Landmark Detection and Tracking in Ultrasound using a CNN-RNN Framework

Conference Paper · December 2016

Tao Xiong
Johns Hopkins University
21 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Arun Nair
Johns Hopkins University
9 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Sang Chin
Boston University
74 PUBLICATIONS **444** CITATIONS
SEE PROFILE

Landmark Detection and Tracking in Ultrasound using a CNN-RNN Framework

Akshay Rangamani[*]
Dept of Electrical and Computer Engineering
Johns Hopkins University
rangamani.akshay@jhu.edu

Tao Xiong[*]
Dept of Electrical and Computer Engineering
Johns Hopkins University
tao.xiong@jhu.edu

Arun Nair[*]
Dept of Electrical and Computer Engineering
Johns Hopkins University
anair8@jhu.edu

Trac D. Tran
Dept of Electrical and Computer Engineering
Johns Hopkins University
trac@jhu.edu

Sang Peter Chin
Department of Computer Science
Boston University
spchin@cs.bu.edu

Abstract

We present a novel framework for landmark detection and tracking in ultrasound data. Our method employs a convolutional neural network (CNN) encoder-decoder for landmark detection, coupled with a recurrent neural network (RNN) for encoding information from previous video frames of the object being tracked. We evaluated our method on the MICCAI CLUST 2015 De Luca u.a. (2015) challenge dataset, and have achieved promising results.

1 Introduction

Ultrasound is commonly used for detection and tracking of tissue landmarks for intervention and therapy . This task is complicated however by a host of factors such as patient respiration, patient movement, operator movement and noise. These issues are often exacerbated as a result of the probe commonly being held by the operator, and the acquisition sequences being long. Tracking algorithms have to to be designed to be robust to such challenges in order to have successful tracking in medical ultrasound applications.

Schnabel (2015) approached this problem by combining logDemons nonlinear registration with a moving window tracking method and leverages dense Scale Invariant Feature Transform as a similarity measure for registration. Makhinya u.a. (2015) uses elliptic and template-based models of vessels in the liver, coupled with a robust optic-flow framework to handle tracking. Kondo (2015) proposes using an extention of the kernelized correlation filter (KCF) for landmark tracking. Nouri u.a. (2015) employs metric learning, and trains a convolutional neural network to generate a low-dimensional embedding space such that patches showing the same landmark at their center have a small ℓ_2 distance in the embedding.

Convolutional Neural Networks (CNNs) (LeCun u.a. (1998)) have facilitated great breakthroughs in various computer vision tasks, including object detection and tracking. Recurrent Neural Networks

[*] Equal Contribution

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

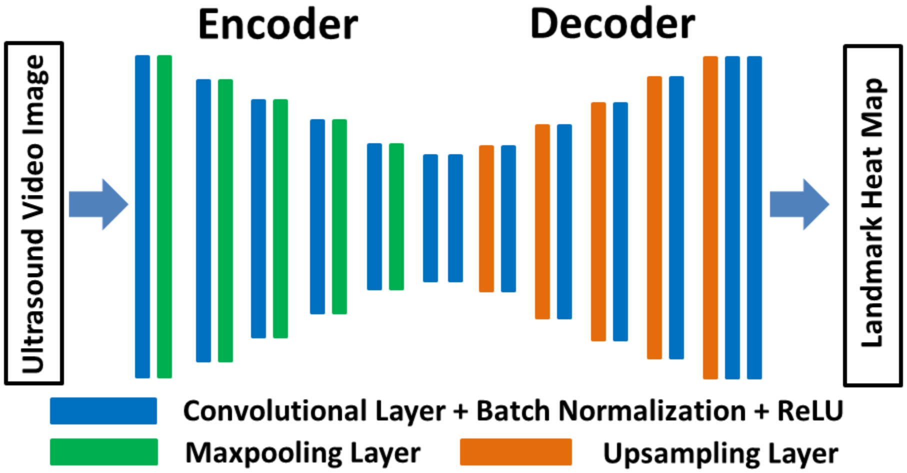

Figure 1: CNN Encoder-Decoder. Given the ultrasound video image, the CNN Encoder-Decoder automatically generates the heat maps of landmarks.

(RNNs) are another category of neural networks well-suited to the modeling of sequential data. Modern RNNs, especially Long Short-Term Memory Units (LSTMs) (Hochreiter u.a. (1997)) have proven to be quite adept at machine translation, language modeling, and image captioning.

We try to use neural networks in order to achieve more robust landmark detection and tracking in ultrasound data. Inspired by Badrinarayanan u.a. (2015), we developed a similar CNN-framework for the detection of landmarks in ultrasound images. We additionally decided to use an RNN to predict future frames of the ultrasound video, and use that model to augment our predictions. This idea is inspired by the use of RNNs in video modeling in Ranzato u.a. (2014) and Srivastava u.a. (2015). Our final output decision is generated based on a combination of predictions from the CNN and RNN. We evaluated our method on the MICCAI CLUST 2015 De Luca u.a. (2015) challenge dataset, and have achieved promising preliminary results.

The remainder of this paper is organized as follows. In Section 2, we elaborate on the CNN-RNN architecture developed and employed . In Section 3, we present the outputs of some of our tracking experiments on the MICCAI CLUST 2015 challenge dataset. We conclude our paper in Section 4, and propose possible future avenues of work.

2 Proposed Method

2.1 CNN Encoder-Decoder

Inspired by the deep encoder-decoder architecture SegNet, we incorporate a convolutional encoder-decoder neural network in our framework as shown in Figure 1. Each encoder layer consists of one convolutional layer with batch normalization and a ReLu non-linearity layer, which is followed by a maxpooling layer for downsampling. Similarly, each decoder consists of one convolutional layer with batch normalization and a ReLu non-linearity layer, followed by a upsampling layer. In order to meet the requirement of large receptive fields, we adopt 6 convolutional layers with the filter size of 3×3 and 5 non-overlapping maxpooling layers with the filter size of 2×2 in the encoder. The final convolutional layer generates the output of size $1 \times c \times m \times n$. In details, c, the channel number, indicates the number of landmarks. In each channel, the matrix of size $m \times n$, which is the same size as the input video frame, indicates the heat map of the corresponding landmark. Given the input ultrasound video frame, the encoder-decoder neural network itself can automatically generate the heat maps of landmarks as shown in Figure 2.

However, the encoder-decoder does not take the temporal dependencies between the ultrasound video's frames into account. Landmarks' heat maps are only decoded from the information which is

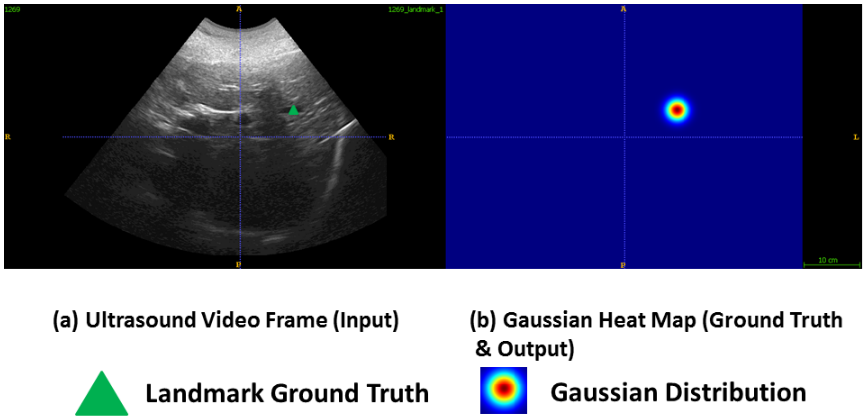
(a) Ultrasound Video Frame (Input)
(b) Gaussian Heat Map (Ground Truth & Output)
Landmark Ground Truth
Gaussian Distribution

Figure 2: (a) indicates the input to our framework, which is the frame of ultrasound video. (b) indicates the Gaussian heat maps as the ground truth in the training scheme or output in the testing.

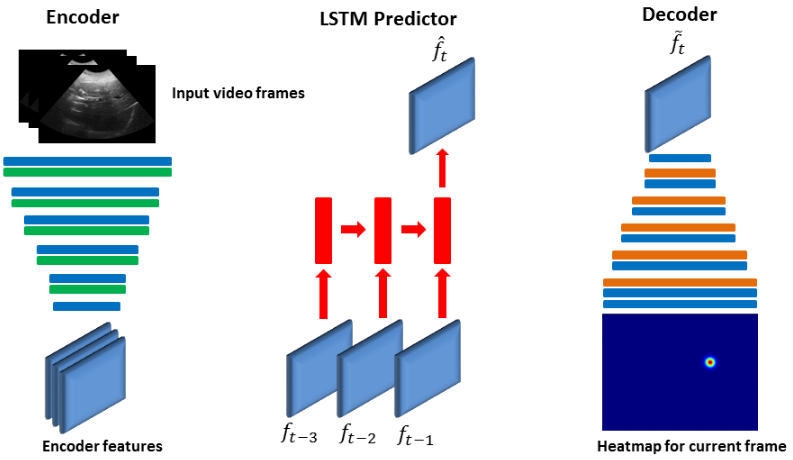

Figure 3: RNN Predictor that uses CNN features from 3 past frames to predict the current feature. The RNN uses the CNN encoder and decoder to generate tracking predictions from the input sequence

encoded from the current ultrasound video's frame. Intuitively, due to the continuous motion of organs, the correlation between ultrasound video's frames is able to further help predict the position of landmarks.

2.2 Temporal Modeling

We approach the Ultrasound Tracking problem for videos using RNNs. We use a subvolume of the past ultrasound frames, to predict the current frame and thus improve ultrasound tracking in the current frame. We integrate this approach with the Encoder-Decoder architecture described above in the following manner. In order to predict the heatmap for a particular frame in the Ultrasound video, we extract features from the past three frames using the encoder ($f_{t-3}, \ldots, f_{t-1}$), and use a 3-layer LSTM network to predict the encoder features for the next frame ($\hat{f}_t$). A combination of the predicted feature and the actual feature ($\tilde{f}_t = \lambda \hat{f}_t + (1-\lambda) f_t$) is then passed through the decoder to produce a heatmap for the current frame. We determine λ using a validation set. The complete architecture is depicted in Figure 3.

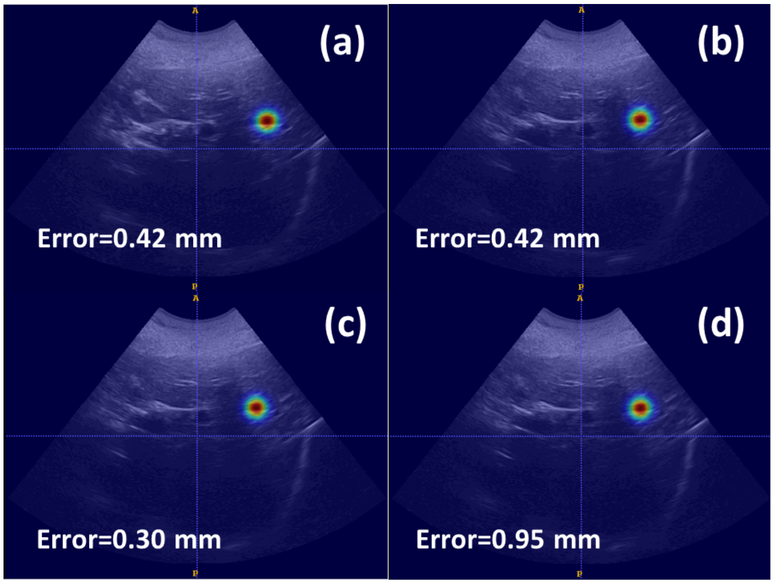

Figure 4: Tracking results from four different test frames. The predicted heatmaps overlay the input frames, and the error is reported.

3 Experiments

The MICCAI CLUST 2015 De Luca u.a. (2015); Banerjee u.a. (2014); Vijayan u.a. (2013); Bell u.a. (2012); Lediju u.a. (2010); Preiswerk u.a. (2014); De Luca u.a. (2013) dataset contains a total of 86 independent Liver Ultrasound tracking datasets

- 64 2D+t sequences
- 22 4D sequences

The data are anonymized and in the format of sequences of 2D images (.png) or 3D images (.mha). The data are split into a training and a test set. Landmark annotations are provided for the training set, to allow for some tuning of the tracking algorithm. Since we lacked access to the test set annotations, we used a subset of the CIL-01 data sequence to train our model and tested on the rest of the sequence.

The CIL-01 dataset has 1342 frames from an ultrasound video, with a landmark to be tracked in each frame. Annotations are provided for 144 frames of the CIL-01 sequence. We used 116 of those to train the CNN, and tested it on the remaining 28 frames. The RNN was trained on 1000 frame feature vectors which did not contain the 28 test frames. When combining the CNN and RNN, we have a hyperparameter λ which determines the linear combination between the prediction from past frames and the feature vector from the current frame. To find the best value for λ (0.2), we used a subset 10 training frames as a validation set, and we report the performance on our 28 frame test set. Our preliminary results (in table 1) seem promising, as compared to the state-of-the-art results on the challenge. Figure 4 shows the results of prediction for four different test frames. **We note that the tracking accuracies reported by those algorithms are on the original testing set, the annotations for which are not available to us.**

We train the CNN encoder-decoder and the RNN predictor separately. The CNN is trained on frame-heatmap pairs, while the RNN predictor is trained on 3 frame feature sequences. We use Adagrad with a learning rate of 0.001 and train both networks for 200 epochs using an NVIDIA Titan X GPU.

4 Conclusion

In this paper, we proposed a CNN-RNN framework for Ultrasound landmark detection and tracking. Using an encoder-decoder and recurrent neural network, we achieved a mean error of 0.51 mm on

Table 1: 2015 CLUST Results

	Tracking Error(mm)		
Participant	Mean	Standard Deviation	95th percentile
Schnabel (2015)	0.91	1.66	2.20
Makhinya u.a. (2015)	1.09	1.75	2.42
Kondo (2015)	1.09	1.35	3.07
Nouri u.a. (2015)	2.83	4.86	13.13
Our approach (without RNN)	0.62	0.35	1.22
Our approach (with RNN, $\lambda = 0.2$)	**0.51**	0.31	0.95

our test set, which is promising compared to the state-of-the-art results on CLUST. In the immediate future, we would like to obtain results on tracking in the entire dataset, not just a subset. We would also like to investigate end to end training of the network architecture, as well as incorporating convolutional layers into the RNN instead of the dense connections that we have now. We would also like to test our networks on other tracking problems, in both medical and non-medical data sets.

References

Hochreiter, Sepp / Schmidhuber, Jürgen(1997): *Long short-term memory*, 8: 1735–1780.

LeCun, Yann / Bottou, Léon / Bengio, Yoshua / Haffner, Patrick(1998): *Gradient-based learning applied to document recognition*, 11: 2278–2324.

Lediju, Muyinatu A / Byram, Brett C / Harris, Emma J / Evans, Philip M / Bamber, Jeffrey C(2010): *3D Liver tracking using a matrix array: Implications for ultrasonic guidance of IMRT*In: 2010 IEEE International Ultrasonics Symposium1628–1631.

Bell, Muyinatu A Lediju / Byram, Brett C / Harris, Emma J / Evans, Philip M / Bamber, Jeffrey C(2012): *In vivo liver tracking with a high volume rate 4D ultrasound scanner and a 2D matrix array probe*, 5: 1359.

De Luca, Valeria / Tschannen, Michael / Székely, Gábor / Tanner, Christine(2013): *A learning-based approach for fast and robust vessel tracking in long ultrasound sequences*In: International Conference on Medical Image Computing and Computer-Assisted Intervention518–525.

Vijayan, Sinara / Klein, Stefan / Hofstad, Erlend Fagertun / Lindseth, Frank / Ystgaard, Brynjulf / Langø, Thomas(2013): *Validation of a non-rigid registration method for motion compensation in 4D ultrasound of the liver*In: 2013 IEEE 10th International Symposium on Biomedical Imaging792–795.

Banerjee, Jyotirmoy / Klink, Camiel / Peters, Edward D / Niessen, Wiro J / Moelker, Adriaan / Walsum, Theo van (2014): *4D liver ultrasound registration*In: International Workshop on Biomedical Image Registration194–202.

Preiswerk, Frank u.a.(2014): *Model-guided respiratory organ motion prediction of the liver from 2D ultrasound*, 5: 740–751.

Ranzato, MarcAurelio / Szlam, Arthur / Bruna, Joan / Mathieu, Michael / Collobert, Ronan / Chopra, Sumit(2014): *Video (language) modeling: a baseline for generative models of natural videos*.

De Luca, V u.a.(2015): *The 2014 liver ultrasound tracking benchmark*, 14: 5571.

Nouri, Daniel / Rothberg, Alex(2015): *Liver Ultrasound Tracking using a Learned Distance Metric*.

Kondo, Satoshi(2015): *Liver Ultrasound Tracking Using Kernelized Correlation Filter With Adaptive Window Size Selection*.

Makhinya, Maxim / Goksel, Orcun(2015): *Motion Tracking in 2D Ultrasound Using Vessel Models and Robust Optic-Flow*.

Schnabel, Julia A(2015): *Robust Liver Ultrasound Tracking using Dense Distinctive Image Features*.

Badrinarayanan, Vijay / Kendall, Alex / Cipolla, Roberto(2015): *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*.

Srivastava, Nitish / Mansimov, Elman / Salakhutdinov, Ruslan(2015): *Unsupervised learning of video representations using lstms*.