Region-aware Deep Localization Framework for Cervical Vertebrae in X-Ray Images

S M Masudur Rahman Al Arif[1], Karen Knapp[2] and Greg Slabaugh[1]

[1]City, University of London
[2]University of Exeter

Abstract. The cervical spine is a flexible anatomy and vulnerable to injury, which may go unnoticed during a radiological exam. Towards building an automatic injury detection system, we propose a localization framework for the cervical spine in X-ray images. The proposed framework employs a segmentation approach to solve the localization problem. As the cervical spine is a single connected component, we introduce a novel region-aware loss function for training a deep segmentation network that penalises disjoint predictions. Using data augmentation, the framework has been trained on a dataset of 124 images and tested on another 124 images, all collected from real life medical emergency rooms. The results show a significant improvement in performance over the previous state-of-the-art cervical vertebrae localization framework.

Keywords: Cervical spine, X-ray, Localization, FCN, Region-aware.

1 Introduction

The cervical spine is a critical part of human body and is vulnerable to high-impact collisions, sports injuries and falls. Roughly 20% of the injuries remain unnoticed in X-ray images and 67% of these missed injuries end in tragic consequences [1,2]. Computer-aided-detection has the potential to reduce the number of undetected injuries on radiological images. Towards this goal, we propose a robust spine localization framework for lateral cervical X-ray radiographs. We reformulated the localization problem as a segmentation problem at a lower resolution. Given a set of high-resolution images and manually segmented vertebrae ground truth, at a lower resolution, the ground truth becomes a single connected region. We train a deep segmentation network to predict this region. To force the network to predict a single connected region, we introduce a novel term in the loss function which penalizes small disjoint areas and encourages single region prediction. This novel loss has produced significant improvement in localization performance. Previous work in vertebrae localization includes generalized Hough transform based approaches [3,4] and more recent random forest based approaches [5–7]. The state-of-the-art (SOTA) work on cervical vertebrae localization [7], uses a sliding window technique to extract patches from the images. A random forest classifier decides which patches belong to the spinal area. Then, a rectangular bounding box is generated to localize the spinal region. In contrast, the proposed framework can produce localization map of arbitrary shape in a one-shot process and provides a localisation result that models the cervical spine better than a rectangular box. We have trained our framework on a dataset of 124 images using data augmentation and tested on a separate 124 images having different shapes, sizes, ages and medical conditions. An average pixel level accuracy of 99.1% and sensitivity of 93.6% was achieved. There are two key contributions of this paper. First, a novel loss function which constrains the segmentation to form a single connected region and second, the adaptation and application of deep segmentation networks to cervical spine localization in real-life emergency room X-ray images. The networks learn from a small dataset and robustly outperform the SOTA both quantitatively and qualitatively.

2 Data

Our dataset contains 248 lateral view emergency room X-ray images collected from Royal Devon and Exeter Hospital. Image size, orientation, resolution, patient position, age, medical conditions and scanning systems all vary greatly in the dataset. Some images can be seen in Fig. 1. Along with the images, our medical partners have provided us with the manual segmentation of the cervical vertebrae, C3-C7. The top two vertebrae, C1-C2, were excluded from the study as ground truth was only available for C3 to C7. The segmentation (green) and localization (blue) ground truth (GT) for the images are highlighted in Fig. 1.

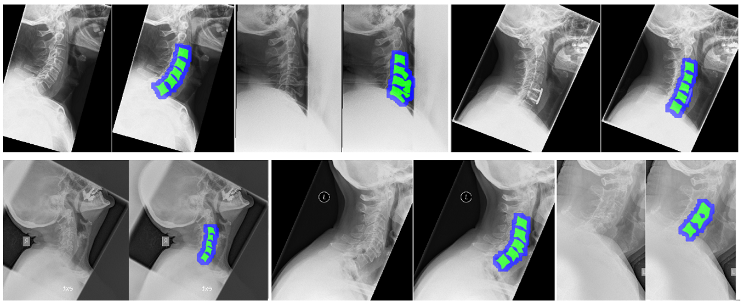

Fig. 1: Examples of X-ray images and corresponding ground truth.

3 Methodology

We have approached the localization problem as a segmentation problem at a lower resolution. The X-ray images are converted into square images by padding an appropriate number zeros in the smaller dimension and the square images are resized to a lower resolution using bicubic interpolation. This resolution can vary based on the available memory and size of the training networks. For our case, we chose this resolution to be 100 × 100 pixel. The corresponding binary segmentations of the vertebrae are also resized to the same resolution. At this resolution, the provided vertebrae segmentation becomes a single localized area encompassing the spine (blue region in Fig. 1). For this work, we have experimented with three different deep segmentation architectures: fully convolutional network (FCN) [8], deconvolutional network (DeConvNet) [9] and UNet [10]. In this work, we train the networks from scratch. The networks take an input X-ray image of 100 × 100 pixels and produce a probabilistic binary segmentation map of the same resolution.

3.1 Localization Ground Truth

As stated earlier, our target is to localize the spinal area in a cervical X-ray image. For this purpose, we convert our manual vertebra segmentations to a localization ground truth. As our networks are designed to produce an output localization map of 100 × 100 pixels, we create our localization ground truth in these dimensions. Since our original image sizes are approximately in the range of 1000 to 5000 pixels, a simple bicubic interpolation based resize of the vertebra segmentation produces a connected localization ground truth in the smaller dimension. To visualize the ground truth, it can be transformed back to the original dimensions. The blue overlay in Fig. 1 shows how much area the localization ground truth covers apart from the actual vertebrae (green).

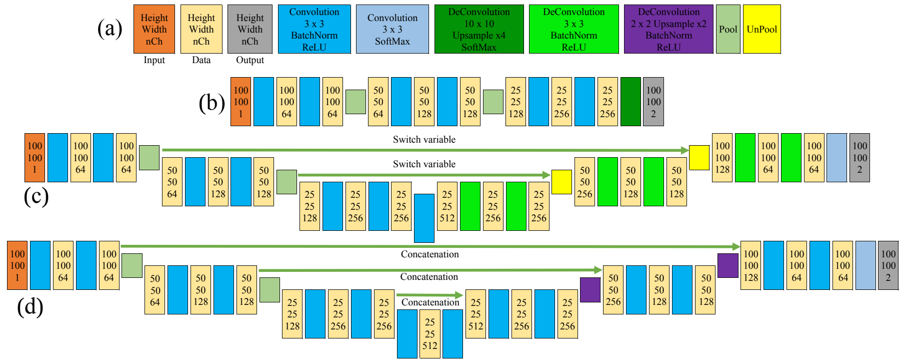

Fig. 2: (a) Legends (b) FCN (c) DeConvNet (d) UNet.

3.2 Network Architectures

The original FCN [8] and DeConvNet [9] were designed to tackle a semantic segmentation problem having multiple classes on natural images. Since our task here is to localize the spinal region, we essentially have a binary segmentation problem. Thus, we use a shallower version having fewer parameters. In our implementation, the FCN network has six convolutional layers and two pooling layers (size 2 × 2, stride 2). The two stages of pooling reduce the dimension from 100 × 100 to 25 × 25 thus creating an activation map of smaller size. The final deconvolutional layer upsamples the 25 × 25 activations to 100 × 100 pixels, producing an output map of the input size. Instead of upsampling from the lower resolution to the input resolution in a single step, DeConvNet uses a deconvolutional network which expands the activations step by step using a series of deconvolutional and unpooling layers. The expanding path forms a mirrored version of the contracting convolutional path. The UNet follows a similar structure but instead of an unpooling layer, it uses deconvolution to upsample the input. Both UNet and DeConvNet use information from the contracting path in the expanding path. DeConvNet does this through switch variables from the pooling layers and UNet uses concatenation of data. Fig. 2 shows the network diagrams that include data sizes after each layer for a single input image. The number of filters in each layer can be tracked from the number of channels in the data blocks. In total, our FCN has 1,199,042 parameters whereas DeConvNet and UNet have 4,104,194 and 6,003,842 parameters, respectively.

3.3 Training

We have a small dataset of only 248 manually segmented images. We divide our data randomly into 124 training and 124 test images. In order to train any network with a large number of parameters, 124 images are not enough. In order to increase the number of training data, we have augmented the images by rotating each image from 5° to 355° with a step of 5°. This results in a training set of 8928 images. It also made the framework rotation invariant. Our choice for data augmentation was only limited to rigid transformations since non-rigid transformation will affect the natural appearance of the spine in the image. All the networks were trained from randomly initialized parameters using a mini-batch gradient descent optimization algorithm from this augmented training dataset.

Given a dataset of training image (x)-segmentation label (y) pairs, training a deep segmentation network means finding a set of parameters $\boldsymbol{W}$ that minimizes a loss function, L_t. The simplest form of the loss function for segmentation problem is the pixel-wise log loss.

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \sum_{n=1}^{N} L_t(\{x^{(n)}, y^{(n)}\}; \boldsymbol{W}) \tag{1}$$

where N is the number of training examples and $\{x^{(n)}, y^{(n)}\}$ represents n-th example in the training set with corresponding manual segmentation. The pixel-wise segmentation loss per image can be defined as:

$$L_t(\{x, y\}; \boldsymbol{W}) = -\sum_{i \epsilon \Omega_p} \sum_{j=1}^{M} y_i^j \log P(y_i^j = 1 | x_i; \boldsymbol{W}) \tag{2}$$

$$P(y_i^j = 1 | x_i; \boldsymbol{W}) = \frac{\exp(a_j(x_i))}{\sum_{k=1}^{M} \exp(a_k(x_i))} \tag{3}$$

where $a_j(x_i)$ is the output of the penultimate activation layer of the network for the pixel x_i, Ω_p represents the pixel space, M is the number of class labels and P are the corresponding class probabilities. However, this term doesn't constrain the predicted maps to be connected. Since the objective of the localization problem is to find a single connected region encompassing the spine area, we add a novel region-aware term in the loss function to force the network to learn to penalize small and disconnected regions.

3.4 Region-aware Term

We translate our domain knowledge into the training by adding a region based term, L_r. This term forces the network to produce a single region by penalizing small disjoint regions. This term can be defined as:

$$L_r(\{x, y\}; \boldsymbol{W}) = \frac{1}{2} \sum_{i \epsilon \Omega_p} \sum_{j=1}^{M} y_i^j E_i P^2(y_i^j = 1 | x_i; \boldsymbol{W}) \tag{4}$$

$$E_i = \begin{cases} \max(N_r - N_t, 0) \frac{A_{max_t} - A_q}{A_{max_t}} & \text{if } i \epsilon R_q \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where N_r is the number of regions predicted as spine regions, N_t is the number of target regions we are looking for, A_q is the area of the q-th region, A_{max_t} is area of the t-th largest region, R_q is the set of pixels in the region q, and q represents the regions having area less than A_{max_t}. In our case, $N_t = 1$. Notice that, if N_r is equal to or less than N_t and/or $A_{max_t} = A_q$, no region based error will be added with the loss function of Eqn. 1.

3.5 Updated Loss Function

Finally, the loss function of Eqn. 1 can be extended as:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \sum_{n=1}^{N} L_t(\{x^{(n)}, y^{(n)}\}; \boldsymbol{W}) + L_r(\{x^{(n)}, y^{(n)}\}; \boldsymbol{W}) \tag{6}$$

The contribution of each term in the total loss can be controlled by introducing a weight parameter in Eqn. 6. However, in our case, the best performance was achieved when both terms contributed equally.

3.6 Experiments and Inference

The networks are trained on a server with two NVidia Quadro M4000 GPUs. Training each network for 30 epochs took 22 to 30 hours. Batch size during training was selected as 10 images and RMSprop [11] version of mini-batch gradient descent algorithm was used to update the parameters in every epoch. We have three different networks and two versions of the loss function, with and without the region-aware term. In total six networks have been trained: FCN, DeConvNet, UNet and FCN-R, DeConvNet-R, UNet-R, '-R' signifying if the region-aware term of Eqn. 6 has been used.

When testing, a test image is padded with zeros to form a square, resized to 100 × 100 pixels and fed forward through the network to produce probabilistic localization map. This map is converted into a single binary map and compared with the corresponding localization ground truth. Pixel level accuracy, object level Dice, sensitivity and specificity are computed. These metrics demonstrate the performance of the trained networks at the lower resolution at which the network generates the prediction. From a practical point of view, the performance of the localization should also be computed at the original resolution with the manually segmented vertebrae ground truth. In order to achieve this, the predicted localization map is transformed (resized and unpadded) back to the original image dimension and sensitivity and specificity are computed by comparing them with vertebrae segmentation.

4 Results and Discussions

The mean and standard deviation of the metrics over 124 test images at lower and original resolutions are reported in Table 1. In all cases and all metrics (other than specificity), inclusion of the region-aware term in the loss function improves the performance. The improvements are statistically significant for most of the metrics according to a paired t-test at a 5% significance level (bold numbers signify statistical significance for that metric over the other version of the same network architecture in the table). It can be noted that as the sensitivity increases, the specificity may decrease. This is because when the predicted region increases in size to cover more spinal regions, it may also start to encompass some other regions. This effect can also be seen in the qualitative results in Fig. 4. However, the specificity is always in the high range of 97.2% to 97.7%. Quantitatively, FCN performs better than UNet and DeConvNet. But qualitatively UNet and DeConvNet produce finer localization maps (Fig. 4b, d). The coarser map for FCN can be attributed to the single stage upsampling strategy of the network. Fig. 4 shows some of the difficult images in the test dataset: osteoporosis (a), image artefacts (b) and severe degenerative change (c, e). In most of these images, our region-aware term has been able to produce better results. It also decreases the standard deviation of the metrics (Table 1 and Fig. 3) proving its usefulness in regularizing the localized maps. However, outliers in the box plot of Fig. 3 show that there are images where all methods fail. Most of these images have severe clinical issues. One example of a complete failure of our algorithm for an image with bone implants is shown in Fig. 4f.

Table 1: Quantitative results (%).

	Lower resolution				Original resolution	
Networks	Pixel Accuracy	Dice	Sensitivity	Specificity	Sensitivity	Specificity
SOTA [7]	Not available				79.9 ± 16.7	98.1 ± 1.1
FCN	98.9 ± 1.0	83.8 ± 15.6	80.2 ± 18.1	**99.7 ± 0.3**	91.5 ± 18.4	**97.5 ± 0.7**
FCN-R	**99.1 ± 1.0**	**85.8 ± 14.7**	**85.0 ± 17.6**	99.6 ± 0.3	**93.6 ± 17.6**	97.2 ± 0.8
DeConvNet	98.6 ± 1.1	79.7 ± 16.4	77.2 ± 19.1	99.5 ± 0.5	88.2 ± 20.4	97.3 ± 0.8
DeConvNet-R	**99.1 ± 1.0**	**85.7 ± 15.6**	**81.0 ± 18.1**	**99.8 ± 0.2**	**91.5 ± 18.3**	**97.7 ± 0.7**
UNet	98.9 ± 1.1	84.1 ± 18.5	79.9 ± 21.9	99.8 ± 0.3	87.5 ± 23.3	**97.6 ± 0.8**
UNet-R	99.0 ± 1.0	85.1 ± 17.0	**82.5 ± 20.8**	99.7 ± 0.2	**89.6 ± 21.1**	97.4 ± 0.8

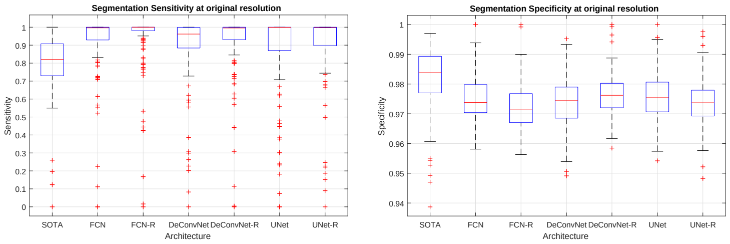

Fig. 3: Box plot of quantitative metrics.

To compare with the previous state-of-the-art (SOTA) in cervical vertebra localization [7], we have implemented and trained the random forest based framework on our dataset. Our algorithm produces a 17.1% relative improvement in average sensitivity with a drop of only 0.9% in specificity. In terms of time required for the algorithm to produce a result, our slowest framework (UNet) is approximately 60 times faster than [7]. To prove the robustness, we have tested

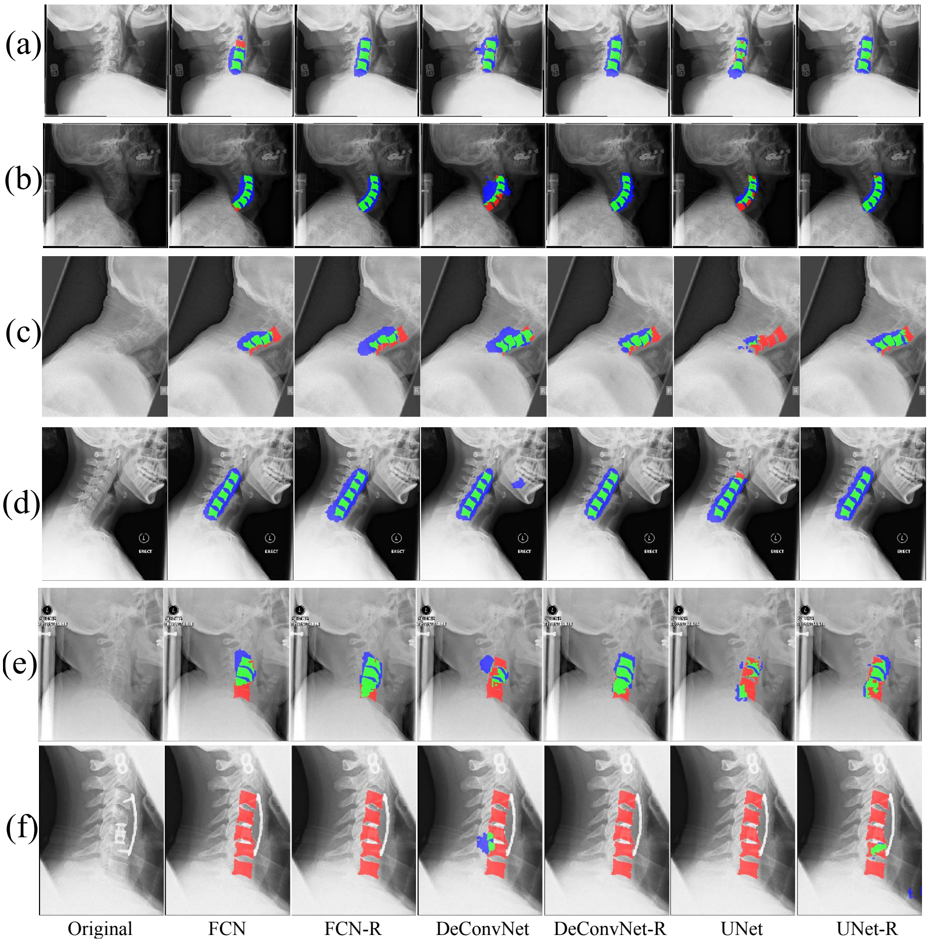

Fig. 4: Qualitative results: true positive (green), false positive (blue) and false negative (red) based on manual vertebrae segmentation.

the proposed framework on 275 cervical X-ray images of NHANES-II dataset [12] and even without any adaptation or transfer learning on the networks, it showed promising capability of generalization in localizing the cervical spine. However, due to insufficient ground truth information, quantitative results are not available. A few qualitative localization results on this dataset are shown in Fig. 5.

5 Conclusion

In this paper, we have proposed a framework for spine localization in cervical X-ray images. The localization problem has been reformulated as a binary

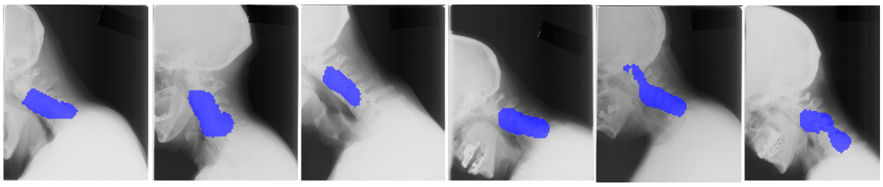

Fig. 5: Localization results (blue overlay) on NHANES-II dataset.

segmentation problem in a lower resolution. Based on the domain knowledge, a novel region-aware term was added to the loss function to produce a single region as localized output. Three segmentation networks were investigated and the novel loss function improved the performance of all these networks significantly. A maximum average sensitivity of 93.6% and specificity of 97.7% was achieved. Currently, we are adapting the proposed method for precise vertebrae segmentation. In future work, we plan to build a fully automatic computer-aided detection system for cervical spine injuries.

References

1. P. Platzer, N. Hauswirth, M. Jaindl, S. Chatwani, V. Vecsei, and C. Gaebler, "Delayed or missed diagnosis of cervical spine injuries," *Journal of Trauma and Acute Care Surgery*, vol. 61, no. 1, pp. 150–155, 2006.
2. C. Morris and E. McCoy, "Clearing the cervical spine in unconscious polytrauma victims, balancing risks and effective screening," *Anaesthesia*, vol. 59, no. 5, pp. 464–482, 2004.
3. A. Tezmol, H. Sari-Sarraf, S. Mitra, R. Long, and A. Gururajan, "Customized Hough transform for robust segmentation of cervical vertebrae from X-ray images," in *Image Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on*, pp. 224–228, IEEE, 2002.
4. M. A. Larhmam, S. Mahmoudi, and M. Benjelloun, "Semi-automatic detection of cervical vertebrae in X-ray images using generalized hough transform," in *Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference on*, pp. 396–401, IEEE, 2012.
5. B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pp. 590–598, Springer, 2012.
6. B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine CT via dense classification from sparse annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 262–270, Springer, 2013.
7. S. M. M. R. Al-Arif, M. Gundry, K. Knapp, and G. Slabaugh, "Global localization and orientation of the cervical spine in x-ray images," in *Computational Methods and Clinical Applications for Spine Imaging (CSI), The Fourth MICCAI Workshop on*, Springer, 2016.
8. E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2016.
9. H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528, 2015.
10. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.
11. S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
12. "NHANES-II Dataset." `https://ceb.nlm.nih.gov/proj/ftp/ftp.php`. Accessed: 2017-02-19.