2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Novel Deep Learning Approach for Tracking Regions of Interest in Ultrasound Images*

Mohammad Wasih[1] and Mohamed Almekkawy[1]

***Abstract*— Due to their great success in learning a universal object similarity metric, Siamese Trackers have been adopted for motion tracking a Region of Interest (ROI) in Ultrasound (US) image sequences. However, these Fully Convolutional Siamese networks (SiamFC) offer no online adaptation of the network and fail to take cues from the input sequence. The more recent Correlation Filter Networks (CFNet) solve this problem by learning the reference template online using a Correlation Filter layer. In this work, we use the CFNet as our backbone model and propose an advanced tracking algorithm (Seq-CFNet) for tracking an ROI in US sequences by constructing a sequential cascade of two identical CFNet. The cascade with CFNet is novel and offers practical benefits in tracking accuracy. Our method is evaluated on 10 different sequences of a Carotid Artery (CA) dataset to track the transverse section of the carotid artery. Results show that Seq-CFNet obtains better Root Mean Square Error (RMSE) values than the baseline CFNet as well as SiamFC, without significantly compromising the speed.**

***Index Terms*— Speckle Tracking, Correlation Filter Network, Siamese Network, Convolutional Neural Network, Cascaded Network**

I. INTRODUCTION

Motion tracking of objects in Ultrasound (US) images is of great clinical advantage as it lends numerous applications in US diagnostics. For instance, real-time tracking of muscles has been used to measure muscle and tendon contraction velocity and strain [1]. Similarly, speckle tracking has applications in assessing myocardial function [2]. Therefore, automated and accurate methods of tracking a given ROI in an US image sequence are beneficial. Furthermore, due to the instancy of the application, sub-pixel level tracking estimates would be ideal.

Conventional techniques like Block Matching (BM) [3] use a similarity metric like Normalized Cross Correlation (NCC) to exhaustively evaluate all possible candidate image blocks for finding the best match with the reference image block. Analysis of correlation-based approaches has been presented in [4]. Using a Kalman filter as a motion model for guiding the search position has also been explored in [5], [6] and [7]. For improving sub-pixel level accuracy estimates, interpolation methods in the Radio Frequency (RF) domain such as [8], kriging interpolation in [9] and iterative projection in [10] have been proposed.

However, the problem with these methods is that they are not largely data-driven and fail to exploit patterns in US images to improve localization. With the advancement of deep learning in the computer vision community, it, therefore, may be beneficial to adopt these deep learning techniques for motion tracking in US images. This is the approach followed in [11]. Here, a Fully Convolutional Siamese network (SiamFC) [12] has been used to track objects. Correlation Filter Network (CFNet) [13] is an advanced version of this network, where the reference template branch uses an adaptive Correlation Filter (CF) layer to update its template in the forward pass. To improve the sub-pixel level accuracy and to model the appearance of objects in a better way, a sequential cascade network that uses SiamFC as the backbone has been implemented in [14]. Taking motivation from this approach, we design a new sequential cascade method with CFNet as the backbone, for tracking an ROI in US images.

II. BACKGROUND

A. CFNet

CFNet is a more advanced version of SiamFC. The major problem with SiamFC is that it does not do any online adaptation. This strategy may not be optimal, since, if we incorporate specific patterns and cues from the video at run-time (forward pass), then, we can use it to tune the internal state of the network to our advantage. Similar to SiamFC, CFNet uses the same Convolutional Neural Network (CNN) to learn the low-dimensional embeddings for the reference and search blocks. The search image is denoted by z, candidate image by x. The CNN learns some embedding function, ϕ_p which has learnable parameters, p. While, SiamFC optimizes for p by minimizing the cross-correlation output, $h_p(z,x) = \phi_p(z) * \phi_p(x)$ with the ground-truth, CFNet however, takes an additional step by learning an updated template, w through Correlation Filter (CF). The embedded template, $\phi_p(z)$, is passed through the CF layer which is a well-known algorithm to efficiently update the template online, by solving a ridge regression problem. This is to say that the output of the reference template branch is, $h_{p,s,b}(z,x) = s\omega(\phi_p(z)) * \phi_p(x) + b$, where s and b are scale and bias parameters respectively and $w = \omega(\phi_p(z))$ is the CF layer [13]. CF finds the optimal template which is robust against translations of the object. The parameters of $\omega(x)$ are updated during the forward pass of the network. Finally, CFNet also updates its reference template using a weighted moving average. A practical benefit of CFNet is that it gives good results even if a weaker embedding network is used, due to the presence of the CF layer.

*This work was not supported by any organization

[1]Mohammad Wasih and Mohamed Almekkawy are with the School of Electrical Engineering and Computer Science, Pennsylvania State University, University Park, PA 16802, USA mvw5820@psu.edu

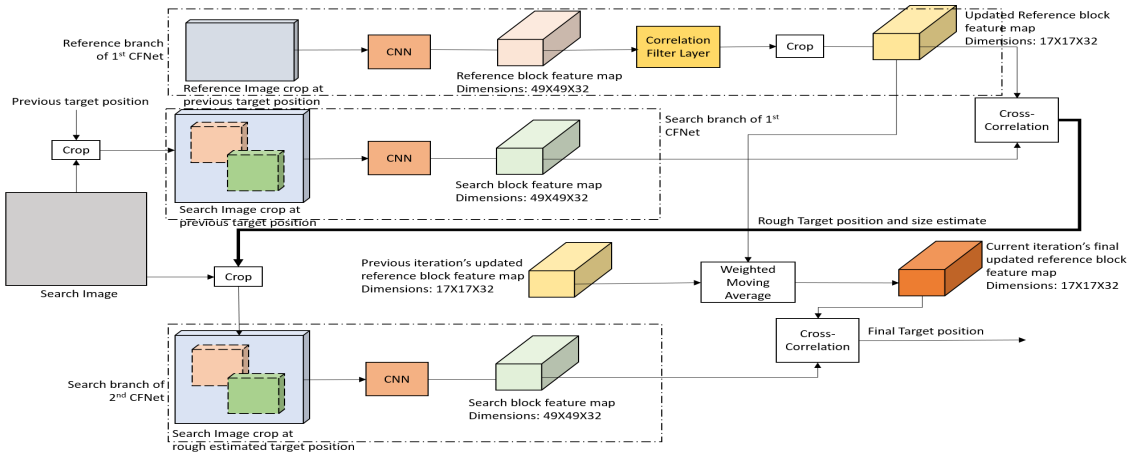

Fig. 1. Architecture diagram of the proposed cascaded sequential correlation filter network. At any given tracking step, the reference branch is used to compute the feature representation of the reference image using a Convolutional Neural Network (CNN). The search branch computes the feature representation of the search image (cropped at the previous target position). The target size is estimated using the scale at which the cross-correlation score was the maximum. The bold arrow in the diagram shows the sequential cascade step. The search image is cropped again at the rough position and passed through the search image branch of the 2^{nd} CFNet. Our architecture is novel due to the connection steps between the two CFNet: The reference branch is not evaluated again for this second step, instead, the previous feature map is combined with the current feature map using a weighted moving average sum.

In this work, we design a new tracking network based on the sequential cascade of two CFNet. Our network, although is motivated by [14], yet differs from it in considerable ways: We use two identical embedding functions as opposed to using slightly different networks. While [14] considers lower resolution images for the second pass, we do not change the resolution, as it affects US image quality and hence negatively affects the tracking accuracy. Finally, our network is based on CFNet as opposed to SiamFC and the connection is novel in the sense that the reference template is only evaluated in the first pass.

III. Methodology

A. Data

We use pre-trained networks (trained on the ILSVRC [15] general images dataset). Therefore, we consider the data for evaluation only. We consider the B-mode US images for the cross-section of the Carotid Artery (CA) as obtained from [16] - [18]. Ten different image sequences from the CA dataset were used to validate the efficacy of our networks. On average, each sequence contains 20 US frames with ground-truth annotations (centroid location and bounding box) of the ROI provided.

B. Architecture of the Neural Network

Fig. 1 shows the architectural diagram of our proposed neural network for tracking which we call, Seq-CFNet. As mentioned earlier, we consider CFNet as the backbone network. The tracking is done in two stages. First, the reference branch of the 1^{st} CFNet is evaluated to compute the updated template. The search image is then cropped at the previous predicted target location and evaluated through the search branch of the 1^{st} CFNet. After cross-correlation, we get the rough estimate of the target position.

To obtain a refined estimate, first, the search image is cropped at the predicted rough target position. Then, the 2^{nd} CFNet is used. However, only the search image and not the reference image, is evaluated through the search branch of the 2^{nd} CFNet. The rationale behind this novel approach is that it may not always be the case that the estimate after the 2^{nd} CFNet would be a local improvement in the rough target estimate. It may happen, for example, that the rough target estimate from the 1^{st} CFNet is slightly offset. In this case, the reference template would get wrongly estimated and this would lead to a further huge drift in the target's predicted position if the reference template is updated through the 2^{nd} CFNet again. Following the same reasoning, the target size is also updated only once. We also confirmed this design choice empirically and found that consecutive updates to the template, deteriorate performance significantly. Finally, the search embedding from the 2^{nd} CFNet is cross-correlated with the updated template to find the new target position.

Seq-CFNet's CNN is composed of 5 blocks of convolutional, batch normalization, and rectified linear activation layers. Furthermore, the same embedding function is used in the 2^{nd} CFNet to prevent over-adaptation of the network to the input sequence, and more importantly to reduce the memory footprint of the network, as the same network object can be evaluated twice without creating additional memory.

C. Experiments

We compared our network against the baseline CFNet (no sequential cascade) and SiamFC. The baseline CFNet is denoted as CFNet-5-Conv (i.e. it has 5 convolutional layers in its embedding function). Our proposed network is Seq-CFNet-5-Conv (both components of the cascade have 5 convolutional layers in the embedding function). We also consider another variant of our proposed network, Seq-CFNet-1-Conv which has a single convolutional layer in

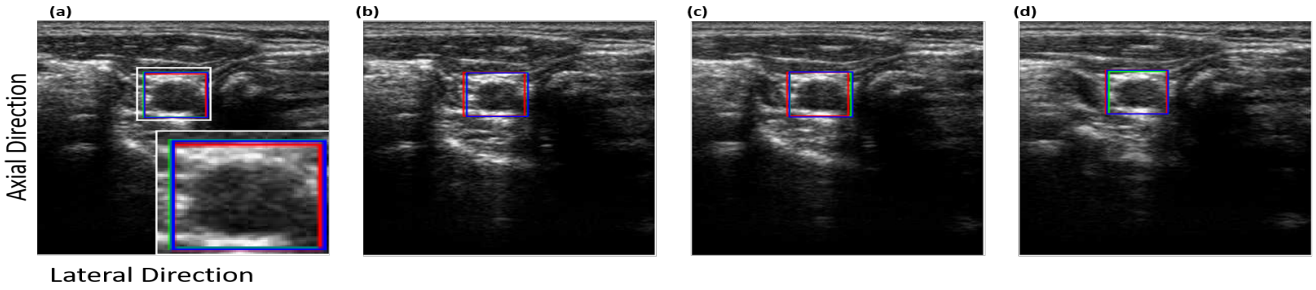

Fig. 2. (a) through (d) show 4 consecutive frames from one of the Carotid Artery (CA) dataset sequences. The red box represents the ground truth. The green box represents the output of the sequential cascade network (Seq-CFNet-5-Conv) and the blue box represents the output of CFNet-5-Conv. Note that the green box stays closest to the red box.

the embedding function to analyze the trade-off between speed and accuracy, as this network would be very light. The metrics we consider are RMSE, (between the network's centroid predictions and available ground truth), IoU (Intersection over Union, of the predicted and ground-truth bounding boxes), and Computational Time per Frame (CTF) (in seconds). For tracking, the first frame with ground truth was passed through the networks as the sole supervision.

IV. Results and Discussion

Fig. 2 shows the results obtained by running our best performing proposed tracker (Seq-CFNet-5-Conv) as well as the baseline tracker (CFNet-5-Conv) on one of the CA sequences.

A. RMSE of Centroid Locations

Fig. 3 shows the Lateral and Axial RMSE scores for each sequence of the dataset (computed with respect to the ground truth) for various networks. We note that SiamFC performs the worst and shows a lot of variance across the sequences. Seq-CFNet-5-Conv gives the best results in both, lateral and axial directions, as confirmed by Table I, which shows the results averaged over all the 10 sequences. Note that the total RMSE, for any given sequence, s_i, is defined as: $total_{RMSE}(s_i) = \sqrt{lateral_{RMSE}(s_i)^2 + axial_{RMSE}(s_i)^2}$. The difference in total RMSE between CFNet-5-Conv and Seq-CFNet-5-Conv is 0.12 which is significant for sub-pixel level accuracy. Seq-CFNet-1-Conv, though gives the lowest lateral RMSE, it performs poor on the axial RMSE. It, however, still does better than CFNet-5-Conv on the total RMSE.

B. Intersection over Union (IoU)

Table I summarizes the IoU results for the networks with respect to the ground truth, averaged over the 10 sequences. We observed that while SiamFC is less accurate, CFNet-5-Conv and Seq-CFNet-5-Conv gave comparable results with Seq-CFNet-1-Conv, slightly behind.

C. Computational Time per Frame (CTF) and Practical Utility

All the experiments were conducted on an Intel Core $i5$ 8^{th} Generation processor CPU running at 2.30 GHz clock rate with 8 GB of RAM. Average CTF over the 10 sequences from CA was noted as can be seen in Table II. It is expected that Seq-CFNet-5-Conv would be slow due to running of the CFNet-5-Conv twice, which itself has a large embedding network (5 convolutional layers). However, Seq-CFNet-1-Conv is more feasible as it is only 1.2 times slower than SiamFC, and gives considerably better results

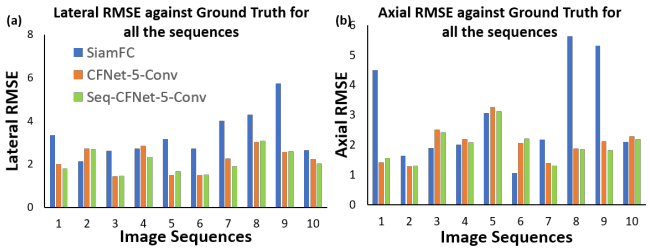

Fig. 3. Represents RMSE plots against ground truth for each of the 10 sequences in the dataset. (a) Lateral RMSE and (b) Axial RMSE

TABLE I
RMSE and IoU values for SiamFC, CFNet-5-Conv, Seq-CFNet-5-Conv and Seq-CFNet-1-Conv averaged over the entire 10 sequences from the CA dataset

Method	Lateral RMSE	Axial RMSE	Total RMSE	IoU(%)
SiamFC	3.550	3.220	4.792	83.080
CFNet-5-Conv	2.208	2.037	3.080	**87.718**
Seq-CFNet-5-Conv	2.110	**1.987**	**2.964**	87.538
Seq-CFNet-1-Conv	**1.969**	2.186	3.032	86.963

TABLE II
Memory Footprint and CTF of various methods considered

Method	Model Size(in MB)	CTF (in seconds)
SiamFC	8.571	**0.29**
CFNet-5-Conv	14.196	0.43
Seq-CFNet-5-Conv	14.196	0.92
Seq-CFNet-1-Conv	**0.136**	0.39

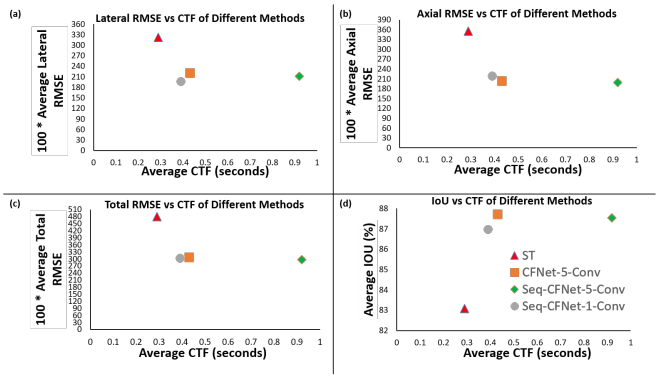

Fig. 4. (a) shows Lateral RMSE vs CTF plot for various methods. It shows that Seq-CFNet-1-Conv is quite efficient as it achieves the lowest RMSE without too much computation. (b) is similarly the plot for Axial RMSE vs CTF for various methods and here Seq-CFNet-5-Conv gives the best performance. (c) is the plot for total RMSE vs CTF for various methods. Note that in (a) through (c), 100 * RMSE is plotted due to low range of RMSE values. (d) shows IoU vs CTF plot for various methods, where CFNet-5-Conv gives slightly better results than Seq-CFNet-5-Conv. Note that all of these results represent the average over the 10 sequences.

than SiamFC in both RMSE and IoU metrics. Seq-CFNet-1-Conv also gives better results than CFNet-5-Conv, achieving a total RMSE of 3.032 as observed from Table I which is 0.05 pixels less than the latter.

Fig. 4 shows the CTF vs accuracy plots across all the metrics for the various methods considered in this study. The points in each plot are representative of the average results over the 10 sequences of the CA dataset. From the plots, it is observed that while Seq-CFNet-5-Conv gives the best RMSE results and near best IoU results, it is computationally more expensive. An efficient alternative is Seq-CFNet-1-Conv which gives the lowest lateral RMSE and slightly lesser overall RMSE than CFNet-5-Conv. At the same time, Seq-CFNet-1-Conv is considerably faster than CFNet-5-Conv and Seq-CFNet-5-Conv. SiamFC performs the worst in terms of accuracy in all the plots.

Table II also shows the model size for each method. As discussed before, we use the same network object for the sequential cascade, which allows us to save large amounts of memory. From the table, it is clear that Seq-CFNet-1-Conv just takes 0.136 MB of the memory which is considerably low. This is because it has a very weak embedding function. Due to the presence of the CF layer, even a weaker embedding network like Seq-CFNet-1-Conv is able to perform competitively.

V. CONCLUSION

In this work, we have designed a novel cascaded neural network based on the correlation filter network and adopted it for US images. CFNet offers a more efficient approach for tracking objects in US images as the accuracy and computational cost could be managed by varying the size of the embedding network. The sequential cascade further improves the CFNet by refining the predictions. Since we use pre-trained networks in our work, one way to further improve the accuracy of our approach would be to use a transfer learning approach by training only a few final layers of the networks on the ultrasound dataset. Our future work would involve improving the cascade further by training on US image datasets in an unsupervised manner and incorporating a motion model.

REFERENCES

[1] Sikdar, Siddhartha1; Wei, Qi1; Cortes, Nelson2 Dynamic Ultrasound Imaging Applications to Quantify Musculoskeletal Function, Exercise and Sport Sciences Reviews: July 2014 - Volume 42 - Issue 3 - p 126-135.

[2] Mondillo, S., Galderisi, M., Mele, D., Cameli, M., Lomoriello, V.S., Zacà, V., Ballo, P., D'Andrea, A., Muraru, D., Losi, M., Agricola, E., D'Errico, A., Buralli, S., Sciomer, S., Nistri, S. and Badano, L. (2011), Speckle-Tracking Echocardiography. Journal of Ultrasound in Medicine, 30: 71-83.

[3] Giachetti, Andrea. (2000). Matching techniques to compute image motion. Image Vision Comput.. 18. 247-260.

[4] B. Rebholz and M. Almekkawy, "Analysis of Speckle Tracking Methods: Correlation and RF Interpolation," 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), 2020, pp. 120-124.

[5] S. Bharadwaj and M. Almekkawy, "Faster Search Algorithm for Speckle Tracking in Ultrasound Images," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020, pp. 2142-2146.

[6] S. Bharadwaj, S. Prasad and M. Almekkawy, "An Upgraded Siamese Neural Network for Motion Tracking in Ultrasound Image Sequences," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control.

[7] S. Bharadwaj and M. Almekkawy, "Motion estimation for ultrasound image sequences using deep learning", The Journal of the Acoustical Society of America 148, 2487-2487 (2020).

[8] B. Rebholz and M. Almekkawy, "Constrained RF Level Interpolation for Normalized Cross Correlation Based Speckle Tracking," 2020 IEEE International Ultrasonics Symposium (IUS), 2020, pp. 1-4.

[9] B. Rebholz and M. Almekkawy, "Efficacy Of Kriging Interpolation In Ultrasound Imaging; Subsample Displacement Estimation," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020, pp. 2137-2141.

[10] Rebholz B, Zheng F, Almekkawy M. Two-dimensional iterative projection method for subsample speckle tracking of ultrasound images. Med Biol Eng Comput. 2020 Dec;58(12):2937-2951.

[11] S. Bharadwaj and M. Almekkawy, "Deep Learning Based Motion Tracking of Ultrasound Image Sequences," 2020 IEEE International Ultrasonics Symposium (IUS), 2020, pp. 1-4.

[12] Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." European conference on computer vision. Springer, Cham, 2016.

[13] Valmadre, Jack, et al. "End-to-end representation learning for correlation filter based tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[14] Fei Liu, Dan Liu, Jie Tian, Xiaoyan Xie, Xin Yang, Kun Wang, Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences, Medical Image Analysis, Volume 65, 2020, 101793, ISSN 1361-8415.

[15] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[16] ŘÍHA, K.; MAŠEK, J.; BURGET, R.; BENEŠ, R.; ZÁVODNÁ, E. Novel Method for Localization of Common Carotid Artery Transverse Section in Ultrasound Images Using Modified Viola-Jones Detector. ULTRASOUND IN MEDICINE AND BIOLOGY. 2013. 39(10). p. 1887 – 1902.

[17] BENEŠ, R.; BURGET, R.; KARÁSEK, J.; ŘÍHA, K. Automatically designed machine vision system for the localization of CCA transverse section in ultrasound images. COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE. 2013. 109(3). p. 92 – 103.

[18] ŘÍHA, K.; BENEŠ, R. Circle Detection in Pulsative Medical Video Sequence. In Proceedings of International Conference on Signal Processing, vol. I. Beijing, IEEE Press. 2010. p. 674 – 677.