A Unified Framework based on Convolutional Neural Networks for Interpreting Carotid Intima-Media Thickness Videos

by

Jaeyul Shin

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2016 by the
Graduate Supervisory Committee:

Jianming Liang, Chair
Ross Maciejewski
Baoxin Li

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of mortality yet largely preventable, but the key to prevention is to identify at-risk individuals before adverse events. For predicting individual CVD risk, carotid intima-media thickness (CIMT), a noninvasive ultrasound method, has proven to be valuable, offering several advantages over CT coronary artery calcium score. However, each CIMT examination includes several ultrasound videos, and interpreting each of these CIMT videos involves three operations: (1) select three end-diastolic ultrasound frames (EUF) in the video, (2) localize a region of interest (ROI) in each selected frame, and (3) trace the lumen-intima interface and the media-adventitia interface in each ROI to measure CIMT. These operations are tedious, laborious, and time consuming, a serious limitation that hinders the widespread utilization of CIMT in clinical practice. To overcome this limitation, this paper presents a new system to automate CIMT video interpretation. Our extensive experiments demonstrate that the suggested system significantly outperforms the state-of-the-art methods. The superior performance is attributable to our unified framework based on convolutional neural networks (CNNs) coupled with our informative image representation and effective post-processing of the CNN outputs, which are uniquely designed for each of the above three operations.

ACKNOWLEDGMENTS

I sincerely thank my advisor Dr. Jianming Liang for his continued support, guidance and encouragement during my masters and while writing this thesis. My sincere thanks to Dr. Ross Maciejewski, to provide me an opportunity to work under his guidance, who have encouraged and guided me throughout the process. His computer graphics and visualization course helped me tremendously in preparing lots of visual elements and gained important insights. I also would like to thank Dr. Baoxin Li for being on my thesis supervisory committee and for providing the guidance and the feedback on my work. Many thanks to Dr. Nima Tajbakhsh for his exceptional knowledge especially in medical imaing field. I am very grateful for the love and support from my wife, who gave a birth to my second child, lovely daughter in 2015.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Background

Carotid intima-media thickness (CIMT) is considered as an early and reliable non-invasive indicator of cardiovascular risk (Stein *et al.*, 2008) and this thesis aims to introduce new methodology that significantly improves CIMT performance using convolutional neural networks (CNNs).

1.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are neural networks with multiple layers that can learn complex hierarchy from the images that recently attracted lots of researches along with advancement with the parallel computing power of graphics processing unit (GPU). CNNs were originally proposed by LeCun in 1989, however, due to slow computing power at that time, CNNs did not gain popularity. However, with the powerful GPUs in scientific computing together with effective regularization techniques (Goodfellow *et al.*, 2013; Hinton *et al.*, 2012; Krizhevsky *et al.*, 2012; Wan *et al.*, 2013; LeCun *et al.*, 2015) has re-ignited research and applications that break records in many computer vision and image analysis tasks. The major power of CNNs is that they learn hierarchical set of image features and that it learns features automatically. This eliminates the need for hand-crafted features which often is suboptimal. Recent image classification challenges such as ImageNet have shown that deep learning machines are now outperforming human and therefore it was a natural choice to use CNNs for medical image analysis, especially for CIMT challenge in this thesis.

CNNs are an extension of multi-layer percentptrons (MLPS) where multiple convolutional layers are placed as hidden layers. Each convolutional layer is connected to a small subset of spatially connected units in the previous layer. The weights in the layer are shared between all the units in convolutional layer and weights are updated during training phase. Weight sharing dramatically reduces the network's width which enable deeper architectures. The pooling layers usually follow the convolutional layers which produce a single output from a neighborhood of units by either taking the average, the maximum or some combinational with learnable weights which mimicks the primary visual cortex (Hubel and Wiesel, 1959). The layers toward the end are made of series of consecutive 1x1 convolutional layer which are also known as fully connected layers. Finally, a softmax or a regression layer generates the outputs.

Convolutional neural networks are trained using the back-progating algorithm like a multi-layer perceptron. If D denotes a set of training images, W denotes a matrix containing the weights of the convolutional layers, and $f_W(D^{(i)})$ denotes the loss for the i^{th} training image, the loss over the entire training set is then computed as

$$\mathcal{L}(W) = \frac{1}{|D|} \sum_{i}^{|D|} f_W(X^{(i)}) \tag{1.1}$$

To minimize the loss function with respect to the unknown weights W, the popular choice is gradient descent, however, due to limited GPU memory, evaluating loss function based on entire training set D is not feasible. Therefore, the loss function is approximated with loss over the mini-batches of training dataset size $N << |D|$. Given the size of mini-batches, ranging from 128 to 1024, one can approximate the loss function as $\mathcal{L}(W) \approx \frac{1}{N} \sum_{i=1}^{N} f_W(X^{(i)})$, and iteratively update the weights of the network with the following equations:

$$\gamma_t = \gamma^{* \frac{tN}{|D|}}$$

$$V_{t+1} = \mu V_t - \gamma_t \alpha \Delta L(W_t)$$

$$W_{t+1} = W_t + V_{t+1} \tag{1.2}$$

where α is the learning rate, μ is the momentum that indicates the contribution of the previous weight update in the current iteration, and γ is the scheduling rate that decreases learning rate α at the end of each epoch.

1.3 Carotid Intima-Media

CIMT is the distance between lumen-intima interface (LI) and the media-adventitia interface (MAI) (Figure 1.1). The CIMT is defined as the distance between the lumen-intima and media-adventitia interfaces at the far wall of the carotid artery (Figure 1.2). Therefore, the lumen-intima and the media-adventitia interfaces must be identified accurately to measure CIMT.

Previous work include hand crafted algorithms as well as machine learning based methods for CIMT image interpretation. Some of the earlier approaches focused on intensity profile analysis and distribution, gradient computation (Pignoli and Longo, 1987; Touboul *et al.*, 1992; Faita *et al.*, 2008), or use of various edge properties through dynamic programming (Liang *et al.*, 2000; Cheng and Jiang, 2008; Rossi *et al.*, 2010). Recent approaches (Loizou *et al.*, 2007; Delsanto *et al.*, 2007; Petroudi *et al.*, 2012; Xu *et al.*, 2012; Ilea *et al.*, 2013; Bastida-Jumilla *et al.*, 2013) are mostly based on active contours (aka, snakes) or their variations (Kass *et al.*, 1988). Some methods require user interaction to adjust the position of the snake control points while other approaches tried to achieve complete automation using special image processing algorithms, such as Hough transform (Molinari *et al.*, 2012) and dynamic programming (Rossi *et al.*, 2010). More recently, Menchn-Lara et al. employed a committee of standard multilayer perceptrons in (Menchón-Lara *et al.*, 2013) and a single standard multilayer perceptron with an auto-encoder in (Menchón-Lara

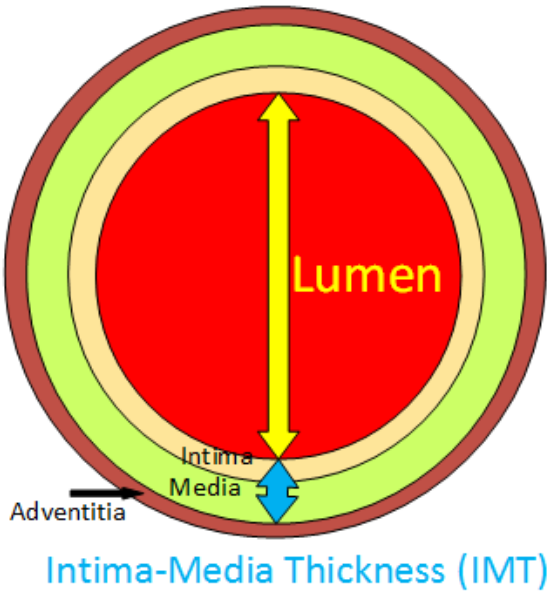

Figure 1.1: Intima-Media Thickness Diagram

and Sancho-Gómez, 2015) for CIMT image interpretation, but both methods did not outperform the snake-based methods from the same research group (Bastida-Jumilla *et al.*, 2013, 2015). More complete survey of various methods for automatic CIMT measurements are found in the review studies conducted by Molinari et al.(Molinari *et al.*, 2010) and (Loizou, 2014).

However, most of above methods are focused on the final operation which is CIMT measurement, which ignores two preceding operations; correct frame selection and ROI localization. To my knowledge, the prior work by Sharma *et al.* (2014), an extension of work by Zhu *et al.* (2011), automatically selects the EUF frame, localizes the ROI in each selected EUF frame, and provides the CIMT measurement in the selected ROI. However, as with other works, this method is also based on hand-crafted algorithms, which often lack the desired robustness for routine clinical use, a weakness that our new proposed method aims to overcome through use of CNNs. As demonstrated in Sections 4 and 5, this new system outperforms the existing methods in all aspects including frame selection, ROI

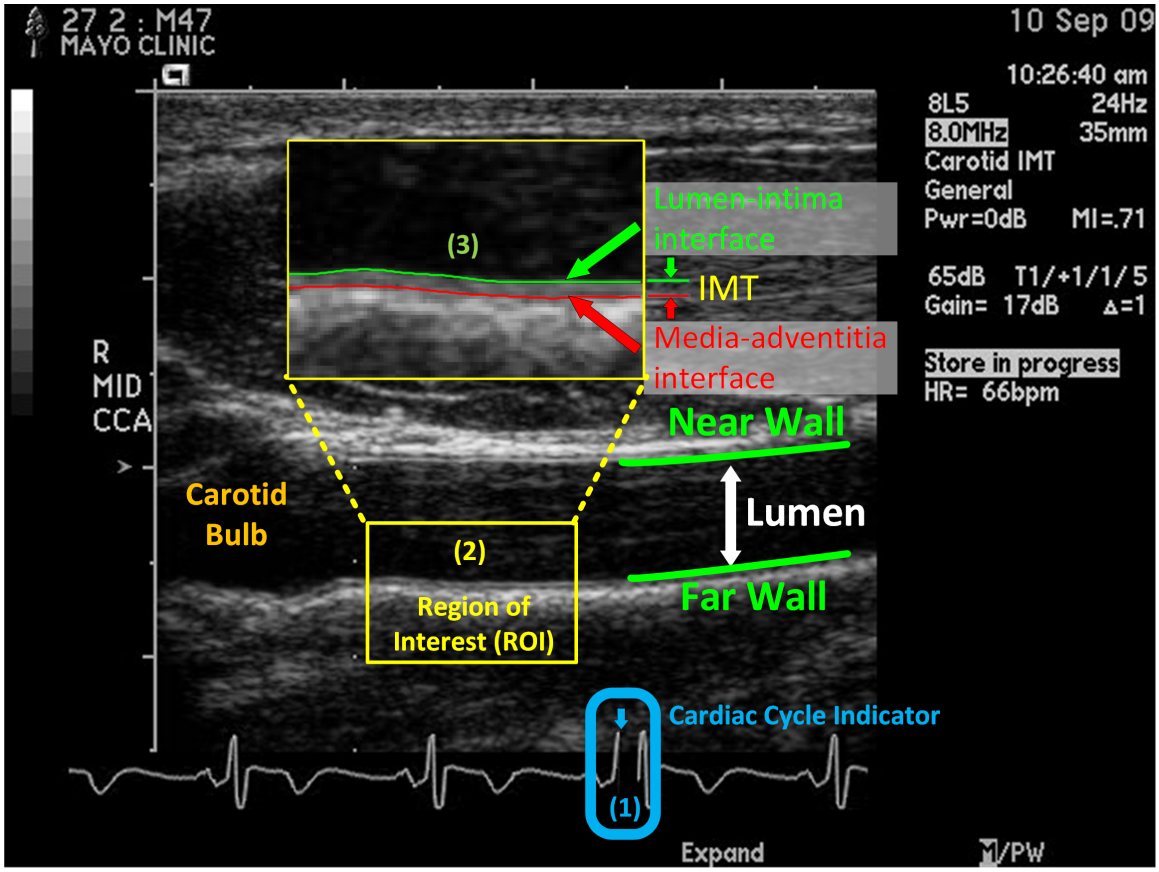

Figure 1.2: Longitudinal view of the carotid artery in an ultrasound B-scan image. CIMT is defined as the distance between the lumen-intima interface and the media-adventitia interface, measured approximately $1\,\mathrm{cm}$ distal from the carotid bulb on the far wall of the common carotid artery at the end of the diastole; therefore, interpreting a CIMT video involves three operations: (1) select three end-diastolic ultrasound frames (EUFs) in each video (the cardiac cycle indicator, a black line, shows to where in the cardiac cycle the current frame corresponds); (2) localize a region of interest (ROI) approximately $1\,\mathrm{cm}$ distal from the carotid bulb in the selected EUF; (3) measure the CIMT within the localized ROI. This thesis aims to automate these three operations simultaneously through a unified framework based on convolutional neural networks.

localization, and CIMT measurements.

A key contribution of this thesis is to automate CIMT by combining all three operations in unified framework using convolutional neural networks (CNNs). By combining pre-processing and post-processing, this new proposed CNN-based method significantly outperforms all the existing methods which includes frame selection, ROI localization and CIMT measurements.

First, the proposed frame selection method uses ECG signals at the bottom of ultrasound frames. Then, the pre-processing of patches and post processing of CNN outputs enabled significant increase in the performance of the frame selection compared to the previous hand-crafted approach (Sharma *et al.*, 2014). Second, a novel method for localizing the ROI for CIMT interpretation is proposed. This method introduces the discriminative power of a CNN with a contextual constraint to accurately localize the ROIs and with the contextual constraint which is also found by CNNs. Third, a framework is proposed that combines active contour models and CNNs for sub-pixel accuracy boundary segmentation. Given a frame and an ROI, two open snakes are initialized with CNNs output which further deforms to the actual intima-media boundary. Lastly, thorough evaluation of leave-one-patient-out cross-validation[1] using the training data only to adjust the parameters of the system, and then thoroughly evaluated performance using a large number of test data set of CIMT videos. This cross validation provides robust confirmation of experimental results.

[1]Leaving all the videos from one patient out for validation, 12-fold cross validation.

Chapter 2

CIMT PROTOCOL

The CIMT exams utilized in this thesis were performed with B-Mode ultrasound using an 8-14MHz linear array transducer utilizing fundamental frequency only (Acuson Sequoia^{TM}, Mountain View, CA, USA) (Hurst *et al.*, 2010). The carotid screening protocol begins with scanning bilateral carotid arteries in a transverse manner from the proximal aspect to the proximal internal and external carotid arteries. The probe is then turned to obtain the longitudinal view of the distal common carotid artery. The sonographer optimizes the 2D images of the lumen-intima and media-adventitia interfaces at the level of the common carotid artery by adjusting overall gain, time gain, compensation and focus position. Once the parameters are optimized, the sonographer captures two CIMT videos focused on the common carotid artery from two optimal angles of incidence. The same procedure is repeated for the other side of neck, resulting in a total of 4 CIMT videos for each subject.

Chapter 3

METHOD

The goal is to automate the three operations in CIMT video interpretation. First, given a CIMT video, the system automatically identifies three EUFs (Section 3.1), localizes an ROI in each EUF (Section 3.2), and segments the lumen-intima and media-adventitia interfaces within each ROI (Section 3.3). Figure 3.1 shows an schematic overview of the new proposed system.

3.1 Frame Selection

First, EUFs are selected based on the ECG signal embedded at the bottom part of a CIMT video. The cardiac cycle indicator is represented by a moving-to-the-right black line in each frame. Since the ECG signal is overlaid on the ultrasound image, there is quite bit of noise around the indicator. The challenge is to reconstruct the original ECG signal from

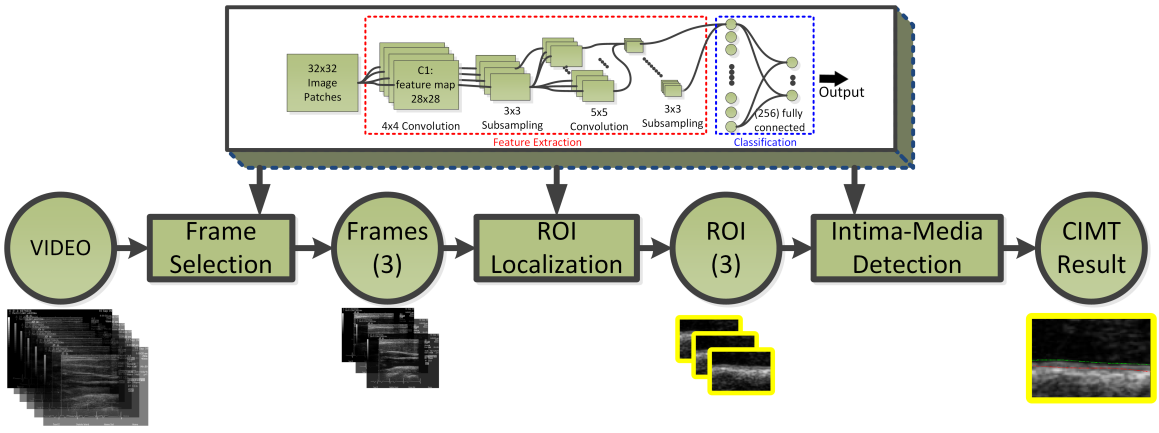

Figure 3.1: Video frames are passed to the system, and three frames are selected by CNNs, and passed to ROI localization which in turn passed to intima-media boundary segmentation system to give final thickness result.

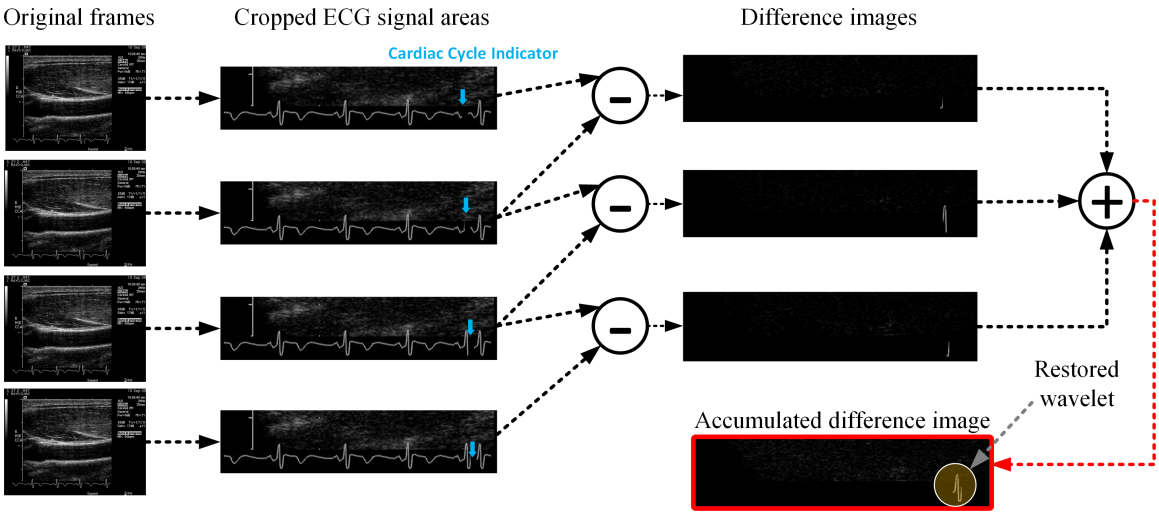

Figure 3.2: An accumulated difference image is generated by adding up three neighboring difference images.

noisy frames and to detect the R peaks from the ECG signal, as the R-peaks correspond to the EUFs. To do so, rather than using a single frame to train or test, the concept of accumulated difference images is proposed. This method combines multiple difference frames, more specifically four frames that carry sufficient information for CNN to learn and distinguish R-peaks from non-R-peaks. Figure 3.3 shows when the difference frames are combined into a single frame that demonstrates the high noise level in many cases that make challenging to extract clean signals.

Training Phase: Let I^t denote an image sub-region selected from the lower part of an ultrasound frame so that it contains the ECG signal. First, construct a set of difference images d^t by subtracting every consecutive pairs of images, $d^t = |I^t - I^{t+1}|$, and then form accumulated difference images by adding up every three neighboring difference images, $D^t = \sum_{i=0}^{2} d^{t-i}$. Accumulated difference image D^t can capture the cardiac cycle indicator at frame t. Figure 3.2 illustrates how an accumulated difference image is generated.

Next, the location of the restored wavelet is determined in each accumulated difference image. For this purpose, the weighted centroid $c = [c_x, c_y]$ of each accumulated difference

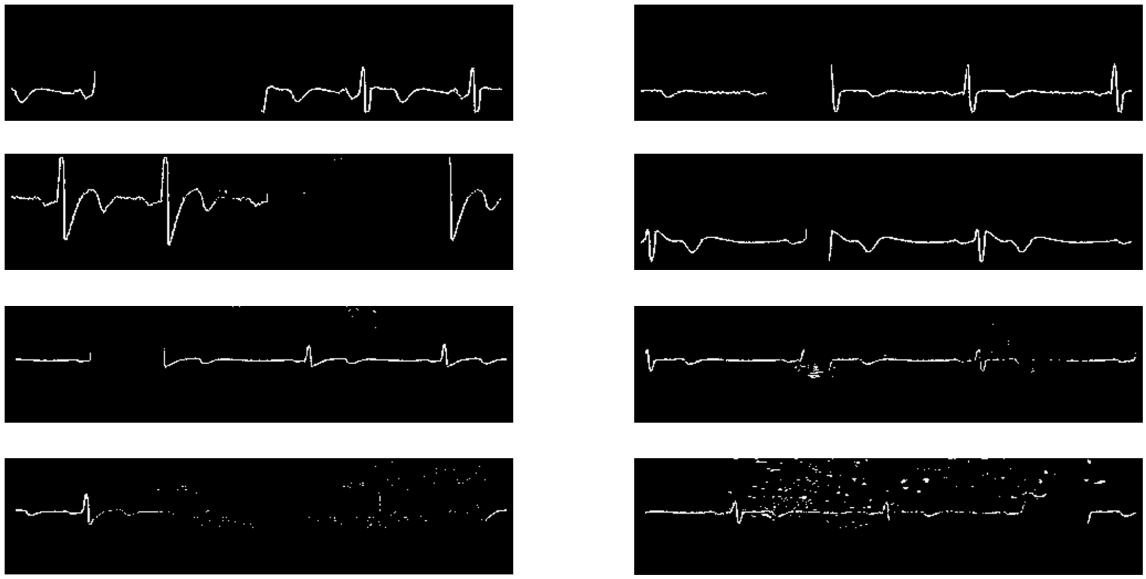

Figure 3.3: 8 cases where the difference images combined into a single image to show clean images vs high noise images.

image D^t is as follows:

$$c = \frac{1}{Z_t} \sum_{p \in D^t} D^t(p_x, p_y) \times p$$

where $p = [p_x, p_y]$ is a pixel in the accumulated difference image and $Z_t = \sum_{p \in D^t} D^t(p_x, p_y)$ is a normalization factor that ensures the weighted centroid stays within the image boundary. Once centroids are identified, patches of size 32×32 are extracted around the centroid locations. Specifically, patches with up to 2 pixel translations from each centroid are extracted. However, the patches are not scaled in data augmentation, because doing so would inject label noise in the training set. For instance, a small restored wavelet may take the appearance of an R-peak after expanding or an R-peak wavelet may look like a non-R-peak wavelet after shrinking. Nor do we perform rotation-based patch augmentation, because the restored wavelets do not appear with rotation in the test image patches. Once collected, patches are binarized using Otsu's method. In Section 4, the choice of binarization method

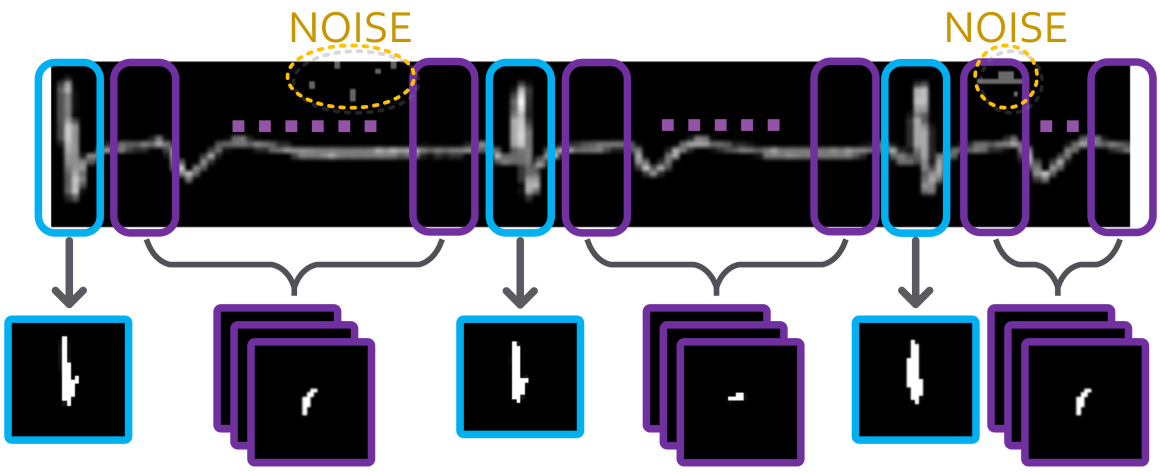

Figure 3.4: The patch extraction for training stage of automatic frame selection scheme (blue is labeled as positives and purple as negatives).

through an extensive set of experiments is presented. Each binary patch is then labeled as positive if it corresponds to an EUF (i.e., an R-peak); otherwise negative. Basically, given a patch, the accumulated difference image is determined from which the patch is extracted. Then it is possible to trace back to the underlying difference images and check whether they are related to the EUF or not. Once the patches are labeled as in Figure 3.4, a stratified set is formed with 96,000 patches to train a 2-way CNN for frame selection.

Testing Phase: Figure 3.5 shows our frame selection system given a test video. First, an accumulated difference image for each frame in the video is computed. Then, image patches are extracted from the weighted centroids of the accumulated difference images. The probability of each frame being the EUF is measured as the average probabilities assigned by the CNN to the corresponding patches. By concatenating the resulting probabilities for all the frames in the video, we obtain a probability signal whose local maxima indicate the locations of the EUFs. However, the generated probability signals often exhibit abrupt changes, which can cause too many local maxima along the signal. We therefore first smooth the probability signal using a Gaussian function, and then find the EUFs by

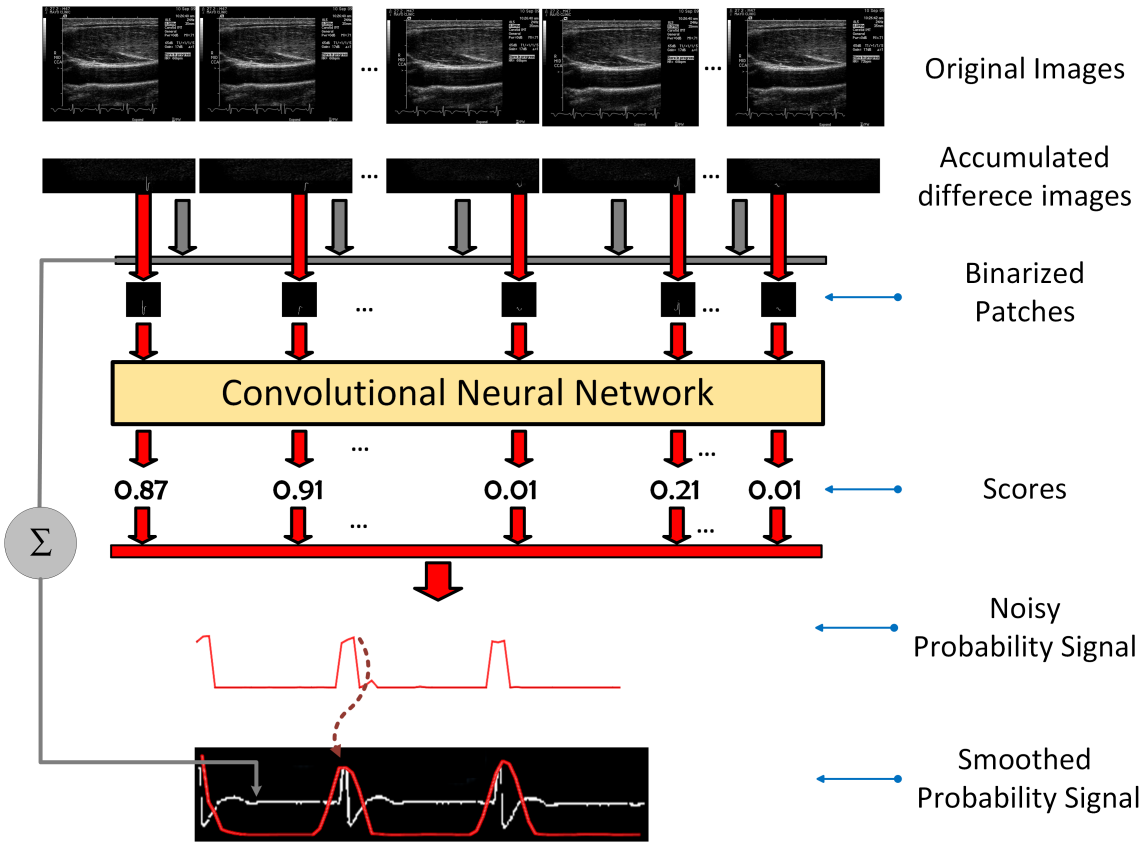

Figure 3.5: The test stage of our automatic frame selection scheme.

locating the local maxima of the smoothed signals. In Figure 3.5, for illustration purposes, the reconstructed ECG signal is shown which is computed as the average of the accumulated difference images, $\frac{1}{N}\sum_{t=1}^{N} D^t$ with N being the number of frames in the video. As seen, the probability of being the EUF reaches its maximum around the R peaks of the QRS complexes (as desired) and then smoothly decays as it distances from the R peaks. By mapping the locations of the local maxima to the frame numbers, the EUFs can be identified in the test video.

3.2 ROI Localization

Accurate localization of the ROI is challenging, because, as seen in Figure 1.2, there are no significant differences that can be observed in image appearance among the ROIs on the far wall of the carotid artery. To overcome this challenge, the location of the carotid bulb as a contextual constraint is utilized. This constraint is chosen for two reasons: 1) the carotid bulb appears as a distinct dark area in the ultrasound frame and thus can be uniquely identified; 2) according to the consensus statement of American society of Electrocardiography for cardiovascular risk assessment, the ROI should be placed approximately $1\,\mathrm{cm}$ from the carotid bulb on the far wall of the common carotid artery. While the former motivates the use of the carotid bulb location as a constraint from a technical point of view, the latter justifies this constraint from a clinical standpoint.

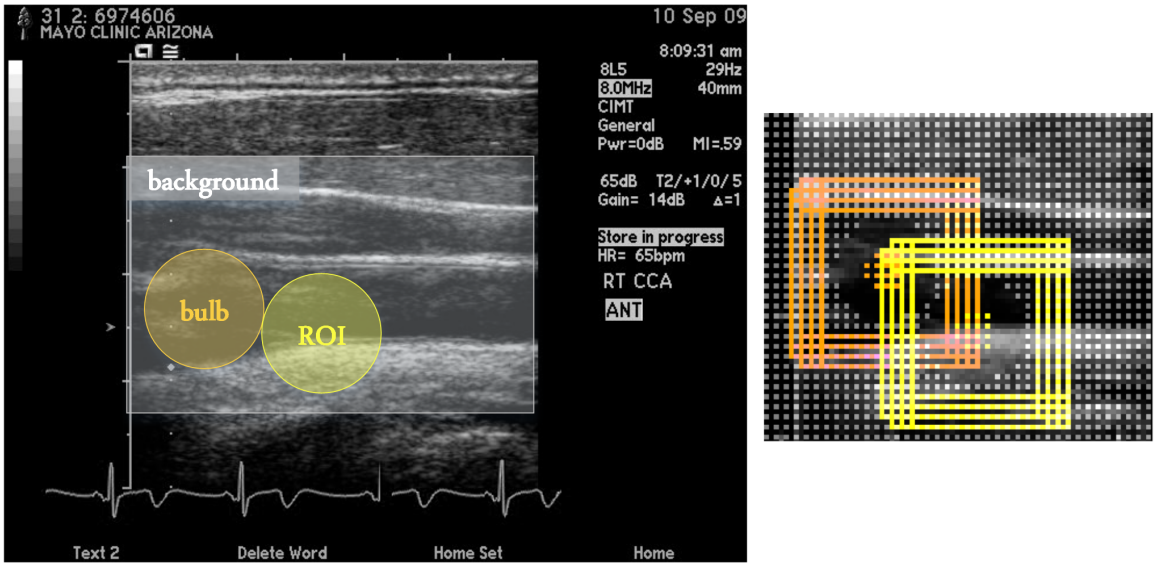

Figure 3.6: For constrained ROI localization, we use a 3-way CNN whose training image patches are extracted from a grid of points on the background and around the ROI and the carotid bulb locations.

Training Phase: We incorporate this constraint in the suggested system by training a 3-

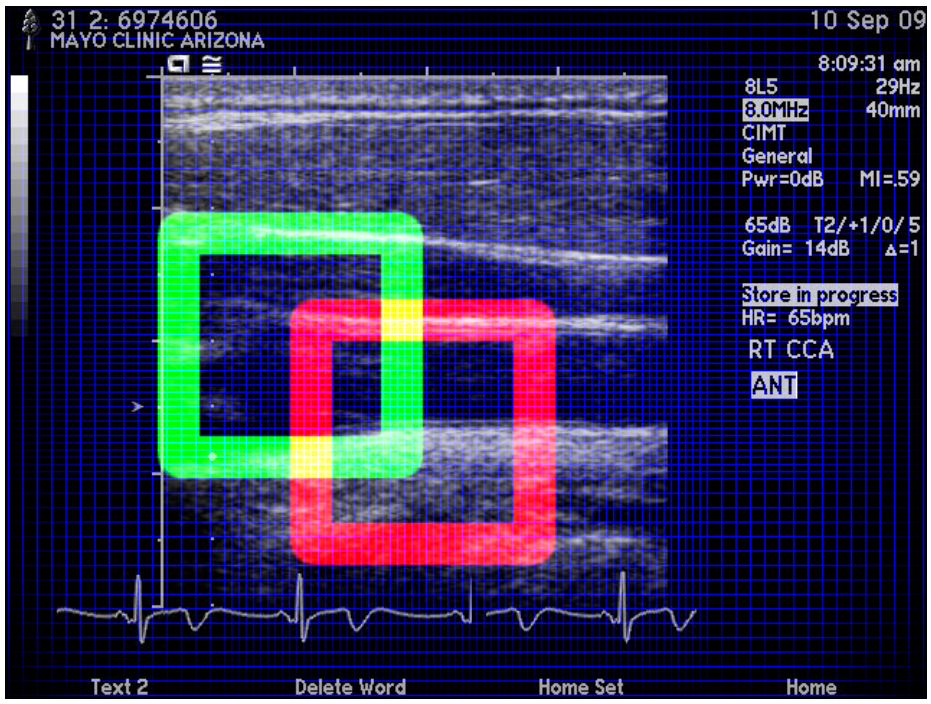

Figure 3.7: Patch extraction boundary rectangle. Green rectangle represents bulb and red for the ROI. Blue represents background or negative patches

way CNNs that simultaneously localizes both ROI and carotid bulb, and then refines the estimated location of the ROI given the location of the carotid bulb (Figure 3.6) illustrates how the image patches are extracted from a training frame. Figure 3.7 shows patch extraction boundary for all 3 classes for easier visualization. We perform data augmentation by extracting the training patches within a circle around the locations of the carotid bulbs and the ROIs. The negative patches are extracted from a grid of points sufficiently far from the locations of the carotid bulbs and the ROIs. Note that the above translation-based data augmentation is sufficient for this application, because our database provides a relatively large number of training EUFs, from which a large set of training patches can be collected. Once the patches are collected, we form a stratified training set with approximately 410,000 patches to train a 3-way CNN for constrained ROI localization.

Testing Phase: Referring to Figure 3.8, during the test stage, the trained CNN is applied

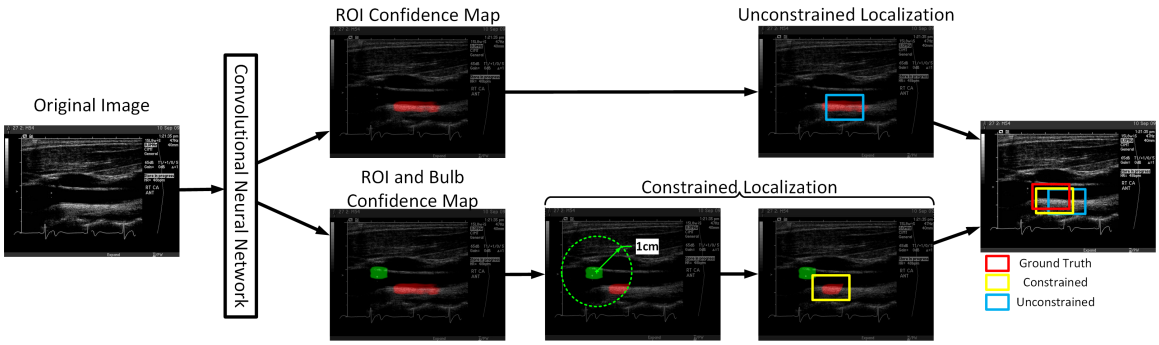

Figure 3.8: The test stage of our ROI localization method. In the unconstrained scenario, we only use the ROI confidence map, which results in relatively large localization error. In the constrained mode, given the estimated location of the carotid bulb, we localize the ROI more accurately.

to all the pixels in the EUF, generating two confidence maps with the same size as the EUF. The first confidence map shows the probability of a pixel being the carotid bulb and the second confidence map shows the probability of a pixel being the ROI. One way to localize the ROI is to find the center of the largest connected component within the ROI confidence map without considering the detected location of the carotid bulb. However, this naive approach may fail to accurately localize the ROI. For instance, a long-tale connected component along the far wall of the carotid artery may cause substantial ROI localization error as seen in Figure 3.9. To compound the problem, the largest connected component of the ROI confidence map may appear far from the actual location of the ROI, resulting in a complete detection failure. To overcome these limitations, we constraint the ROI location l_{roi} by the location of the carotid bulb l_{cb}. For this purpose, we first determine the location of the carotid bulb as the centroid of the largest connected component within the first confidence map. ROI localization can be obtained using the following formula,

$$l_{roi} = \frac{\sum_{p \in C^*} M(p) \cdot p \cdot I(p)}{\sum_{p \in C^*} M(p) \cdot I(p)} \tag{3.1}$$

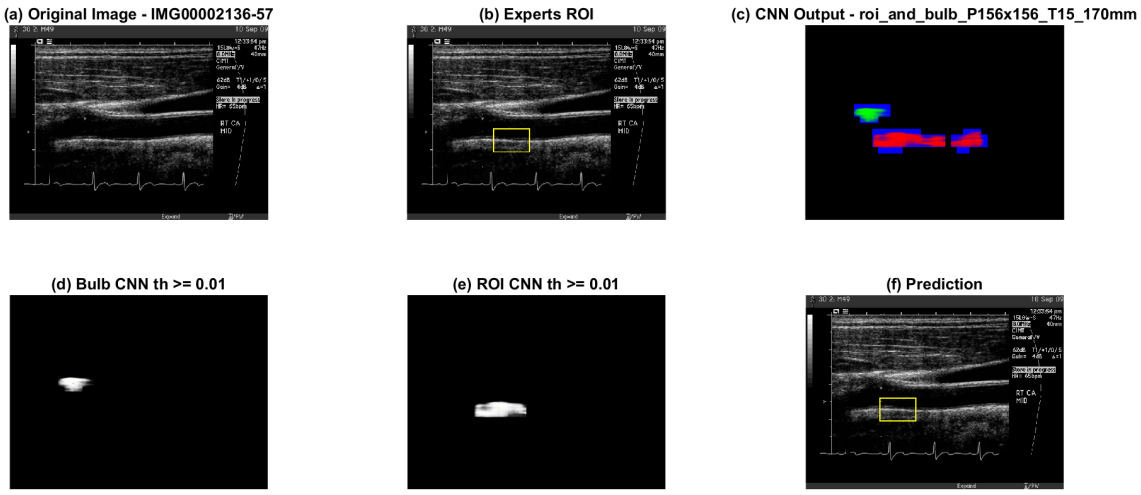

Figure 3.9: The step by step test case for ROI localization process. (a) Original image. (b) ground truth by expert. (c) 3-way CNN confidence map. (d) carotid bulb confidence map. (e) ROI confidence map. (f) final ROI location is determined.

where l_{roi} denotes the ROI location, l_{cb} denotes the center of the carotid bulb, M denotes the confidence map of being the ROI, C^* is the largest connected component in M that is the nearest to the carotid bulb, and I(p) is an indicator function for pixel $p = [p_x, p_y]$ that is defined as

$$I(p) = \begin{cases} 1, & \text{if } \|p - l_{cb}\| < 1\,\text{cm} \quad (3.2) \\ 0, & \text{otherwise} \quad (3.3) \end{cases}$$

The indicator function I(p) is binary function that simply includes pixel when the value is 1 as in Eq. 3.2, otherwise excludes pixel when the value is 0 as in Eq. 3.3. In simple term, this function excludes confidence map of being the ROI that is farther than 1 cm from the center of carotid bulb location.

3.3 Intima-Media Thickness Measurement

Measuring intima-media thickness require a continuous and one-pixel precise boundary for lumen-intima and media-adventitia. Lumen-intima is relatively easier to detect because of strong gradient change at the border (large dark region above lumen-intima interface), however, detecting media-adventitia interface is quite challenging due to its subtle image gradients and noise around its border. The proposed 3-way classification used is: 1) lumen-intima interface, 2) media-adventitia interface, and 3) background.

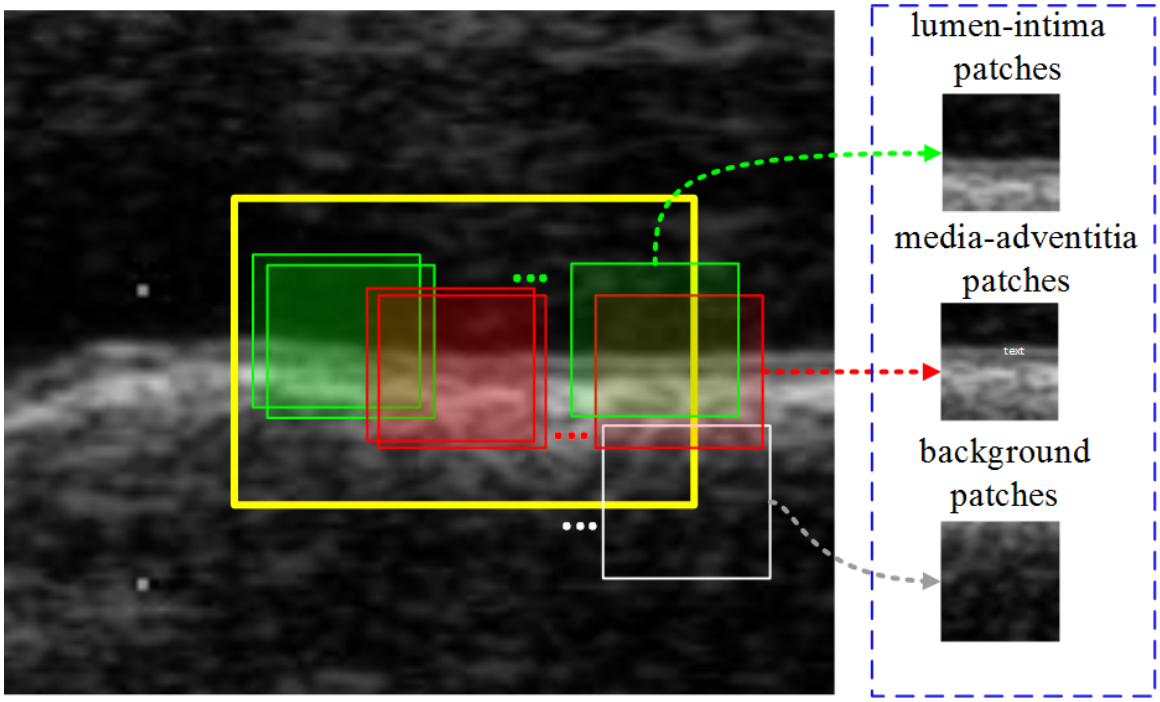

Figure 3.10: For lumen-intima and media-adventitia interface segmentation, we use a 3-way CNN whose training image patches are extracted from the background and around the lumen-intima and media-adventitia interfaces.

Training Phase: To train 3-way CNN, collecting pixel-by-pixel patches were inefficient and un-necessary. Instead, sparse background patches and then pixel-by-pixel image patches around lumen-intima interface and media-adventitia interface with additional patches ± 3

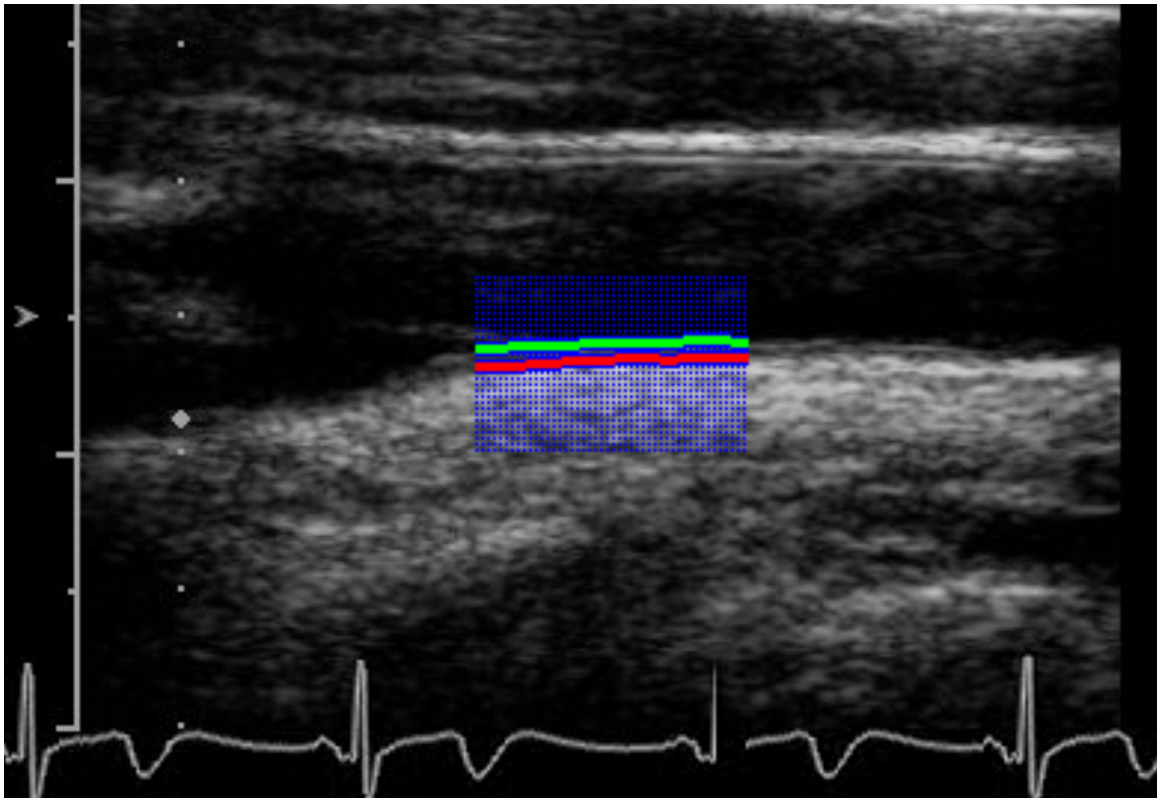

Figure 3.11: The patch extraction for training stage of IMT. Each color dot represents the center of 32x32 patches being extracted. Green represents the lumen, red for intima interface and blue for background patches.

pixels from the ground truth provided showed very similar performance in experimental results. Using ± 3 pixels for additional patches around intima-media boundary was necessary to balance number of patches with background patches and so that the confidence map could produce thicker scores along the two interfaces which could then used to select the center since there are odd number (exactly three in this case) per interface. In the images given, lumen-intima interface and media-adventitia interface had about six to eight pixels apart so going more than ± 3 pixels for patches may cause overlapping and poorer performance. Figure 3.10 illustrates how the training patches are collected from an ROI.

Testing Phase: Figure 3.12 illustrates the testing process. The 3-way trained CNN is

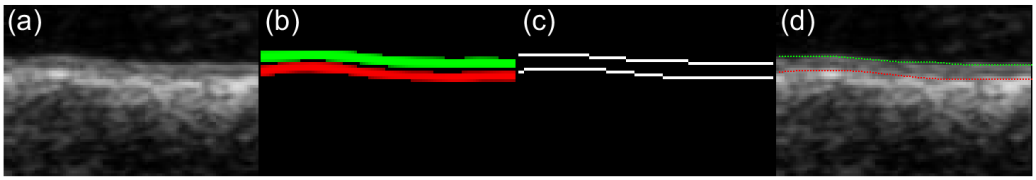

Figure 3.12: The test stage of lumen-intima and media-adventitia interface detection. (a) a test ROI. (b) The trained CNN generates a confidence map where the green and red colors indicate the likelihood of lumen-intima interface and media-adventitia interface, respectively. (c) The thick probability band around each interface is thinned by selecting the largest probability for each interface in each column. (d) The step-like boundaries are refined through two open snakes.

applied in a sliding-window fashion for a given test ROI and generates two confidence maps (Figure 3.12(b)) with the same size as the ROI. Since confidence map is thicker than a pixel, we choose the maximum response column-by-column and generate a new binary image as shown in Figure 3.12(c). Finally, we use two active contour models (a.k.a, snakes) (Liang *et al.*, 2006) for segmenting lumen-intima and media-adventitia interfaces. The input image to the snakes is the binary image only, not the original image. Figure 3.12(d) shows two final converged snakes and measurements are taken as the average vertical distance between the two snakes.

For illustration purpose, one of the best test case is shown in Figure 3.13 as well as worst test case in Figure 3.14 which represent step-by-step processes how the proposed system work for a given ROI. In Figure 3.14, even with the missing pixel boundary information, sub-optimal CNN prediction output is nicely augmented by use of snakes and CNN prediction output helps with the initial snake position.

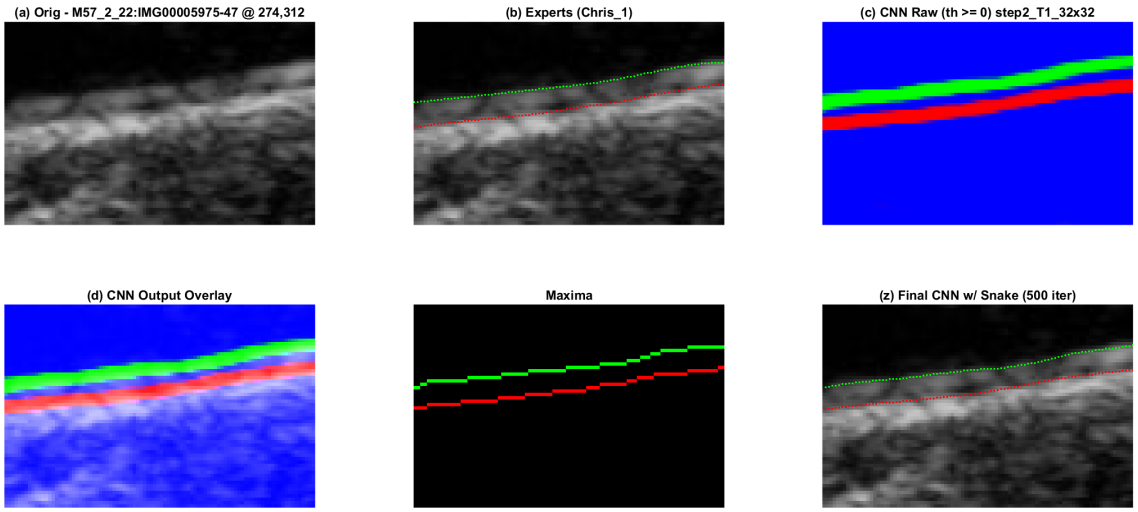

Figure 3.13: One of the best test case of lumen-intima and media-adventitia interface detection. (a) Original image. (b) Expert's ground truth. (c) CNN raw predictions output. (d) CNN raw outputs overlaid with original image. (e) Maxima for each column for initial snake position. (d) Snake after 500 iterations.

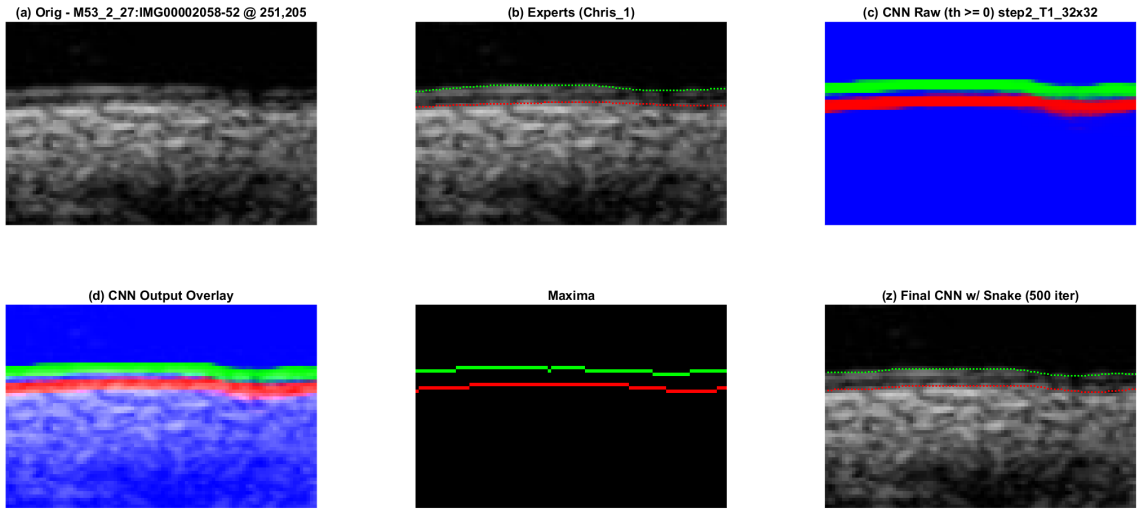

Figure 3.14: One of the "worst" test case of lumen-intima and media-adventitia interface detection. (a) Original image. (b) Expert's ground truth. (c) CNN raw predictions output. (d) CNN raw outputs overlaid with original image. (e) Maxima for each column for initial snake position. (d) Snake after 500 iterations.

Chapter 4

EXPERIMENTS

A database of 92 CIMT videos captured from 23 subjects with 2 CIMT videos from the left and 2 CIMT videos from the right carotid artery of each subject are used in this experiments. The ground truth for each video contains the EUF number, the locations of ROI, and the segmentation of lumen-intima and media-adventitia interfaces. For consistency, we use the same training set and the same test set (no overlap with training) for all three tasks. The training set contains 48 CIMT videos of 12 subjects with a total of 4,456 frames and our test set contains 44 CIMT videos of 11 subjects with a total of 3,565 frames. For each task, leave-one-patient-out cross-validation (12-fold cross validation) is performed based on the *training subjects* to tune the parameters, and then the performance of the tuned system using the test subjects is evaluated.

Architecture: As shown in Table 4.1, a CNN architecture with 2 convolutional layers, 2 sub-sampling layers, and 2 fully connected layers (see Section 5 for our justifications) is used. In addition, a softmax layer to the last fully connected layer is appended so as to generate probabilistic confidence values for each class. This CNN architecture has input patches of size 32x32, and the collected patches are re-sized to 32x32 prior to the training process. For the CNNs used in our experiments, an initial learning rate of $\alpha = 0.001$, a momentum of $\mu = 0.9$, and a constant scheduling rate of $\gamma = 0.95$ (the rate of learning rate decrease at each epoch) is used.

Pre- and post-processing for frame selection: Interestingly, use of binarized image patches for training CNNs improved the quality of convergence and accuracy of frame selection which was found out experimentally. Furthermore, the parameter of standard deviation of

Table 4.1: The CNN architecture used in the experiments. Note that C is the number of classes, which is 2 for frame selection and 3 for both ROI localization and intima-media thickness measurements.

layer	type	input	kernel	stride	pad	output
0	input	32x32	N/A	N/A	N/A	32x32
1	convolution	32x32	5x5	1	0	64x28x28
2	max pooling	64x28x28	3x3	2	0	64x14x14
1	convolution	64x14x14	5x5	1	0	64x10x10
2	max pooling	64x10x10	3x3	2	0	64x5x5
2	fully connected	64x5x5	5x5	1	0	250x1
2	fully connected	250x1	1x1	1	0	Cx1

the Gaussian function used for smoothing the probability signals, can also substantially influence frame selection accuracy. Therefore, leave-one-patient-out cross-validation was conducted based on the training subjects to find the best binarization method and the optimal standard deviation of the Gaussian function. For binarization, a fixed set of thresholds and adaptive thresholding using Otsu's method(Otsu, 1975) was used. For smoothing the output scores, a Gaussian function with different standard deviation (σ_g) as well as the scenario where no smoothing is applied were performed. For each configuration of parameters, a free-response ROC (FROC) analysis was done. For labeling, if the frame is found within one frame from the expert-annotated EUF, then that frame is considered as a true positive, otherwise, a false positive.

The leave-one-patient-out cross-validation study, summarized in Figure 4.2, indicates that the use of a Gaussian function with $\sigma_g = 1.5$ for smoothing the probability signals and

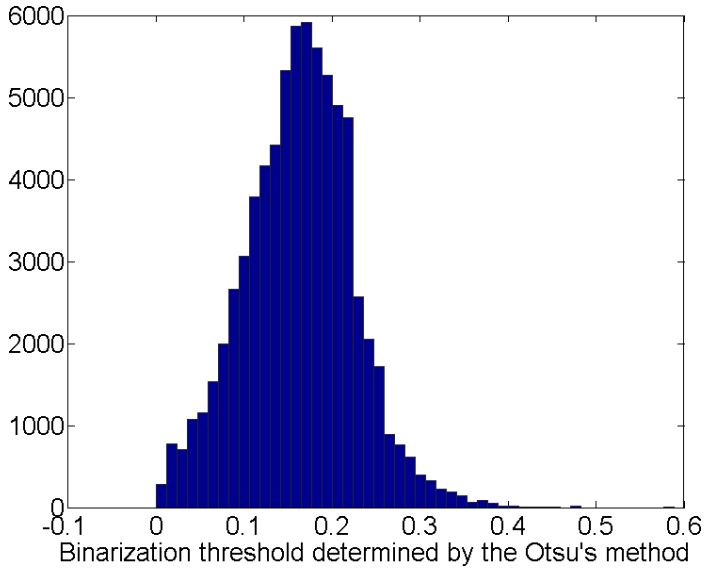

Figure 4.1: Histogram of Otsu's binarization thresholds in frame patches. Y-axis is frequency and x-axis is threshold ranging from 0 to 1.

adaptive thresholding using Otsu's method achieve the highest performance. For completeness, Figure 4.1 shows histogram of Otsu's binarization thresholds for all the extracted patches. Figure 4.3 shows the FROC curve of our system for the test subjects using the above parameters. For comparison, the operating point of the hand-crafted approach (Sharma *et al.*, 2014) is shown, which is significantly outperformed by the suggested system.

Constrained ROI Localization: A leave-one-patient-out cross-validation study was conducted based on the training subjects to find the optimal size of the training patches. The size of patch is important because the 32x32 pixel patch, equivalent to 0.35×0.35 cm patch only captures small part of the overall image and therefore does not have enough contextual information. Figure 4.4 shows the various sizes of patches that were used as well as ROI size for comparison. The cross-validation analysis, summarized in Figure 4.5, indicates that the use of 1.8×1.8 cm patches achieves the most stable performance, yielding low ROI localization error with only a few outliers. Figure 4.6 shows the ROI localization error

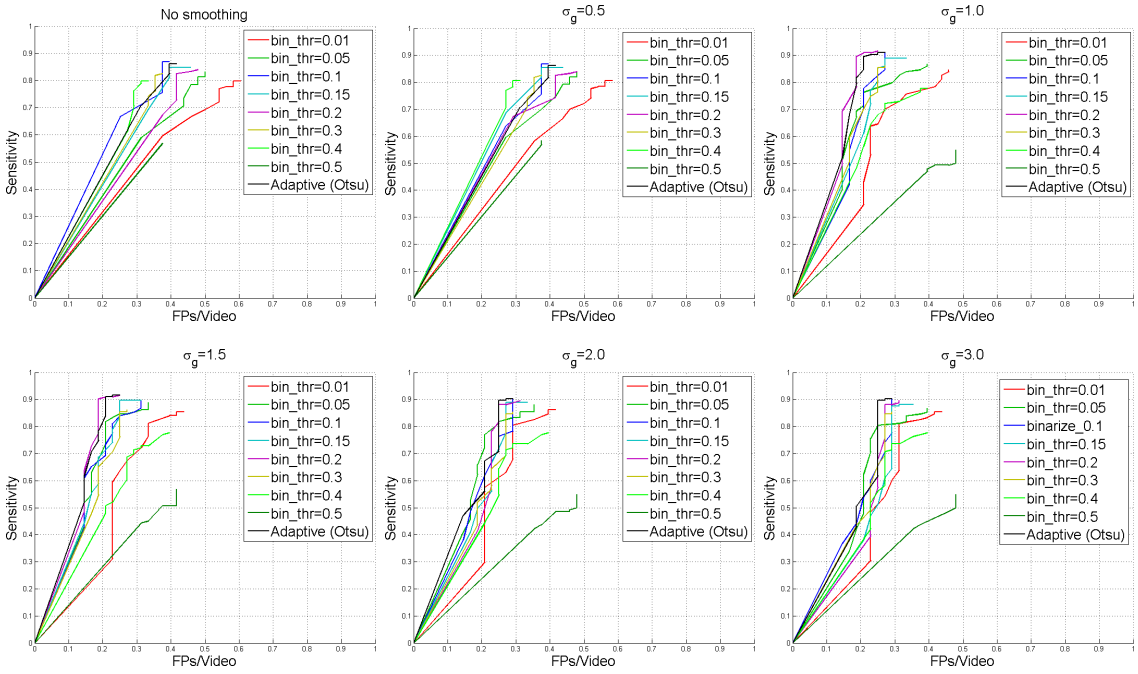

Figure 4.2: FROC curves of our system for automatic frame selection. Each plot shows FROC curves for different binarization thresholds and different levels of Gaussian smoothing.

of the proposed system for the test subjects using the optimal size of training patches. To demonstrate the effectiveness of our constrained ROI localization method, the performance of the unconstrained system is also included. In the constrained mode, The Eq. 3.1 for ROI localization is used whereas in the unconstrained mode, only the ROI is localized as the center of the largest connected component in the corresponding confidence map without considering the location of the carotid bulb. The constrained method achieves an average localization error of $0.19\,\mathrm{mm}$ and $0.35\,\mathrm{mm}$ in the constrained and unconstrained modes, respectively. The decrease in localization error is statistically significant ($p < 0.01$). Also as seen in Figure 4.6, the unconstrained method has resulted in 3 complete localization failures (outliers), which have been corrected in the constrained mode. Furthermore, compared with the hand-crafted approach (Sharma *et al.*, 2014), the proposed system using the con-

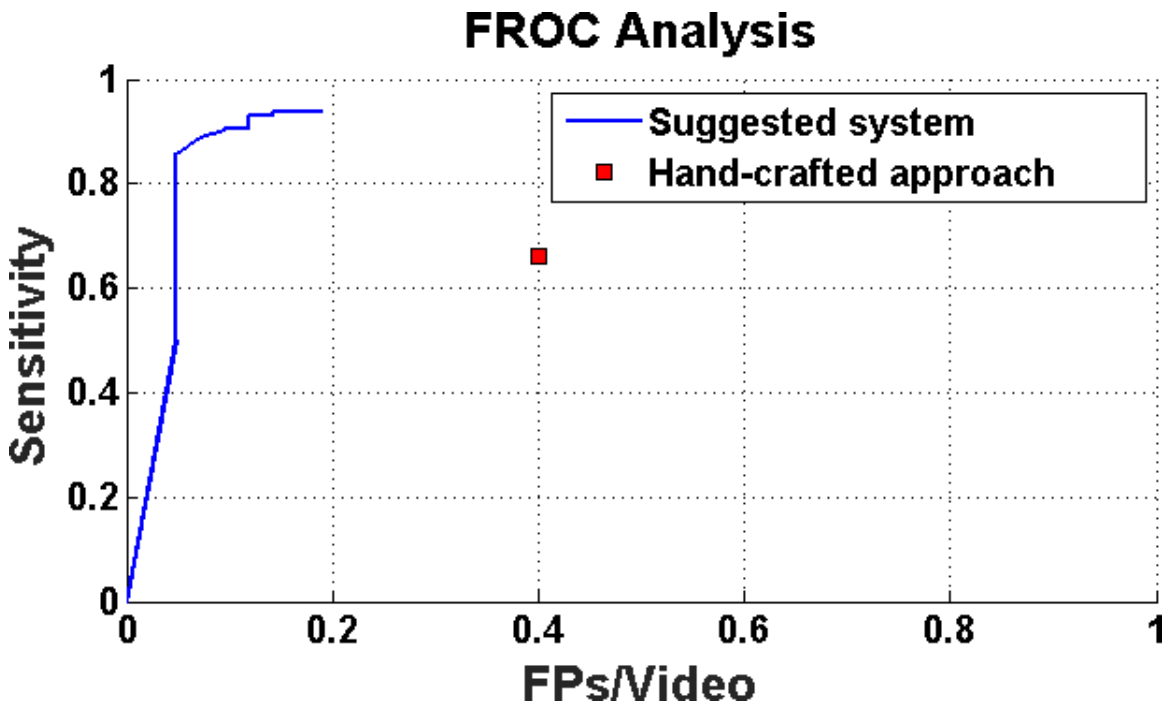

Figure 4.3: FROC curve of our frame selection system for the test subjects using the tuned parameters. For comparison, we have also shown the operating point of the prior hand-crafted approach (Sharma *et al.*, 2014), which is significantly outperformed by the suggested system.

strained mode shows a decrease of $0.1\,\mathrm{mm}$ in ROI localization error, which is statistically significant ($p < .00001$).

Intima-Media Thickness Measurement: The optimal image patch size by leave-one-patient-out cross-validation using various image patch sizes was 360×360 $\mu\mathrm{m}$ which achieved slightly lower localization error and fewer outliers in Figure 4.5-4.7. Figure 4.9 shows the interface localization error of our system on the test subjects, where we break down the overall localization error for lumen-intima and that of the media-adventitia interface as well as the hand-crafted approach (Sharma *et al.*, 2014) for each interface. We further analyzed agreement between our system and the expert with the Bland-Altman plot in Figure 4.8.

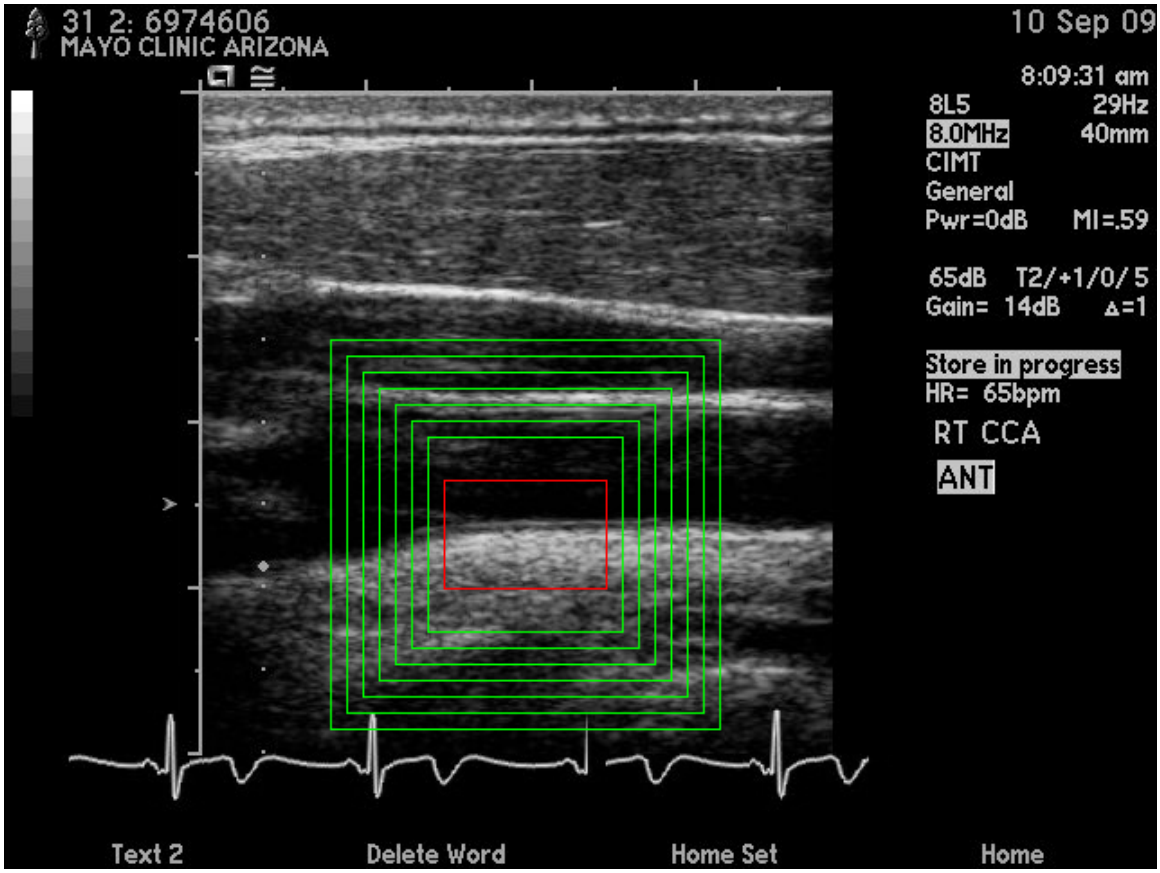

Figure 4.4: Various patches size from 1.2×1.2 cm to 2.4×2.4 cm shown in blue, and expert's ROI (1.0×0.75 cm) shown in red

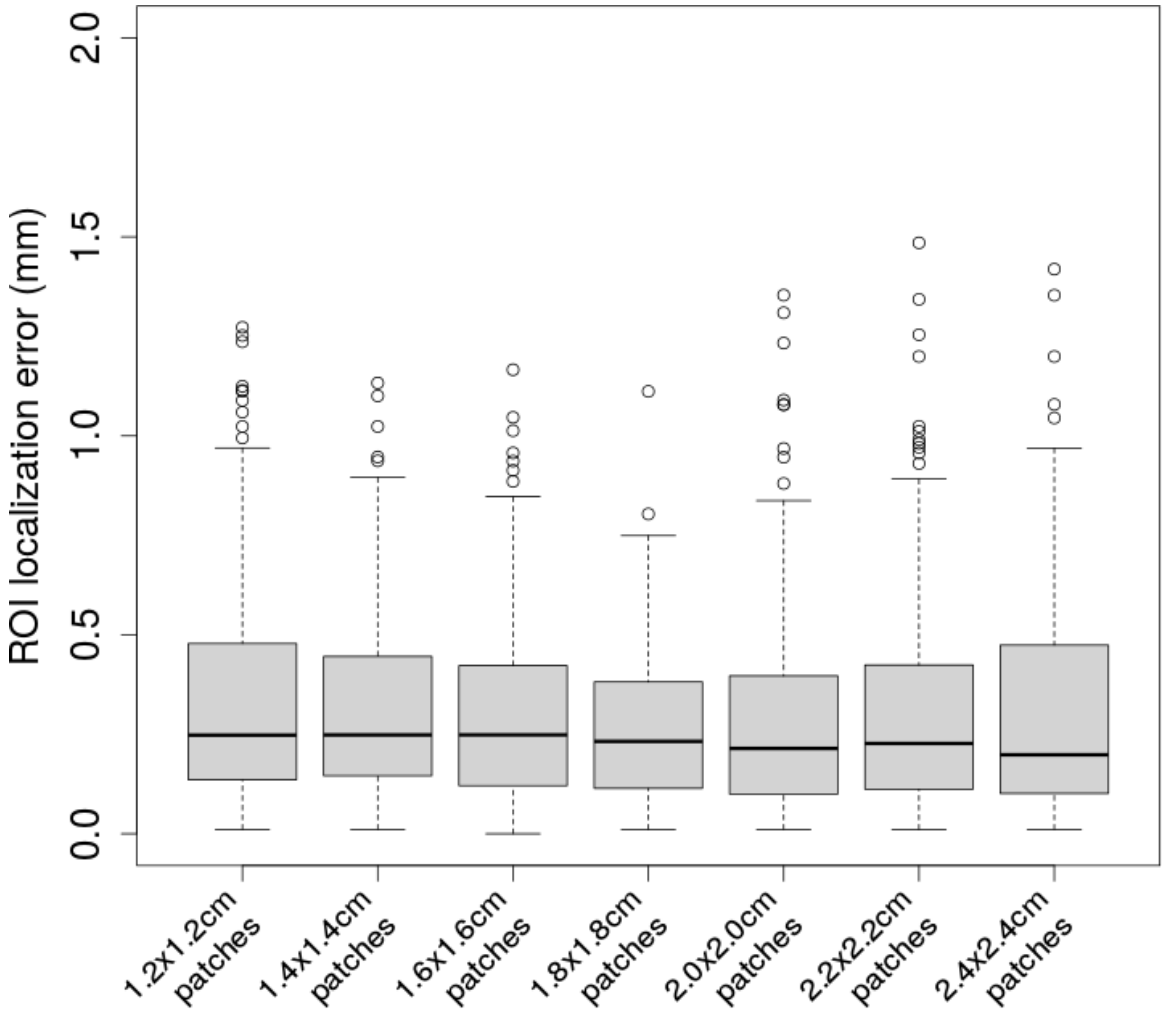

Figure 4.5: ROI localization error for different sizes of patches.

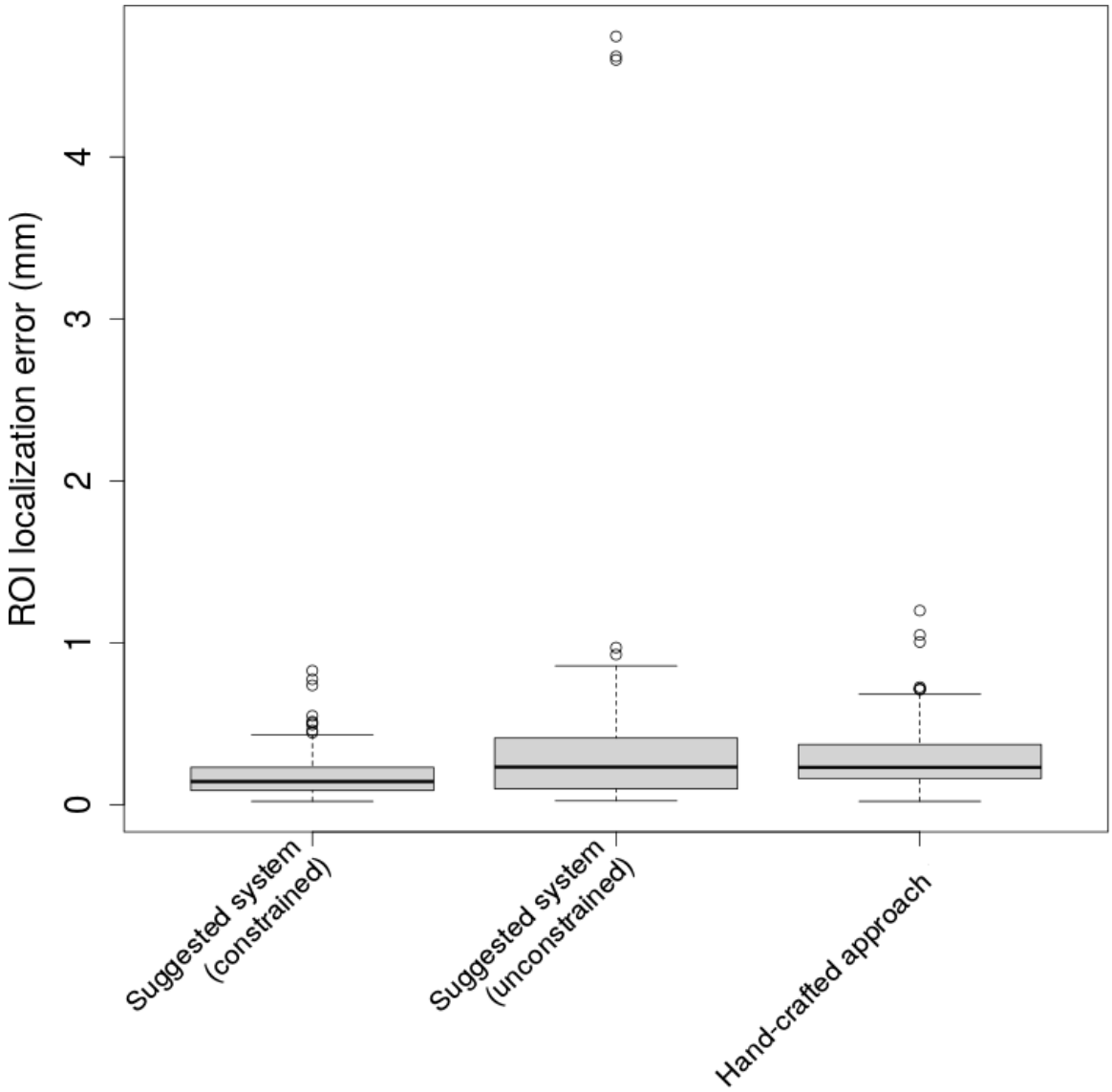

Figure 4.6: ROI localization error for the test subjects. Our method in the constrained mode outperforms both the unconstrained counterpart and the prior hand-crafted approach (Sharma *et al.*, 2014).

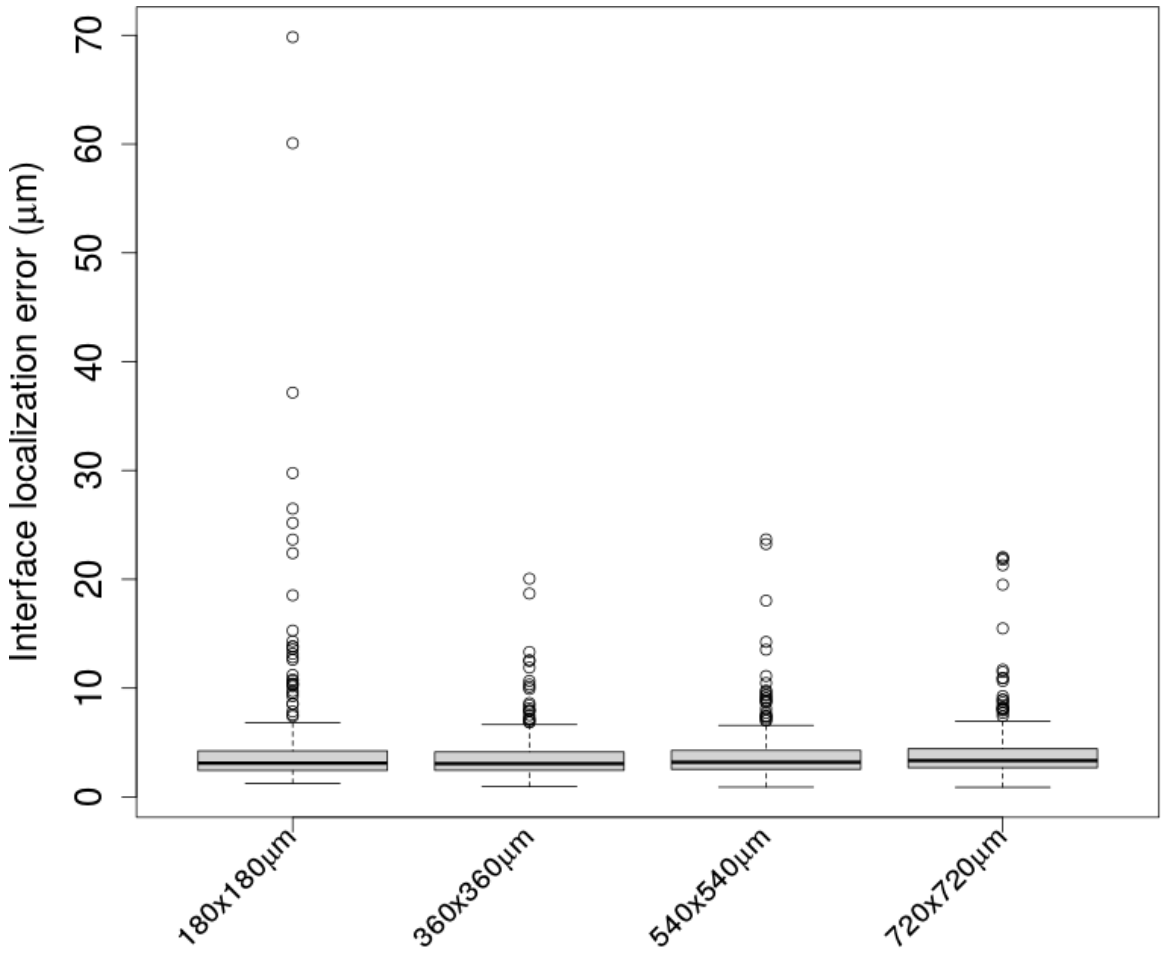

Figure 4.7: Combined interface localization error for different sizes of patches. The results are produced through a leave-one-patient-out cross-validation study based on the training subjects. Each box plots show the combined localization error of lumen-intima and media-adventitia interfaces for a different size of patches. In our analyses, we determine the localization error as the average of absolute vertical distances between our detected boundaries and the expert-annotated boundaries for the interfaces. As can be seen, while our system shows a high degree of robustness against different sizes of input patches, the use of patches of size 360×360 µm achieves slightly lower localization error and fewer outliers. Furthermore, this choice of patches yields higher computational efficiency compared to the larger counterpart patches.

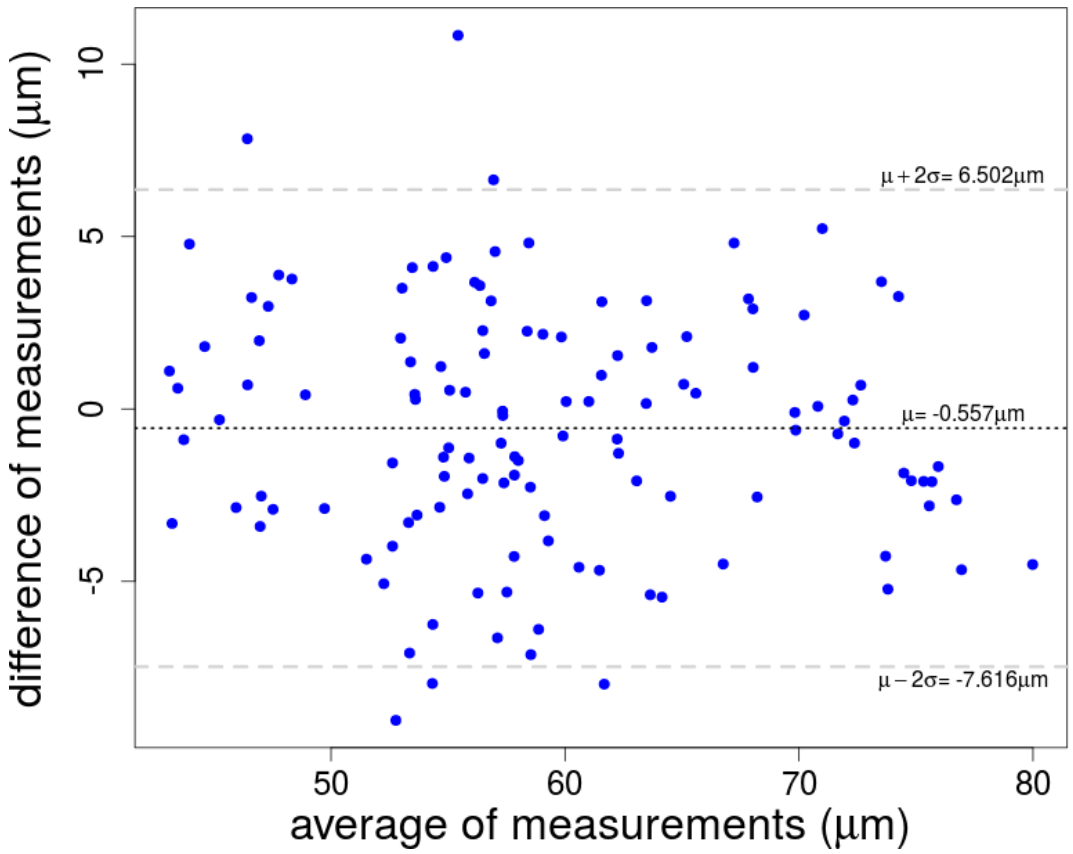

Figure 4.8: The Bland-Altman plot shows high agreement between our system and the expert for the assessment of intima-media thickness. Each circle in this plot represents a pair of thickness measurements from our method and the expert for a test ROI. In this plot, we have a total of 126 circles corresponding to 44 test videos.

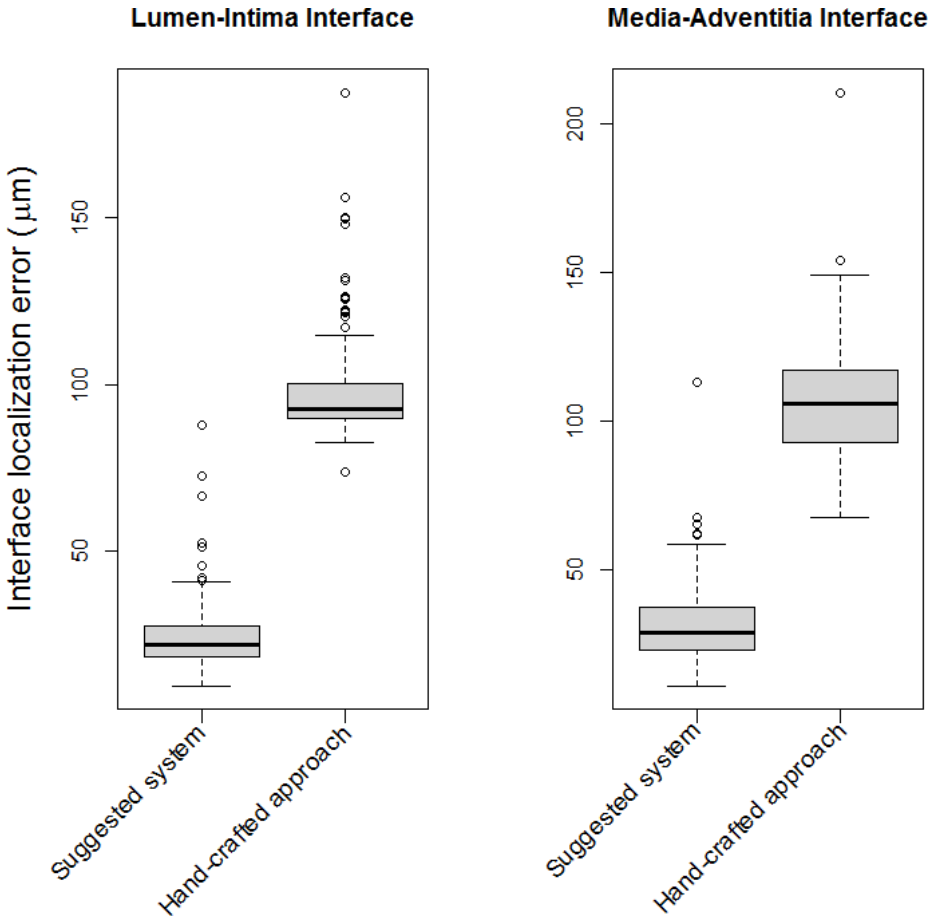

Figure 4.9: Localization error of the lumen-intima and media-adventitia interfaces for the suggested system and the prior hand-crafted approach (Sharma *et al.*, 2014). The results are obtained for the test subjects.

Chapter 5

DISCUSSIONS

5.1 Frame Selection

In Section 4, the choice of patch binarization and degree of Gaussian smoothing affected the accuracy of frame selection. Here, these findings provide important insights about the hyperparameters that were chosen. The binarization of the patches were chosen, because it reduces appearance variability and suppress the low-magnitude noise content in the patches. Without patch binarization, one can expect a large amount of variability in the appearance of wavelets can deteriorate the performance of the subsequent CNN (see Figure 4.2). The choice of binarization threshold is another important factor. The use of a high threshold results in the partial appearance of the wavelets in the resulting binary patches, reducing the discriminatory appearance features of the patches. A low threshold, on the other hand, can intensify noise content in the images, which decreases the quality of training samples and consequently a drop in classification performance. According to our analyses, it is difficult to find a fixed threshold that can both suppress the noise content and keep the shapes of the restored wavelets intact in all the collected patches. Otsu's method seems to overcome this limitation by adaptively selecting a binarization threshold according to the intensity distribution of each individual patch. For patches with intensity values between 0 and 1, the adaptive thresholds have a mean of 0.15 and standard deviation of 0.05. The wide range of adaptive thresholds explains why a constant threshold may not perform as desirably.

Gaussian smoothing of the probability signals is also essential for accurate frame selection. This is because the probability signals prior to smoothing exhibit high frequency

fluctuations, which may complicate the localization of the local maxima in the signals. The first cause of such high frequency changes is patch misplacement in the accumulated difference images. Recall that we extract the patches around the weighted centroids of the accumulated difference images. However, a large amount of noise content in the difference images may cause the weighted centroid to deviate from the center of the restored wavelet. In this case, the extracted patch may partially or completely miss the restored wavelet. This can manifest itself as a sudden change in the CNN output and as a result in the corresponding probability signal. The second cause of high frequency changes is the inherited high variance of CNNs. Use of ensemble of CNNs and data augmentation can alleviate this problem at a significant computation cost. Alternatively, we choose to mitigate these undesirable fluctuations using Gaussian smoothing for computational efficiency.

5.2 ROI

As described in Section 3.2, ROI localization constrained by the location of the carotid bulb method was chosen. This is because the bulb area appears as a relatively distinct dark area in the ultrasound frame. The distinct appearance of the carotid bulb is also confirmed by our experiments, where the average bulb localization error of $0.26\,\mathrm{mm}$ is obtained for the test subjects with only one failure case, which is more favorable than the average unconstrained ROI localization error of $0.38\,\mathrm{mm}$ with 3 failure cases. Therefore, the localization of the bulb area can be done more reliably than the localization of the ROI, which motivates the use of the bulb location as a guide for more accurate ROI localization. This constraint is integrated into the localization system through a post-processing mechanism (see Eq. 3.1). Alternatively, a regression CNN could be used where each pixel in the image directly votes for the location of the ROI. However, this approach may be hindered by lack of stable anatomical structures in noisy ultrasound images. A regression-based CNN for ROI localization is left as future work.

5.3 IMT

In Section 4, a high level of agreement between the proposed system and the expert for the assessment of intima-media thickness is shown. The suggested system achieves a mean absolute error of $28\,\mu\mathrm{m}$ with a standard deviation of $2.1\,\mu\mathrm{m}$ for intimia-media thickness measurements. However, this level of measurement error cannot hurt the interpretation of the vascular age, because there exists a minimum difference of $400\,\mu\mathrm{m}$ between the average intima-media thickness of healthy and high-risk population ($600\,\mu\mathrm{m}$ for healthy and ≥ 1000 $\mu\mathrm{m}$ for high-risk population) (Jacoby *et al.*, 2004).

5.4 CNN Architectures

The proposed system uses LeNet-like CNN architecture, but it does not limit the suggested framework to this architecture. In fact, deeper CNN architectures such as AlexNet (Krizhevsky *et al.*, 2012), VGGNet (Simonyan and Zisserman, 2014) and GoogleNet (Szegedy *et al.*, 2015) were explored in both training and fine-tuning modes; however, there was not any significant performance gain. This was probably because the higher level semantic features detected by the deeper networks are not very relevant to the tasks in our CIMT applications. Meanwhile, the concomitant computational cost of deep architectures may hinder the applicability of the proposed system, because it lowers the speed—a key usability factor of the system. A shallower architecture may not offer the performance required for clinical practice. This is because a network shallower than the LeNet has only one convolutional layer and thus limited to learning primitive edge-like features. Detecting the carotid bulb and the ROI, and segmenting intima-media boundaries are relatively challenging tasks, requiring more than primitive edge-like features. Similarly, for frame selection, classifying the restored wavelets into R-peak and non-R-peak categories is similar to digit recognition, for which LeNet is a common choice of architecture. Therefore, LeNet-like

CNN architecture seems to represent an optimal balance between efficiency and accuracy for CIMT video analysis.

5.5 Performance

On a desktop computer with a 3.6 Ghz quad core Intel with an Nvidia GTX 970 GPU, the proposed system detects each EUF in 2.9 seconds, localizes each ROI in 12.1 seconds, and measures intima-media thickness in 8.2 seconds. While the current speed is suitable for off-line processing of the CIMT videos, further performance speedup is required for an interactive use in the clinical practice. However, the use of CNNs does not hinder the interactive use of the proposed system in terms of time, rather, extracting a large number of patches from a dense set of locations in the ultrasound images causes a computational bottleneck. Therefore, significant performance speedup can be achieved by using fully convolutional networks (Long *et al.*, 2014) for patch extraction within GPU, which eliminates the need for computationally expensive image patch extraction in CPU. Further performance speedup can also be obtained using faster and/or more dedicated graphics cards.

One of the significant achievement is that all performance evaluations were performed without involving any user interactions. However, the goal of the proposed system is not to exclude the user (for example, sonographer) from the loop rather to relieve him from the three tedious, laborious, and time consuming operations by automating them while still offering the user a highly, user-friendly interface to bring his indispensable expertise onto CIMT interpretation through refining the automatic results easily at the end of each of the automated operations. For instance, the proposed system is expected to automatically locate a EUF within one frame, which is clinically acceptable, but in case the automatic selected EUF is not the one the user wants, the user can simply move one frame forward or backward in the interactive system. The automatically localized ROI by the proposed system is usually acceptable as long as there is a small distance from the ground truth

location, but the user still can easily drag the ROI and move it around as desired. Finally, in refining the automatically identified lumen-intima and media-adventitia interfaces, the original snake formulation comes with spring forces for user interaction (Kass *et al.*, 1988), but given the small distance between the lumen-intima and media-adventitia interfaces, the "movable" hard constraints as proposed in (Liang *et al.*, 2006) are far more effective than the spring forces in measuring CIMT.

Chapter 6

CONCLUSION

In this thesis, a unified framework to fully automate and accelerate CIMT video interpretation is presented. Specifically, a computer-aided CIMT measurement system with three components which are: (1) automatic frame selection in CIMT videos, (2) automatic ROI localization within the selected frames, (3) automatic intima-media boundary segmentation within the localized ROIs. Each of the above components on a CNN with a LeNet-like architecture is used and then boosted the performance of the employed CNNs with effective pre- and post-processing techniques. For frame selection, how patch binarization as a preprocessing step and smoothing the probability signals as a post-processing step improve the results generated by the CNN are analyzed. For ROI localization, the location of the carotid bulb, as a constraint in a post-processing setting, significantly improves ROI localization accuracy by proving experimentally. For intima-media boundary segmentation, open snakes were employed as a post processing step to further improve the segmentation accuracy. Then, the results produced by the suggested system were compared with those of the major prior works, demonstrating more accurate frame selection, ROI localization, and CIMT measurements. This superior performance is attributed to the effective use of CNNs coupled with pre- and post- processing steps, uniquely designed for the three CIMT tasks.

REFERENCES

Bastida-Jumilla, M., R. Menchón-Lara, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz and J. Sancho-Gómez, "Frequency-domain active contours solution to evaluate intima–media thickness of the common carotid artery", Biomedical Signal Processing and Control **16**, 68–79 (2015).

Bastida-Jumilla, M. C., R. M. Menchón-Lara, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz and J. L. Sancho-Gómez, "Segmentation of the common carotid artery walls based on a frequency implementation of active contours", Journal of digital imaging **26**, 1, 129–139 (2013).

Cheng, D.-C. and X. Jiang, "Detections of arterial wall in sonographic artery images using dual dynamic programming", Information Technology in Biomedicine, IEEE Transactions on **12**, 6, 792–799 (2008).

Delsanto, S., F. Molinari, P. Giustetto, W. Liboni, S. Badalamenti and J. S. Suri, "Characterization of a completely user-independent algorithm for carotid artery segmentation in 2-d ultrasound images", Instrumentation and Measurement, IEEE Transactions on **56**, 4, 1265–1274 (2007).

Faita, F., V. Gemignani, E. Bianchini, C. Giannarelli, L. Ghiadoni and M. Demi, "Real-time measurement system for evaluation of the carotid intima-media thickness with a robust edge operator", Journal of Ultrasound in Medicine **27**, 9, 1353–1361 (2008).

Goodfellow, I. J., D. Warde-Farley, M. Mirza, A. Courville and Y. Bengio, "Maxout networks", arXiv preprint arXiv:1302.4389 (2013).

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", arXiv preprint arXiv:1207.0580 (2012).

Hubel, D. H. and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex", The Journal of physiology **148**, 3, 574–591 (1959).

Hurst, R. T., R. F. Burke, E. Wissner, A. Roberts, C. B. Kendall, S. J. Lester, V. Somers, M. E. Goldman, Q. Wu and B. Khandheria, "Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players", The American journal of cardiology **105**, 8, 1107–1111 (2010).

Ilea, D. E., C. Duffy, L. Kavanagh, A. Stanton and P. F. Whelan, "Fully automated segmentation and tracking of the intima media thickness in ultrasound video sequences of the common carotid artery", Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on **60**, 1 (2013).

Jacoby, D., I. Mohler, EmileR. and D. Rader, "Noninvasive atherosclerosis imaging for predicting cardiovascular events and assessing therapeutic interventions", Current Atherosclerosis Reports **6**, 1, 20–26, URL `http://dx.doi.org/10.1007/s11883-004-0112-8` (2004).

Kass, M., A. Witkin and D. Terzopoulos, "Snakes: Active contour models", International journal of computer vision **1**, 4, 321–331 (1988).

Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in "Advances in neural information processing systems", pp. 1097–1105 (2012).

Le Cun, B. B., J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Handwritten digit recognition with a back-propagation network", in "Advances in neural information processing systems", (Citeseer, 1990).

LeCun, Y., Y. Bengio and G. Hinton, "Deep learning", Nature **521**, 7553, 436–444 (2015).

Liang, J., T. McInerney and D. Terzopoulos, "United snakes", Medical image analysis **10**, 2, 215–233 (2006).

Liang, Q., I. Wendelhag, J. Wikstrand and T. Gustavsson, "A multiscale dynamic programming procedure for boundary detection in ultrasonic artery images", Medical Imaging, IEEE Transactions on **19**, 2, 127–142 (2000).

Loizou, C. P., "A review of ultrasound common carotid artery image and video segmentation techniques", Medical & biological engineering & computing **52**, 12, 1073–1093 (2014).

Loizou, C. P., C. S. Pattichis, M. Pantziaris and A. Nicolaides, "An integrated system for the segmentation of atherosclerotic carotid plaque", Information Technology in Biomedicine, IEEE Transactions on **11**, 6, 661–667 (2007).

Long, J., E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", arXiv preprint arXiv:1411.4038 (2014).

Menchón-Lara, R.-M., M.-C. Bastida-Jumilla, A. González-López and J. L. Sancho-Gómez, "Automatic evaluation of carotid intima-media thickness in ultrasounds using machine learning", in "Natural and Artificial Computation in Engineering and Medical Applications", pp. 241–249 (Springer, 2013).

Menchón-Lara, R.-M. and J.-L. Sancho-Gómez, "Fully automatic segmentation of ultrasound common carotid artery images based on machine learning", Neurocomputing **151**, 161–167 (2015).

Molinari, F., K. M. Meiburger, L. Saba, G. Zeng, U. R. Acharya, M. Ledda, A. Nicolaides and J. S. Suri, "Fully automated dual-snake formulation for carotid intima-media thickness measurement a new approach", Journal of Ultrasound in Medicine **31**, 7, 1123–1136 (2012).

Molinari, F., G. Zeng and J. S. Suri, "A state of the art review on intima–media thickness (imt) measurement and wall segmentation techniques for carotid ultrasound", Computer methods and programs in biomedicine **100**, 3, 201–221 (2010).

Otsu, N., "A threshold selection method from gray-level histograms", Automatica **11**, 285-296, 23–27 (1975).

Petroudi, S., C. Loizou, M. Pantziaris and C. Pattichis, "Segmentation of the common carotid intima-media complex in ultrasound images using active contours", Biomedical Engineering, IEEE Transactions on **59**, 11, 3060–3069 (2012).

Pignoli, P. and T. Longo, "Evaluation of atherosclerosis with b-mode ultrasound imaging.", The Journal of nuclear medicine and allied sciences **32**, 3, 166–173 (1987).

Rossi, A. C., P. J. Brands and A. P. Hoeks, "Automatic localization of intimal and adventitial carotid artery layers with noninvasive ultrasound: a novel algorithm providing scan quality control", Ultrasound in medicine & biology **36**, 3, 467–479 (2010).

Sharma, H., R. G. Golla, Y. Zhang, C. B. Kendall, R. T. Hurst, N. Tajbakhsh and J. Liang, "Ecg-based frame selection and curvature-based roi detection for measuring carotid intima-media thickness", in "SPIE Medical Imaging", pp. 904016–904016 (International Society for Optics and Photonics, 2014).

Simonyan, K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556 (2014).

Stein, J., C. Korcarz, R. Hurst, E. Lonn, C. Kendall, E. Mohler, S. Najjar, C. Rembold and W. Post, "American society of echocardiography carotid intima-media thickness task force. use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: a consensus statement from the american society of echocardiography carotid intima-media thickness task force. endorsed by the society for vascular medicine", J Am Soc Echocardiogr **21**, 2, 93–111 (2008).

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 1–9 (2015).

Touboul, P.-J., P. Prati, P.-Y. Scarabin, V. Adrai, E. Thibout and P. Ducimetière, "Use of monitoring software to improve the measurement of carotid wall thickness by b-mode imaging.", Journal of hypertension **10**, S37–S42 (1992).

Wan, L., M. Zeiler, S. Zhang, Y. L. Cun and R. Fergus, "Regularization of neural networks using dropconnect", in "Proceedings of the 30th International Conference on Machine Learning (ICML-13)", pp. 1058–1066 (2013).

Xu, X., Y. Zhou, X. Cheng, E. Song and G. Li, "Ultrasound intima–media segmentation using hough transform and dual snake model", Computerized Medical Imaging and Graphics **36**, 3, 248–258 (2012).

Zhu, X., C. B. Kendall, R. T. Hurst and J. Liang, "A user friendly system for ultrasound carotid intima-media thickness image interpretation", in "SPIE Medical Imaging", pp. 79681G–79681G (International Society for Optics and Photonics, 2011).