a measure of tissue structure and for detection of small structures, and 4) the Euclidean distances to the lumen and outer vessel wall (mm), and the product of these two distances, to include spatial location within the plaques as a feature. This resulted in a total of 23 features. We used four classes: fibrous tissue (FT), LRNC, CA, and IPH. All classifier training and evaluation was performed using Matlab (Release 2011b, MathWorks, Inc., Natick, MA, USA) and the prtools toolbox [34].

As the classification model, linear discriminant classification (LDC) was chosen for all experiments, since this classifier has proven to be successful for atherosclerotic plaque segmentation previously [20], [21], [32]. With LDC the density of each tissue class is modelled by a normal distribution with equal covariance for all classes, by calculating the mean and covariance of previously calculated features on a training set. The logarithm of the class conditional density ρ_k for class k is defined as follows by LDC [35]:

$$\rho_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k. \quad (1)$$

Here $\mathbf{x}$ is the feature vector to classify, $\boldsymbol{\Sigma}$ the covariance matrix that is pooled over the classes, $\boldsymbol{\mu}_k$ are the class means, and π_k the class prior probabilities. For classification, each sample is assigned to the class with the highest class conditional density ρ_k.

B. Adaptive Histogram Binning

We propose this feature normalization step to account for nonlinear differences in intensity scaling between imaging protocols. We used all voxels from the vessel walls from all patients scanned with the same imaging protocol. An adaptive histogram binning using piece-wise linear rescaling was then applied to each feature independently. Each percentile of samples (voxels) was linearly scaled over one out of 100 equal bins from 0 to 1000. Here we assumed that the data from each center had similar patient characteristics, and hence a similar fraction of FT, CA, LRNC, and IPH voxels was present in the entire dataset from each center. Additionally, for each feature the ordering of tissue components in the imaging protocols was assumed to be the same, while the contrast between tissues may vary between imaging protocols. This procedure performs histogram equalization and also affects the distribution of samples in the feature space as regions in the original histograms with high density are stretched out over more bins.

C. Transfer Learning

We propose two forms of transfer learning for use with LDC, inspired by the sample weighting transfer-learning approaches of Wu *et al.* [29] and Van Opbroek *et al.* [31]. For both approaches the training data was composed of 1) a large labeled dataset acquired in a different center than the data we aim to segment (called the different-center data), and 2) a small number (n) of labeled samples from data acquired in the center for which we aim to segment the data (called same-center data). The labeled samples can for example be obtained by manually indicating a few locations of the different tissue types, or by manually segmenting a number of slices. We propose *weighted-LDC* and *reweighted-LDC*, which both use LDC as provided in (1). In both methods individual samples get different weights, based on their representativeness of the test data. Sample weighting allows tuning the contribution of individual samples to a classifier, and therefore seems an appropriate approach when training data from different sources is used. In case of LDC, this weighting affects the classifier by weighting the estimated class means $\boldsymbol{\mu}_k$ and the pooled covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu}_k^{'} = \frac{\sum\limits_{i:y_i=k} w_i \mathbf{x}_i}{\sum\limits_{i:y_i=k} w_i} \tag{2}$$

$$\boldsymbol{\Sigma}^{'} = \sum_{k=1}^{4} \pi_k \frac{\sum\limits_{i:y_i=k} w_i \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{'}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{'}\right)^T}{\sum\limits_{i:y_i=k} w_i} \tag{3}$$

with y_i the label, and w_i the weight of sample i.

With *weighted-LDC*, we aim to balance the contribution of the large amount of different-center data, and the smaller, but more representative, amount of same-center data to the classifier. To achieve this, samples from the two sources receive different weights. We set the total sum of the sample weights of the different-center data (ΣW_{dc}) and the sum of the sample weights of the same-center data (ΣW_{sc}), while all samples from the same center had the same weight. In our experiments we kept ΣW_{dc} fixed at 1, and varied (ΣW_{sc}).

With *reweighted-LDC*, we assume that part of the different-center data may be more representative of the same-center data than the rest. The sample density for each class may be similar for different- and same-center data in some areas of the feature space, but not for other parts. Therefore, we aim to give a larger weight to the representative different-center samples that provide relevant information, and a lower weight to different-center samples in areas with low same-center density. We first applied weighted-LDC according to (1)–(3) to estimate the density of the classes based on all data. After this step, for each different-data sample $\rho_k(\mathbf{x})$ was determined for its label k, and used as new sample weight. Per class the weights of all samples were linearly rescaled such that their sum equalled the initial total weight of that class. So, also the ratio between ΣW_{sc} and ΣW_{dc} remained the same as for weighted-LDC, only the weights of the different-center data varied between samples

$$w_i(\mathbf{x}_i) = \frac{\rho_{y_i}(\mathbf{x}_i) \cdot \pi_{y_i}}{\sum\limits_{j:y_j=k} \rho_{y_j}}. \tag{4}$$

Here π_{y_i} was determined based on the different-center data only. For reweighted-LDC the classifier was retrained using the updated weights w_i. This way the different-center sample weights are linearly scaled with their corresponding initial class densities.

III. Experimental Set-Up

A. Image Data

We used image data acquired within the multi-center PARISK study [36]. This is a large prospective multi-center imaging study to improve risk stratification in patients with mild to moderate carotid atherosclerosis. Inclusion criteria were a recent (<3 months) transient ischemic attack (TIA), amaurosis fugax or minor stroke, and a symptomatic carotid artery plaque of at least 2–3 mm with a stenosis <70% as determined on Doppler ultrasound or CT angiography. All patients underwent MRI imaging of the carotid artery. For the present study we selected the first 20 patients from the Maastricht University Medical Center (center 1) and the first 22 patients from the Erasmus Medical Center (center 2), for whom a complete MRI session was available. MR imaging was performed on 3.0-T whole-body scanners. Center 1 used an Achieva TX scanner (Philips Healthcare, Best, The Netherlands) with an eight-channel phased-array coil (Shanghai Chenguang Medical Technologies Co., Shanghai, China). Center 2 used the Discovery MR 750 system (GE Healthcare, Milwaukee, MI, USA) with a four-channel phased-array coil with an angulated setup (Machnet B.V., Roden, The Netherlands).

The MRI protocol has been described previously [36], and is summarized in Table I. The main differences between the two centers, apart from the differences in scanner model and coil, are the voxel sizes (both acquired and reconstructed), the use of a T1w IR-TFE (center 1) versus a SPGR scan (center 2) to visualize IPH, and a TOF FFE (center 1) versus a FSPGR (center 2) scan. This FSPGR sequence has been designed specifically to visualize calcification in a single image sequence, making it possible to visually or automatically detect calcification without the

TABLE I
MRI SCAN PARAMETERS

Pulse sequence	T1w QIR TSE (Both pre- and postcontrast)	T1w DIR FSE (Both pre- and postcontrast)	TOF FFE	FSPGR	IR-TFE	SPGR	T2w TSE	T2w DIR FSE
Center	1	2	1	2	1	2	1	2
TR (ms)	800	1RR	20	3.3	9.1	9	4800	2RR
TE (ms)	10	5.2	5	2.1	5.5	1.3	49	50
FA (°)	-	-	20	5	15	30	-	-
Acquired voxel size (mm)	0.62×0.67	0.55×0.71	0.62×0.62	1.00×1.25	0.62×0.63	1.00×1.25	0.62×0.63	0.55×0.71
Reconstructed voxel size (mm)	0.30×0.30	0.55×0.63	0.30×0.30	0.63×0.63	0.30×0.24	0.63×0.63	0.30×0.30	0.55×0.63

* Abbreviations: TR = repetition time, TE = echo time, FA = flip angle, RR = R wave to R wave interval (1 heart beat), QIR = quadruple inversion recovery, TSE = turbo spin echo, DIR = double inversion recovery, FSE = fast spin echo, TOF = time of flight, FFE = fast field echo, FSPGR = fast spoiled gradient echo, IR = inversion recovery, TFE = turbo field echo, SPGR = spoiled gradient echo.

need to combine information from multiple sequences. The TOF scan can identify hypointense regions near the lumen border as calcifications, while they may be considered as lumen on the black-blood sequences. For center 1 for all sequences 15 axial slices were acquired; for center 2 the SPGR and FSPGR images were acquired in the coronal direction. Examples images from both centers are provided in Fig. 2.

B. Manual Reference

Manual contours of the symptomatic artery in each image were obtained for training and validation of the automatic methods. For center 1, the 20 scans were annotated by two observers with three years of experience with carotid MRI, using vessel wall analysis software (MRI-Plaque View, VP-Diagnostics Inc., Seattle, WA, USA). First the images were semi-automatically aligned by registering the four other images to the T1w precontrast scan using the built-in tool for in-plane rigid registration [18]. This registration was manually adjusted for errors. Subsequently, lumen and outer vessel wall were semi-automatically segmented using active contours [37], requiring one lumen seed point, and manually adjusted. Plaque components (CA, LRNC, and IPH) were fully manually delineated, based on previously determined criteria [38]–[40] as agreed on by both observers on beforehand. IPH was defined as a hyperintense area in the IR-TFE scan, LRNC as a region that shows no contrast enhancement on the postcontrast T1w scan and is iso- or hyperintense on the precontrast T1w scan, and CA as hypointense on at least three image sequences. The remaining tissue within the vessel wall was considered as FT. All slices for which the five image sequences were available after co-registration, were annotated. One patient was excluded due to excessive patient movement.

As we used VesselMass (Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands) for further analysis, editing and visualization, the annotated contours were converted into a format for use in VesselMass. Contour coordinates on the fixed T1w image could be extracted from the PlaqueView contours, and rigid registration [41] with manual adjustments was repeated in VesselMass until the alignment of contours with images was determined to be correct by one of the observers. Due to significant inter-observer variability for this center, a consensus contour set was created by five experienced observers including the first two observers, with knowledge of the two initial segmentations, but without knowledge of any automatic segmentation results that were obtained using the first two contour sets. For this dataset segmentation and registration was approved by all observers.

The 22 scans from center 2 were manually annotated by one observer with three years of experience, using VesselMass. A subset of 10 scans was annotated by a second observer with one year of experience. First, the lumen was manually annotated on the T1w precontrast scan. Subsequently the remaining four image sequences were automatically registered to the T1w precontrast scan using a previously described algorithm for 3D rigid registration [41]. All contours (lumen, outer wall, CA, LRNC, and IPH) were fully manually drawn. Similar to center 1, LRNC was defined as a hypointense area on the postcontrast T1w scan that is iso- or hyperintense on the precontrast T1w scan. IPH was defined as a hyperintense region on the SPGR scan. The criterium for CA segmentation was different from center 1: all hypointense regions in the FSPGR sequence were defined as calcium, without taking information from the other image sequences into account. In addition, areas with hypointensity in two or more of the other sequences without hypointensity in the FSPGR sequence were annotated as calcium if this was thought to be related to misregistration of the FSPGR volume. For both centers, by definition, all IPH lesions were drawn within a region of LRNC. However, for our experiments LRNC and IPH were considered mutually exclusive, so the IPH regions were not considered as LRNC as well for classification.

C. Experiments

Five different training approaches for voxel classification were evaluated. For all approaches for center 1, the consensus contours were used for training and evaluation. For center 2 we used the contours of the observer who annotated all 22 datasets.

I Same-center training: Methods were trained and evaluated on data from the same center, and thus acquired using the same hardware and imaging protocol. For both centers we performed leave-one-subject-out cross-validation. This uses the same approach as published state-of-the-art methods [20], [21] in terms of classifier, features and training, and is therefore used as the reference method.

II Different-center training: Here a classifier developed on all vessels from center 1 was applied to segment the data from center 2, and vice versa, without the use of any labeled same-center samples during training. This resembles the situation in which one would apply the same-center classifier to previously unseen data. In this study this represents the reference for the situation in which no fully annotated same-center dataset is available.

III–IV Weighted and reweighted transfer learning: We simulated the situation in which a few slices from a larger set of same-center data are selected and manually segmented. This is practically feasible, and allows the use of transfer-learning methods to tune the segmentation algorithm for use on the same-center data while most of the training data originates from the different-center dataset. In order to do this we selected a number of slices from both datasets that were considered suitable for training in such a setting. The selection criterium was presence of at least one of the three components CA, LRNC and IPH with a size of at least 10 voxels. This led to a selection of 118 out of the 285 slices for center 1 and 128 out of 359 slices for center 2. Experiments were performed with a random selection of 1, 3, 5, and 10 slices of same-center data, where those slices together contained at least 10 voxels of all components. Those slices were randomly selected from the other vessels from the target data in a leave-one-patient-out fashion. For each slice set selection, weighted- and reweighted-LDC were performed with five different settings of weighting between different-center and same-center data. ΣW_{sc} was set to 0.1, 0.2, 1, 5, and 10, while ΣW_{dc} was always 1. The prior probabilities for the classes (π_k) were set to the prior obtained from the fully annotated different-center dataset in all experiments. These experiments were repeated 100 times, to account for the variability between slice set selections. In each of the 100 iterations all vessels from the target data were segmented once.

V Training on x same-center slices. For comparison we also performed segmentation by training on the 1, 3, 5, or 10 selected same-center slices only, without adding different-center data. This was done for all 100 repetitions. Similar to the transfer-learning experiments, the prior probabilities were obtained from the different-center dataset for a more accurate comparison.

D. Evaluation

Segmentation results were evaluated by 1) two-way intraclass correlation coefficients (ICC) for the volume of FT, CA, LRNC, and IPH per vessel, 2) the error per component: the absolute difference between the amount of that component in the ground truth and the segmentation result per vessel, 3) accuracy as % of correctly classified voxels, and 4) confusion matrices. All experiments were performed both with and without feature normalization by adaptive histogram binning, to also assess the effect of adaptive histogram binning on same-center training. In the transfer-learning experiments with a limited number of same-center slices, the entire set of same-center data was used to determine the histogram bins and normalization of the individual slices was performed using these parameters. Results were compared with the inter-observer variability as determined between the two observers and the consensus reading for center 1, and between the two observers for center 2. For a more fair comparison of the automatic segmentation results with the inter-observer variability, the contours of the observers were evaluated within the consensus contours for the vessel wall for center 1, and for center 2 the contours of observer 2 were evaluated within the vessel wall contours of observer 1. Only voxels annotated within this wall were considered, and voxels not annotated within this wall were considered to be fibrous tissue. This is similar to how the automatic segmentation works, which also takes the reference vessel wall as an input.

Statistical comparisons were made between 1) same-center training, 2) different-center training, 3) different-center training with adaptive histogram binning, 4) training on a few same-center slices, 5) transfer learning, and 6) transfer learning with adaptive histogram binning. We compared the mean error of the four components, and the voxelwise accuracy per vessel. The analysis was done for the two centers combined, for the setting where 5 labeled same-center slices are available. For transfer learning we selected the method (weighted- or reweighted LDC and ΣW_{sc}) that overall performed best. For this method for each patient we took the median error and voxelwise accuracy for each vessel over the 100 repeated experiments to use in the statistical analysis. Comparisons were made by Friedman analysis, followed by Tukey-Kramer testing for individual differences taking multiple comparisons into account. A p-value <0.05 was considered significant.

IV. Results

In this section, we will first present the results obtained by the reference method [21], both for same-center (Section IV-A) and different-center (Section IV-B) training and thereafter the results using transfer learning (Section IV-C). All results are presented with and without feature normalization by adaptive histogram binning. An overview of all results on all 41 subjects of the two centers combined is provided in Table III. A subset of the results is provided for the two centers separately (Table IV).

A. Same-Center Training

Correlations of tissue component volumes per vessel for the two centers combined, using same-center training without adaptive histogram binning are provided in the top row of Table III. Good ICC values were obtained for FT, CA, and IPH. A lower correlation was found for LRNC. Table IV indicates that a good correlation for LRNC was obtained for the data from center 1, but a considerable underestimation with low correlation was obtained for center 2. A confusion matrix of both centers combined is provided in Table II(a) to assess voxelwise agreement. This shows a low sensitivity for LRNC (15%) and a moderate sensitivity for CA (43%) which both were often misclassified as FT. A good sensitivity for IPH (73%) and a high sensitivity for FT (97%) were found.

Further results (volume errors, ICC values, and accuracy) for same-center training are summarized in Tables III (centers combined) and IV (two centers separately). The accuracy of the automated same-center methods was similar to the inter-observer agreement for both centers (Table IV). However, for center 2 the errors were slightly larger than the differences between observers.

The results for training on 5 same-center slices only are also provided in Table III. The obtained volume errors were similar to same-center training on the full dataset. However, a lower ICC for CA and IPH was obtained, and the voxelwise accuracy

TABLE II
CONFUSION MATRICES SHOWING AGREEMENT BETWEEN THE GROUND TRUTH AND SEGMENTATION RESULTS. EACH VALUES IS GIVEN AS A PERCENTAGE OF THE TOTAL AMOUNT OF VOXELS INCLUDED. SENSITIVITY IS PROVIDED FOR EACH COMPONENT AS WELL. (A) SAME-CENTER TRAINING, (B) DIFFERENT-CENTER TRAINING, (C) DIFFERENT-CENTER TRAINING AND ADAPTIVE HISTOGRAM BINNING, (D) PROPOSED TRANSFER-LEARNING APPROACH: WEIGHTED-LDC WITH $\Sigma W_{\mathrm{sc}} = 0.1$ AND ADAPTIVE HISTOGRAM BINNING. FOR (D) THE MEAN OF 100 REPETITIONS IS TAKEN. AHB = ADAPTIVE HISTOGRAM BINNING

		Same-center result				
		FT	LRNC	CA	IPH	Sensitivity
Ground truth	FT	83.7	0.8	1.7	0.3	97%
	LRNC	2.3	0.5	0.2	0.3	15%
	CA	3.2	0.0	2.6	0.2	43%
	IPH	0.7	0.4	0.0	3.0	73%

(a)

		Different-center result				
		FT	LRNC	CA	IPH	Sensitivity
Ground truth	FT	74.8	0.7	10.7	0.4	86%
	LRNC	1.9	0.4	0.7	0.3	12%
	CA	3.2	0.1	2.5	0.2	42%
	IPH	0.6	0.5	0.1	2.9	71%

(b)

		Different-center + AHB result				
		FT	LRNC	CA	IPH	Sensitivity
Ground truth	FT	82.0	0.7	2.8	1.0	95%
	LRNC	2.0	0.7	0.2	0.4	21%
	CA	3.6	0.1	2.1	0.2	35%
	IPH	1.1	0.2	0.1	2.8	67%

(c)

		Transfer-learning result				
		FT	LRNC	CA	IPH	Sensitivity
Ground truth	FT	82.9	0.6	2.1	1.1	96%
	LRNC	2.1	0.6	0.2	0.4	14%
	CA	3.5	0.1	2.2	0.1	37%
	IPH	1.1	0.2	0.0	2.8	68%

(d)

(median 88%) was significantly lower than for reference same-center training (median 91%).

B. Different-Center Training

Training on only different-center data resulted in an extreme overclassification of FT as CA, resulting in large errors for FT and CA (Tables III and IV). This can also be seen from the confusion matrix in Table II(b) and was mainly due to a large overclassification of FT as CA for center 1. This can be explained since the FSPGR scan from center 2 has low intensity for calcification only, while the corresponding TOF from center 2 shows low intensity in almost the entire vessel wall. The errors were much lower when adaptive histogram binning was applied. The confusion matrix in Table II(c) shows that a slight overclassification of CA remains, but a large improvement with respect to Table II(b) is seen. Statistical analysis of the combined errors and voxelwise accuracy showed that different-center training without adaptive histogram binning had significantly larger volume errors and lower accuracy than same-center training. Despite the large improvement obtained by adaptive histogram binning, the error and accuracy remained significantly different from same-center training.

C. Transfer Learning

Results for the transfer-learning experiments, for the two centers combined, are summarized in Fig. 3. It can be seen that using most approaches, and having at least three same-center slices, reasonable volume errors were obtained. The effect of the number of slices, and the weight given to the same-center data differs between approaches. For weighted-LDC, these parameters did not have a large influence on the volume errors. ICC values were more sensitive to the same-center weight, especially after adaptive histogram binning. Here giving little weight to the same-center data yielded the most accurate results, suggesting that not enough same-center data is available to let it contribute equal to or more than the large amount of different-center data. This is supported by the fact that ICC decreases faster when less same-center slices are used.

With reweighted-LDC, the lowest errors were obtained when same- and different-center data were given the same weight. Reweighted-LDC was also more strongly dependent on the amount of same-center data available. This indicates that for reweighting it is more important to have enough same-center data to accurately model the data and to adjust the different-center sample weights accordingly.

Based on these findings, in Table III the results are specified per component for weighted-LDC with $\Sigma W_{\mathrm{sc}} = 0.1$, and reweighted-LDC with $\Sigma W_{\mathrm{sc}} = 1$, both with and without adaptive histogram binning. After considering volume error, ICC and accuracy together, we decided to focus on weighted-LDC with $\Sigma W_{\mathrm{sc}} = 0.1$ and adaptive histogram binning for further analysis. The corresponding confusion matrix is given in Table II(d) and shows results very similar to same-center training. When looking at the centers individually in Table IV, the improvement over different-center AHB is clear for the accuracy in center 1. For center 2 the differences are smaller. In the statistical analysis we included weighted-LDC with $\Sigma W_{\mathrm{sc}} = 0.1$, both with and without adaptive histogram binning. The mean volume error (mm^3) of both transfer-learning approaches, but also of training on five same-center slices only as mentioned above, was not significantly different from full same-center training. This error was significantly larger when a conventional classifier was trained on different-center data, either with or without adaptive histogram binning. More importantly, only the voxelwise accuracy of transfer learning and adaptive histogram binning combined was not significantly lower than for same-center training. The accuracy of transfer learning with adaptive histogram binning was also significantly better than transfer learning without adaptive histogram binning, than training on five same-center slices, and than training on different-center data only.

D. Visualization of Results

Segmentations for three slices from both centers are shown in Fig. 4. The results for transfer learning were obtained using five same-center slices, adaptive histogram binning on the features and weighted-LDC with $\Sigma W_{\mathrm{sc}} = 0.1$. Of the 100 repeated experiments with random selection of five target slices, for each vessel we used the selection for which the total error over the four components was closest to the median total error of the 100 experiments for the examples shown. The segmentations

TABLE III
AUTOMATED SEGMENTATION RESULTS INCLUDING ALL 41 PATIENTS

	Median error mm^3 (IQR)				ICC (95% Confidence interval)				Accuracy (%)
	FT	LRNC	CA	IPH	FT	LRNC	CA	IPH	(Median (IQR))
Same-center (=Ref)	55 (23-88)	22 (7-34)	28 (15-50)	0 (0-25)	0.96 (0.93-0.98)	0.43 (0.14-0.65)	0.86 (0.76-0.93)	0.96 (0.93-0.98)	91 (87-95))
Same-center AHB	43 (17-67)	20 (7-34)	22 (9-36)	2 (0-25)	0.97 (0.95-0.99)	0.68 (0.47-0.81)	0.90 (0.83-0.95)	0.96 (0.92-0.98)	91 (85-95)
Different-center (=Ref)	92 (57-178)	22 (10-41)	93 (47-183)	1 (0-26)	0.88 (0.78-0.93)	0.57 (0.32-0.74)	0.25 (-0.06-0.51)	0.93 (0.87-0.96)	80 (75-90)
Different-center AHB	63 (28-105)	13 (5-36)	34 (20-68)	6 (0-22)	0.94 (0.89-0.97)	0.76 (0.59-0.86)	0.64 (0.41-0.79)	0.95 (0.90-0.97)	88 (84-95)
5 target slices	49 (23-94)	20 (8-38)	30 (15-63)	4 (0-27)	0.94 (0.90-0.97)	0.44 (0.16-0.66)	0.65 (0.43-0.80)	0.88 (0.78-0.93)	88 (82-94)
5 target slices, AHB	49 (21-104)	19 (7-39)	28 (13-60)	8 (1-31)	0.93 (0.86-0.96)	0.45 (0.17-0.67)	0.64 (0.42-0.79)	0.75 (0.57-0.86)	88 (82-94)
Transfer Learning									
Weighted, ΣW_{sc}=0.1	51 (28-89)	22 (9-38)	47 (21-70)	0 (0-27)	0.94 (0.90-0.97)	0.56 (0.30-0.74)	0.59 (0.35-0.76)	0.94 (0.88-0.97)	88 (82-94)
Weighted, ΣW_{sc}=0.1, AHB	53 (26-110)	16 (7-37)	28 (9-57)	6 (0-24)	0.95 (0.91-0.97)	0.73 (0.54-0.84)	0.72 (0.54-0.84)	0.95 (0.90-0.97)	90 (85-95)
Reweighted, ΣW_{sc}=1	48 (21-89)	17 (7-33)	31 (14-65)	4 (0-27)	0.94 (0.90-0.97)	0.57 (0.32-0.75)	0.70 (0.50-0.83)	0.91 (0.83-0.95)	88 (83-94)
Reweighted, ΣW_{sc}=1, AHB	46 (20-90)	15 (6-32)	26 (12-52)	11 (1-31)	0.95 (0.90-0.97)	0.71 (0.52-0.84)	0.80 (0.65-0.89)	0.88 (0.78-0.93)	88 (83-94)

* For transfer learning all results are given as the median over the 100 experiments where 5 labeled same-center slices were used. FT = Fibrous tissue, LRNC = Lipid-rich necrotic core, CA = Calcification, IPH = Intraplaque hemorrhage, IQR = Interquartile range, ICC = Intraclass correlation coefficient, AHB = Adaptive histogram binning.

TABLE IV
SEGMENTATION RESULTS PER CENTER

	Median error mm^3 (IQR)				ICC (95% Confidence interval)				Accuracy (%)
	FT	LRNC	CA	IPH	FT	LRNC	CA	IPH	(Median (IQR))
*Center 1**									
Inter-observer									
Obs1 – Obs2	58 (25-72)	34 (12-75)	24 (11-37)	1 (0-23)	0.96 (0.89-0.99)	0.34 (-0.17-0.71)	0.42 (-0.1-0.75)	0.19 (-0.31-0.62)	89 (85-94)
Obs1	41 (21-61)	20 (13-35)	17 (7-28)	0 (0-16)	0.97 (0.92-0.99)	0.75 (0.42-0.91)	0.87 (0.67-0.95)	0.26 (-0.25-0.66)	90 (89-94)
Obs2	14 (7-26)	10 (5-32)	7 (1-20)	0 (0-14)	0.99 (0.96-0.99)	0.71 (0.39-0.88)	0.81 (0.56-0.92)	0.92 (0.82-0.97)	94 (92-97)
Automated									
Same-center (=Ref)	37 (9-64)	16 (1-24)	23 (12-33)	0 (0-15)	0.98 (0.65-0.99)	0.93 (0.83-0.97)	0.66 (0.31-0.85)	0.96 (0.90-0.99)	91 (90-95)
Same-center AHB	27 (13-50)	15 (3-29)	15 (7-25)	2 (0-24)	0.96 (0.91-0.99)	0.77 (0.50-0.91)	0.83 (0.62-0.93)	0.76 (0.48-0.90)	93 (89-95)
Different-center (=Ref)	169 (97-204)	21 (4-33)	185 (143-218)	0 (0-6)	0.90 (0.75-0.96)	0.56 (0.15-0.80)	0.25 (-0.22-0.63)	0.98 (0.94-0.99)	74 (70-79)
Different-center AHB	42 (18-78)	6 (1-13)	24 (10-35)	6 (1-24)	0.93 (0.84-0.97)	0.92 (0.81-0.97)	0.66 (0.30-0.85)	0.66 (0.31-0.85)	87 (83-92)
Weighted, ΣW_{sc}=0.1, AHB	30 (9-55)	11 (2-20)	10 (4-26)	6 (0-28)	0.94 (0.85-0.98)	0.86 (0.67-0.94)	0.70 (0.37-0.87)	0.69 (0.35-0.87)	90 (85-93)
*Center 2 ***									
Inter-observer									
Observer 2	22 (17-37)	8 (7-27)	18 (8-30)	5 (0-18)	0.97 (0.86-0.99)	0.76 (0.29-0.93)	0.75 (0.28-0.93)	0.98 (0.92-0.99)	88 (85-91)
Automated									
Same-center (=Ref)	80 (42-126)	29 (14-53)	38 (22-67)	7 (0-28)	0.95 (0.89-0.98)	0.30 (-0.13-0.64)	0.91 (0.80-0.96)	0.96 (0.90-0.98)	90 (84-95)
Same-center AHB	51 (28-93)	28 (9-40)	32 (12-44)	4 (0-24)	0.98 (0.94-0.99)	0.66 (0.33-0.84)	0.90 (0.78-0.96)	0.98 (0.95-0.99)	90 (85-95)
Different-center (=Ref)	60 (37-92)	24 (10-50)	49 (28-71)	7 (0-31)	0.97 (0.94-0.99)	0.56 (0.19-0.79)	0.85 (0.68-0.94)	0.92 (0.83-0.97)	88 (80-94)
Different-center AHB	86 (53-130)	30 (11-50)	51 (27-69)	6 (0-21)	0.97 (0.94-0.99)	0.76 (0.50-0.89)	0.85 (0.67-0.93)	0.99 (0.97-0.99)	89 (85-95)
Weighted, ΣW_{sc}=0.1, AHB	83 (49-132)	25 (11-52)	49 (27-70)	6 (0-22)	0.97 (0.94-0.99)	0.72 (0.44-0.88)	0.87 (0.71-0.94)	0.99 (0.97-0.99)	90 (86-95)

* All values shown are with respect to the consensus reading. Inter-observer variability is calculated within the annotated consensus vessel wall contours. ** Inter-observer variability is only on 10 vessels. Contours from observer 2 are considered within the annotated vessel wall by observer 1. For transfer learning all results are given as the median over the 100 experiments where 5 labeled same-center slices were used. FT = Fibrous tissue, LRNC = Lipid-rich necrotic core, CA = Calcification, IPH = Intraplaque hemorrhage, IQR = Interquartile range, ICC = Intraclass correlation coefficient, AHB = Adaptive histogram binning.
† In previous studies, IPH was frequently considered as being part of the LRNC. If we combine the segmentations for these components, ICC between the two observers equals 0.53 for center 1 and 0.99 for center 2. For the automatic results the ICC of LRNC and IPH combined is similar or slightly larger than for IPH on itself.

show that same-center training sometimes yields the smoothest results (columns 1, 2, and 5 eg.). The transfer-learning segmentations have a slightly better detection of CA (column 4) and IPH (column 1) than different-center training in some of the examples.

V. DISCUSSION AND CONCLUSION

In this work we performed carotid plaque-component segmentation in a two-center MRI study. Whereas traditional supervised approaches would use a considerable amount of training data from each center, we developed two approaches that were shown to improve segmentation accuracy when only few annotations from the dataset to segment are available. To achieve this, a much larger annotated dataset with slightly different feature distributions is used, in our case from the other center. Our results showed that using extensive feature normalization by adaptive histogram binning, and transfer-learning algorithms, performed significantly better than applying a method trained on different-center data. Moreover, these results were not significantly different from training on the complete set of manually annotated same-center data. For both centers these obtained segmentations showed a similar agreement with manual annotations as the inter-observer agreement.

Applying a reference classifier optimized using training data from one center directly to image data from the other center yielded large errors. The largest errors were obtained for CA, which can mainly be explained by the differences in image acquisition for this component. In center 2, hypointense regions within the furthermore isointense vessel wall in the FSPGR scan were annotated as CA, while for center 1 the corresponding TOF-FFE has an overall hypointense vessel wall, resulting in large CA overestimation when using the classifier developed for center 2. Appearance of IPH and LRNC was more similar between centers and raised fewer problems. Training on only five same-center slices was not significantly different from same-center training or transfer learning when considering the mean volume error of the components, but the voxelwise accuracy was significantly lower. It should also be noted that the

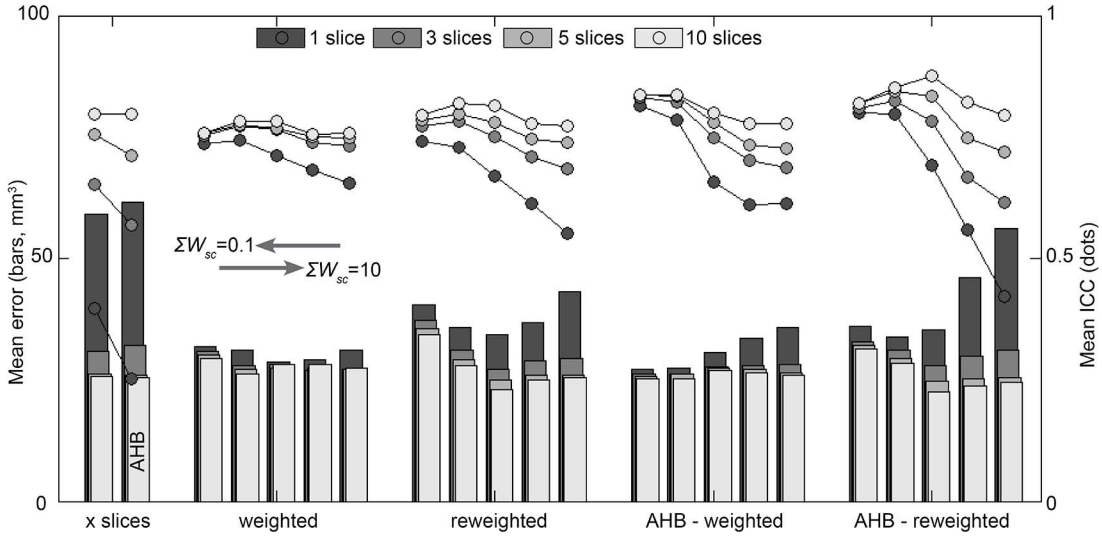

Fig. 3. Mean error (mean of median per component) and ICC when only a limited amount of same-center training data is available. For each set of five transfer learning experiments ΣW_{sc} from left to right is 0.1, 0.2, 1, 5, 10. AHB = adaptive histogram binning. Bars and dots for "x slices" are the result for training on the labeled same-center slices only.

class prior probabilities here were also determined on the different-center data, as the presence of CA, LRNC, and IPH would be overestimated based on the selected slices.

The first method proposed in this paper is piecewise-linear feature normalization by adaptive histogram binning. Normalizing both same- and different-center data by this method yielded a large improvement of segmentations over direct different-center training. This approach was chosen to overcome nonlinear differences in the probability density of features between the two datasets. It is similar to the histogram matching of Nyúl *et al.* [23] for brain tissue segmentation. They perform histogram matching per image and only for the intensity, whereas we normalize all features, for all subjects per center combined. We chose to combine subjects because the tissue distribution in the vessel wall differs more between patients than the tissue distribution of the brain. For example not all classes are present in each of the images. Additionally, a different approach would be to map the percentiles of the features from one center to match the corresponding percentiles obtained from the other center. This has a smaller influence on the density distribution of the feature histograms, but yielded larger errors in a pilot experiment on a small subset of the data than the histogram equalization approach that we propose.

The second proposed method, a transfer-learning classifier with sample weighting, improved accuracy over different-center training, by only obtaining labels for a small number of slices (1–10). Several approaches for transfer learning have been described, of which sample weighting and feature selection and/or transformation are the most common [28]. Sample weighting is an appropriate method to handle a nonlinear change in the distribution of features over the feature space. Adaptive histogram binning can partly solve this, but not completely. Reweighted-LDC can also partly handle differences in the (manual) labeling procedure, for example when different features are used to segment a certain class. It can then reduce the weight of different-center samples that do not correspond to the combined distribution of same- and different-center samples. Feature selection, or learning a low-dimensional representation of features that are similar between centers, is another common transfer-learning approach. However, in our case the features that differ most between centers are essential for accurate classification of all classes, and such an approach would increase the risk of obtaining low accuracy for those classes.

There are certain requirements to the data for both adaptive histogram binning and transfer learning to be successful. One requirement to successfully use adaptive histogram binning is that a representative set of data needs to be available for both centers, such that the distribution of disease stage and therefore prior class probabilities is similar. For both adaptive histogram binning and the used transfer-learning algorithms it is important that there is a direct link between each feature in one dataset with one of the features in the other dataset, with the same ordering of classes. In the multi-center study that we used, the MRI protocol was designed to be comparable between the two centers; differences occurred only due to use of a different scanner and institutional preferences. If no obvious link is present, either only the sequences that are comparable in both datasets can be used, or measures of histogram similarity such as the Kullback-Leibler divergence could be used to determine which sequences are most similar in appearance [42]. The assumption that the classes had the same ordering did not fully hold for the TOF-FFE/FSPGR sequence, since IPH had the highest intensity in the FSPGR scan, but the second-highest (after FT) in the TOF-FFE scan. However, if sufficient weight is given to the same-center data, such features will contribute less to the classifier, which aims to optimize the discrimination between classes.

The performance of transfer learning depends on the selected slices, which need to be representative of all classes in the target data. As a selection criterium we used a minimum of 10 voxels for each class for the total set of selected slices, where each slice has at least one component with at least 10 voxels besides

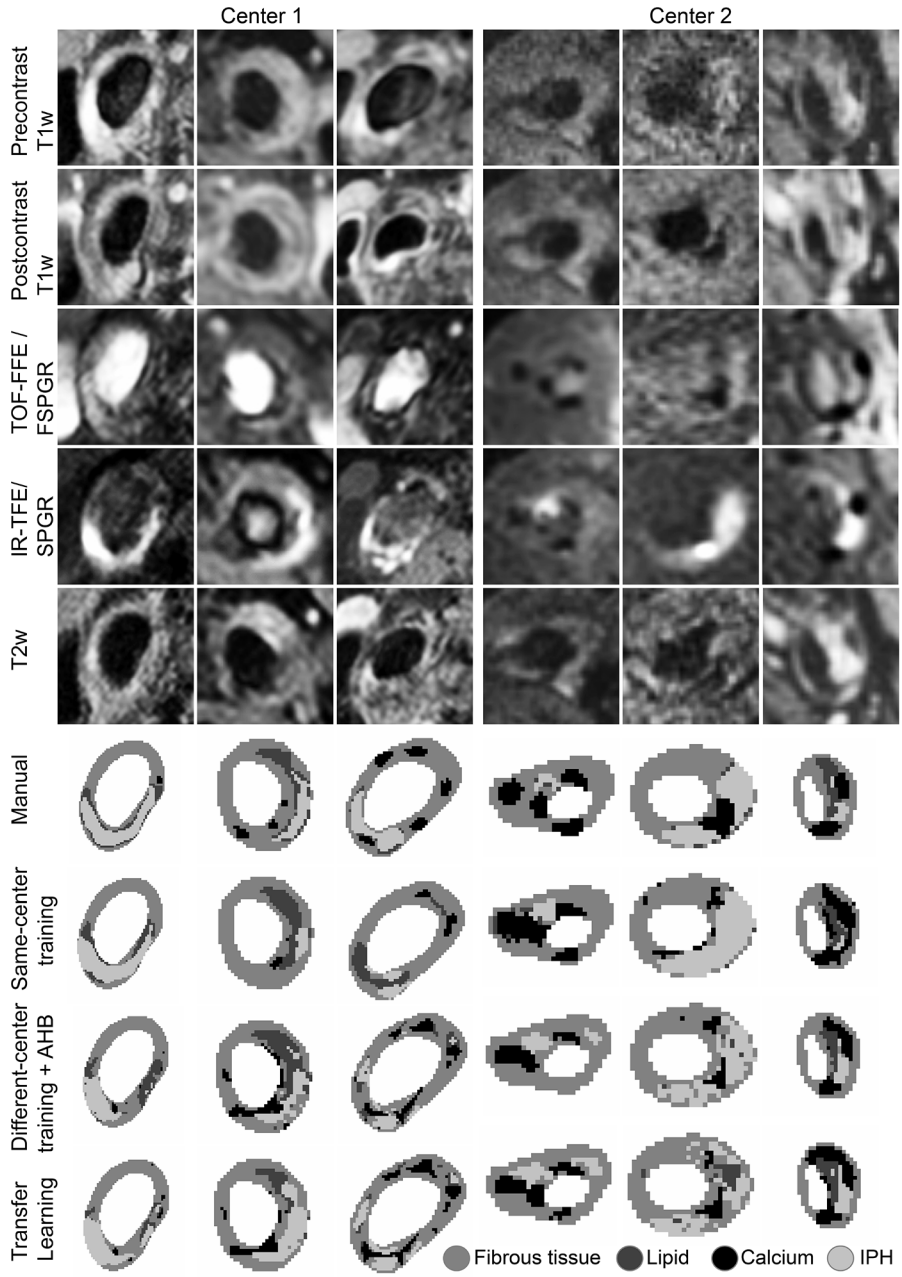

Fig. 4. Segmentation results for slices selected from three patients from center 1 and three patients from center 2. Results for transfer learning were obtained using five selected target slices and both adaptive histogram binning and weighted-LDC with $\Sigma W_{\mathrm{sc}} = 0.1$.

FT. In practice this is not expected to raise problems, as slices with large areas of CA, LRNC or IPH can easily be recognized by human observers. The number of same-center training samples in the smallest class was significantly correlated with the average error for both centers and all transfer-learning experiments combined ($\mathrm{R} = -0.04, \mathrm{p} < 10^{-21}$). The suitability of adaptive histogram binning and the used sample (re)weighting algorithms depends on which classifier is used. LDC is optimal if all classes follow a Gaussian distribution with equal covariance, which cannot be assumed after the performed histogram stretching. In this study adaptive histogram binning did not negatively affect our results for same-center training, except for IPH segmentation in center 1. Concerning the sample (re)weighting, almost all classifiers can deal with sample weighting, however, our reweighting procedure was more tuned to density-based classifiers. For other classifiers, such as SVM, different criteria for reweighting have been used [31] which may be more suitable for that specific classifier. For instance, in [31] misclassified different-data samples received a lower weight. With SVM outliers that are classified correctly have little effect. In LDC, however, these outliers have a large effect on class means and covariance, which is why we decided to lower the weight of those samples.

A combination of adaptive histogram binning and transfer learning performed best, and yielded good results for both centers. Overall, weighted-LDC with adaptive histogram binning and $\Sigma W_{\mathrm{sc}} = 0.1$ performed best and its results did not differ significantly from same-center training. However, the best performing algorithm differed between the two centers. Instead of

using the same transfer-learning approach on every dataset, it may be useful to re-evaluate what method and settings work best on new data. If the proposed approach of weighted-LDC with a relatively low weight on the same-center data yields visually unsatisfactory results, a different weighting-ratio can be considered. Based on our results reweighting seems most appropriate when a considerable amount of same-center annotations can be obtained. The optimum weights vary based on how representative the different-center data is of the same-center data.

In our data there was a slight difference in prior probabilities for the tissue components between the two centers, which could have influenced the classifiers based on adaptive histogram binning, and different-center training. More CA (8% versus 4%) and IPH (5% versus 3%) was manually segmented for center 2. This could be due to a difference in patient population and vessel wall composition in the included patients, but likely also results from differences in the imaging protocol. Although this has not been studied in detail, it seems the FSPGR sequence to image CA in center 2 leads to an overestimation of CA, similar to the blooming effect that is seen in CT. Use of this dedicated sequence to image CA did on the other hand yield larger ICC values for CA than the traditional MR protocol used in center 1. This difference in prior probabilities could also have contributed to the overestimation of CA in center 1 when trained on center 2. However, prior probabilities from the different-center data were used in the transfer-learning experiments as well, where no overestimation of CA was seen. It is therefore likely that differences in the imaging protocol contribute more to the large overestimation of CA in Table II(b) than the difference in prior.

The features used for classification in this study yielded good results on data from both sites, as well as in previous studies using data with a slightly different imaging protocol [20], [21]. This indicates that these features are appropriate, both for traditional supervised classifiers, and on slightly different previously unseen data, when using the methods proposed in this paper.

Compared to previous literature on plaque-component segmentation, our same-center segmentation and the proposed combination of transfer learning and adaptive histogram binning, have similar [18] or slightly better [19]–[21] accuracy than previously published results on same-center training and evaluation. Our results imply that such methods can more easily be implemented in multi-center studies, although standardization of image protocols remains advantageous. Our results also had a similar agreement with the ground truth manual annotations as the inter-observer agreement for both centers. This suggests that these segmentations could replace manual annotation in large research studies, for example to study the relation between composition and prognostic outcome parameters such as plaque progression and cerebrovascular events. In previous studies such relations have been found for presence of IPH and LRNC [5], [17] in MRI, and CA [6] in CT. Automatic segmentation would allow studying vulnerability based on component volumes, which may be more sensitive than presence versus absence. Moreover, use in clinical practice would be feasible with similar accuracy as manual annotation. Although this study was performed on the carotid artery, similar results can be expected when applied to MRI studies of atherosclerotic plaques in other vessels such as the aorta and femoral artery [43], [44]. MR imaging of the coronary vessel wall is still very challenging due to the small size and extensive cardiac and respiratory motions.

In conclusion, good plaque-component segmentations with similar agreement as inter-observer agreement were obtained for carotid MRI data from two centers. We showed that when no labeled same-center data is available extensive feature normalization by means of adaptive histogram binning improves results, and secondly that transfer-learning classifiers improve results when a few labeled same-center examples are available. These approaches yield results with similar accuracy to the reference of same-center training and significantly better than different-center training. The combination of feature normalization and transfer learning can facilitate segmentation across scanners. This can stimulate the wide implementation of automated image analysis methods in large-scale multi-center studies and in clinical practice.

ACKNOWLEDGMENT

The authors would like to thank Z. Kassar for his manual annotations, and F. Schreuder for his assistance in creating the consensus segmentation.

REFERENCES

[1] A. S. Go *et al.*, "Heart disease and stroke statistics-2013 update: A report from the American Heart Association," *Circulation*, vol. 127, no. 1, pp. e6–e245, 2013.

[2] G. W. Petty *et al.*, "Ischemic stroke subtypes: A population-based study of incidence and risk factors," *Stroke*, vol. 30, no. 12, pp. 2513–2516, 1999.

[3] M. M. Mughal *et al.*, "Symptomatic and asymptomatic carotid artery plaque," *Exp. Rev. Cardiovasc. Ther.*, vol. 9, no. 10, pp. 1315–1330, 2011.

[4] M. Naghavi *et al.*, "From vulnerable plaque to vulnerable patient: A call for new definitions and risk assessment strategies: Part I," *Circulation*, vol. 108, no. 14, pp. 1664–1672, 2003.

[5] T. Saam *et al.*, "Meta-analysis and systematic review of the predictive value of carotid plaque hemorrhage on cerebrovascular events by magnetic resonance imaging," *J. Am. Coll. Cardiol.*, vol. 62, no. 12, pp. 1081–1091, 2013.

[6] R. M. Kwee, "Systematic review on the association between calcification in carotid plaques and clinical ischemic symptoms," *J. Vasc. Surg.*, vol. 51, no. 4, pp. 1015–1025, 2010.

[7] T. G. Brott *et al.*, "2011 ASA/ACCF/AHA/AANN/AANS/ACR/ASNR/CNS/SAIP/SCAI/SIR/SNIS/SVM/SVS guideline on the management of patients with extracranial carotid and vertebral artery disease: Executive summary," *Stroke*, vol. 42, no. 8, pp. e420–e463, 2011.

[8] K. I. Paraskevas, D. P. Mikhailidis, and F. J. Veith, "Comparison of the five 2011 guidelines for the treatment of carotid stenosis," *J. Vasc. Surg.*, vol. 55, no. 5, pp. 1504–1508, 2012.

[9] T. J. Wang, "Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction," *Circulation*, vol. 123, no. 5, pp. 551–565, 2011.

[10] J. J. Ricotta *et al.*, "Updated society for vascular surgery guidelines for management of extracranial carotid disease," *J. Vasc. Surg.*, vol. 54, no. 3, pp. e1–e31, 2011.

[11] M. R. Makowski and R. M. Botnar, "MR imaging of the arterial vessel wall: Molecular imaging from bench to bedside," *Radiology*, vol. 269, no. 1, pp. 34–51, 2013.

[12] J. Sanz and Z. A. Fayad, "Imaging of atherosclerotic cardiovascular disease," *Nature*, vol. 451, no. 7181, pp. 953–957, 2008.

[13] T. S. Hatsukami, R. Ross, N. L. Polissar, and C. Yuan, "Visualization of fibrous cap thickness and rupture in human atherosclerotic carotid plaque in vivo with high-resolution magnetic resonance imaging," *Circulation*, vol. 102, no. 9, pp. 959–964, 2000.

[14] T. Saam *et al.*, "Quantitative evaluation of carotid plaque composition by in vivo MRI," *Arterioscler. Thromb. Vasc. Biol.*, vol. 25, no. 1, pp. 234–239, 2005.

[15] C. Yuan *et al.*, "In vivo accuracy of multispectral magnetic resonance imaging for identifying lipid-rich necrotic cores and intraplaque hemorrhage in advanced human carotid plaques," *Circulation*, vol. 104, no. 17, pp. 2051–2056, 2001.
[16] N. Takaya *et al.*, "Association between carotid plaque characteristics and subsequent ischemic cerebrovascular events: A prospective assessment with MRI—Initial results," *Stroke*, vol. 37, no. 3, pp. 818–823, 2006.
[17] A. Gupta *et al.*, "Carotid plaque MRI and stroke risk: A systematic review and meta-analysis," *Stroke*, vol. 44, no. 11, pp. 3071–3077, 2013.
[18] F. Liu *et al.*, "Automated in vivo segmentation of carotid plaque MRI with morphology-enhanced probability maps," *Magn. Reson. Med.*, vol. 55, no. 3, pp. 659–668, 2006.
[19] J. M. A. Hofman *et al.*, "Quantification of atherosclerotic plaque components using in vivo MRI and supervised classifiers," *Magn. Reson. Med.*, vol. 55, no. 4, pp. 790–799, 2006.
[20] R. van't Klooster *et al.*, "Automated versus manual in vivo segmentation of carotid plaque MRI," *Am. J. Neuroradiol.*, vol. 33, no. 8, pp. 1621–1627, 2012.
[21] A. van Engelen *et al.*, "Atherosclerotic plaque component segmentation in combined carotid MRI and CTA data incorporating class label uncertaintyy," *PLOS One*, 2014.
[22] B. Fischl *et al.*, "Sequence-independent segmentation of magnetic resonance images," *Neuroimage*, vol. 23, pp. S69–S84, 2004.
[23] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.
[24] Y. Zhuge and J. Udupa, "Intensity standardization simplifies brain MR image segmentation," *Comput. Vis. Image Understand.*, vol. 113, no. 10, pp. 1095–1103, 2009.
[25] Y. Artan, A. Oto, and I. Yetik, "Cross-device automated prostate cancer localization with multiparametric MRI," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5385–5394, Dec. 2013.
[26] R. H. H. M. Philipsen, P. Maduskar, L. Hogeweg, and B. van Ginneken, "Normalization of chest radiographs," in *Proc. SPIE Med. Imag.*, 2013, vol. 8670.
[27] L. G. Estrella, B. van Ginneken, and E. M. van Rikxoort, "Normalization of CT scans reconstructed with different kernels to reduce variability in emphysema measurements," in *Proc. SPIE Med. Imag.*, 2013, vol. 8670.
[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[29] P. Wu and T. Dietterich, "Improving SVM accuracy by training on auxiliary data sources," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 110.
[30] V. Ablavsky, C. Becker, and P. Fua, Transfer learning by sharing support vectors School Comput. Commun. Sci. Swiss Fed. Inst. Technol., Lausanne (EPFL), Tech. Rep., 2010.
[31] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, "Supervised image segmentation across scanner protocols: A transfer learning approach," in *MICCAI 2013 Workshop on Mach. Learn. Med. Imag.*, 2012, pp. 160–167.
[32] A. van Engelen *et al.*, "Multi-feature-based plaque characterization in ex vivo MRI trained by registration to 3D histology," *Phys. Med. Biol.*, vol. 57, no. 1, pp. 241–256, 2012, 1.
[33] N. Tustison *et al.*, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.
[34] R. P. W. Duin *et al.*, PRTools4.1, A MATLAB Toolbox for pattern recognition Delft Univ. Technol., 2007 [Online]. Available: http://www.prtools.org
[35] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2003, ch. 4.
[36] M. T. B. Truijman *et al.*, "PARISK (Plaque At RISK): Prospective multicenter study to improve diagnosis of high risk carotid plaques," *Int. J. Stroke*, vol. 9, pp. 747–754, 2013.
[37] W. Kerwin *et al.*, "Magnetic resonance imaging of carotid atherosclerosis: Plaque analysis," *Topics Magn. Reson. Imag.*, vol. 18, no. 5, pp. 371–378, 2007.
[38] V. C. Cappendijk *et al.*, "Comparison of single-sequence T1w TFE MRI with multisequence MRI for the quantification of lipid-rich necrotic core in atherosclerotic plaque," *J. Magn. Reson. Imag.* vol. 27, no. 6, pp. 1347–1355, 2008.
[39] J. Cai *et al.*, "In vivo quantitative measurement of intact fibrous cap and lipid-rich necrotic core size in atherosclerotic carotid plaque: Comparison of high-resolution, contrast-enhanced magnetic resonance imaging and histology," *Circulation*, vol. 112, no. 22, pp. 3437–3444, 2005.
[40] R. M. Kwee *et al.*, "Reproducibility of fibrous cap status assessment of carotid artery plaques by contrast-enhanced MRI," *Stroke*, vol. 40, no. 9, pp. 3017–3021, 2009.
[41] R. van't Klooster *et al.*, "Automated registration of multispectral MR vessel wall images of the carotid artery," *Med. Phys.*, vol. 40, no. 12, p. 121904, 2013.
[42] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, "A transfer-learning approach to image segmentation across scanners by maximizing distribution similarity," in *MICCAI 2013 Workshop on Mach. Learn. Med. Imag.*, 2013, pp. 49–56.
[43] M. Yukihiko and Z. A. Fayad, "Aortic plaque imaging and monitoring atherosclerotic plaque interventions," *Topics Magn. Reson. Imag.*, vol. 18, no. 25, pp. 349–355, 2007.
[44] M. S. Galizia *et al.*, "Wall morphology, blood flow and wall shear stress: MR findings in patients with peripheral artery disease," *Eur. Radiol.*, pp. 1–7, 2013.