Probabilistic Spatial Regression using a Deep Fully Convolutional Neural Network

S M Masudur Rahman Al Arif[1]
http://www.student.city.ac.uk/~acjv981
Karen Knapp[2]
http://emps.exeter.ac.uk/staff/kmk201
Greg Slabaugh[1]
http://www.staff.city.ac.uk/~sbbh653

[1] Department of Computer Science
City, University of London
London, UK

[2] University of Exeter Medical School
Exeter, UK

Abstract

Probabilistic predictions are often preferred in computer vision problems because they can provide a confidence of the predicted value. The recent dominant model for computer vision problems, the convolutional neural network, produces probabilistic output for classification and segmentation problems. But probabilistic regression using neural networks is not well defined. In this work, we present a novel fully convolutional neural network capable of producing a spatial probabilistic distribution for localizing image landmarks. We have introduced a new network layer and a novel loss function for the network to produce a two-dimensional probability map. The proposed network has been used in a novel framework to localize vertebral corners for lateral cervical X-ray images. The framework has been evaluated on a dataset of 172 images consisting 797 vertebrae and 3,188 vertebral corners. The proposed framework has demonstrated promising performance in localizing vertebral corners, with a relative improvement of 38% over the previous state-of-the-art.

1 Introduction

Deep convolutional neural networks (CNNs) have revolutionized the field of computer vision and artificial intelligence. Since 2012, different deep neural networks have produced state-of-the-art performance in image classification [9, 11, 18, 19] and segmentation [14, 16, 17, 20] problems. Classification and segmentation networks produce a probabilistic distribution over the output classes in the dataset. However, regression using CNNs is not usually probabilistic [8, 13]. A CNN based probabilistic regression method was proposed in [15] to address this issue. It utilizes a probabilistic interpretation of the Euclidean regression loss function to enforce a set of known constraints on the output space. Here, we propose a novel CNN based approach to produce a 2D spatial probabilistic distribution for localizing image landmarks. Instead of finding a set of constraints on the output space like [15], we convert the output space into a 2D probability distribution having the same spatial resolution of the input and train a CNN to learn the direct modelling from the input image to the spatial probability map. The motivation behind this work came from the need for a probabilistic corner

localization algorithm for cervical vertebrae in X-ray images. Corners provide vital information about different pathological conditions of the subject and can be used for initialization of vertebral segmentation methods [2, 7, 10].

Motivated by the success of fully convolutional networks (FCN) on medical image segmentation problems [4, 6, 16], we modify an FCN model, UNet, to generate a spatial probability map for vertebral corners over the input image space. In order to generate a probability map instead of a segmentation, a new spatial normalization layer has been introduced replacing the softmax layer used in classification and segmentation networks. A novel Bhattacharyya coefficient based loss function is proposed which can quantify the similarity between two probability distributions. The network learns to predict a probability map for vertebral corners in the image. Multiple corner locations can be predicted from a single input image patch. A complete semi-automatic framework has been designed which uses the patch-level corner prediction capability of the proposed probabilistic spatial regressor network to localize all vertebral corners in an X-ray image. The network is trained on a dataset of only 124 X-ray images. A total of 70,620 training patches were generated using data augmentation. The complete framework has been tested on 172 images consists of 797 cervical vertebrae and 3,188 corners. An average corner localization error of 1.56 mm has been achieved, signifying a 38% relative improvement over the previous state-of-the-art [1].

2 Data

A total of 296 lateral cervical spine X-ray images were collected from Royal Devon and Exeter Hospital in association with the University of Exeter. The collected data was not taken under a controlled environment. The age of the patients, scanning systems, X-ray intensity, image resolution, size, zoom, crop, spine position and patient position all varied according to the situation of the emergency department. The images include examples of vertebrae with fractures, degenerative changes and bone implants. Five vertebrae, C3 to C7 are considered for this study. C1 and C2 have been excluded from the study due to their ambiguous appearance in lateral cervical radiographs, similar to other cervical spine image analysis research [1, 2, 3, 12]. The images were received in two sets. The first set of 124 images are used for training and the rest, 172, are kept for testing. Each vertebral body from

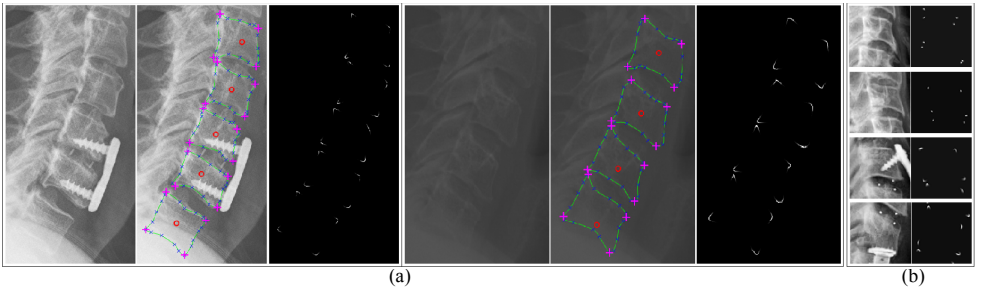

Figure 1: (a) Zoomed X-ray images (left), manual annotations (middle): center (o), manually clicked boundary points (x), corner points (+) and splined vertebrae curve (-) and corner distributions (right) (b) Image patches and patch level ground truth probability distribution.

the images was manually annotated for the vertebral boundaries and centers by expert radiographers. Two examples with the corresponding manual annotations are shown in Fig. 1a. The corner point of a cervical vertebra is often not well defined because of the smooth transition of the vertebral body. Thus manually clicked corner points vary substantially from expert to expert and from vertebrae to vertebrae. This variation makes it difficult for machine learning algorithms to learn a single deterministic model for corner prediction. This led us to consider probability distributions to represent the corners instead of a single point. The probability distribution is generated by setting a Gaussian distribution along the splined vertebral boundary centered at each corner. The variance of the Gaussian is proportional to the vertebral height and width and aligned to the vertebral orientation. The height, width and orientation are computed based on the manually annotated center points. To create the training image patches and the ground truth probability distributions, a set of uniformly distributed grid points are generated using the manually clicked vertebrae centers. From each point, multiple image patches are extracted with different scales and rotations. A total of 73,620 training patches were extracted from 124 training images. The patches were resized to a size of 64×64 pixel at which the proposed network is trained. A few vertebra patches are shown in Fig. 1b with their corresponding ground truth distributions.

3 Methodology

The overview of the proposed deep FCN-based corner localization framework is summarized in Fig. 2. The framework is semi-automatic. Given a test image, an operator will manually click on the vertebrae centers. From these manually clicked center points, a set of patches is generated. Each of these image patches is sent forward through the novel probabilistic spatial regression network described in Sec 3.1. The network generates patch level spatial probability distributions for corners in each patch. The patches are then transformed back on the original image space using their known location, orientation and size. Finally, the vertebral corners are localized from the accumulated patch distribution. These last steps are part of the post-processing phase and described in Sec. 3.3.

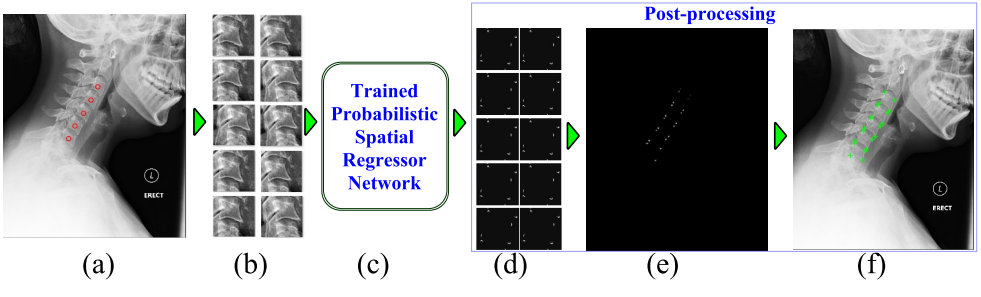

Figure 2: Framework block diagram: (a) Input image with manually clicked vertebrae centers (b) Image patches (c) Proposed network (d) Patch level predictions (e) Image level prediction (f) Localized corners.

3.1 Network

The network in Fig. 2c is trained on the training image patches to learn a model for predicting high probabilities at the vertebral corner locations. We want to keep the size of the input image patch and the output distribution space same. For this purpose, we chose to use a UNet-like architecture which has been successfully applied to medical image segmentation [2, 16]. Our network takes a single channel input patch of size 64×64 and learns to generate a single channel probabilistic distribution of the same size. The size, 64×64, is arbitrary and can be changed based on the available memory in the system if desired. For our dataset and available computational system, 64×64 was a break-even point between losing too much detail and too long training time. The network consists of a downsampling path and an upsampling path. The downsampling is done by max-pooling operations and upsampling is achieved by deconvolutional layers. The down and upsampling paths share information in the form of concatenation of data matrices. Our network has a total of 19 convolutional layers. Each convolutional layer is followed by a batch normalization and a rectified non-linear units (ReLU) except the last convolutional layer. In similar classification and segmentation network architectures, the final convolutional layer is followed by a softmax layer which converts final multi-channel activation of the convolutional layer into a probability distribution over the possible class labels. However, in our case, the final activation of the network is a single channel output which will be compared with a 2D spatial probability distribution over the input image space. Thus a new layer is needed to convert the final activation into a valid spatial probability distribution. One choice could have been doing softmax-like operation spatially, but as our input patches have multiple corners with high probabilities (Fig. 1b), the exponential nature of the softmax function often results in a single localized corner. Thus we introduce a new spatial normalization layer, which converts the final activation of the network into a valid spatial probability distribution using a simple mathematical operation by forcing the minimum to be zero and the integration to be unity. The network is shown in Fig. 3. The number of kernels in each convolutional and deconvolutional layer can be tracked from the number of channels in the intermediate data blocks. The total number of parameters in the network is 24,237,633.

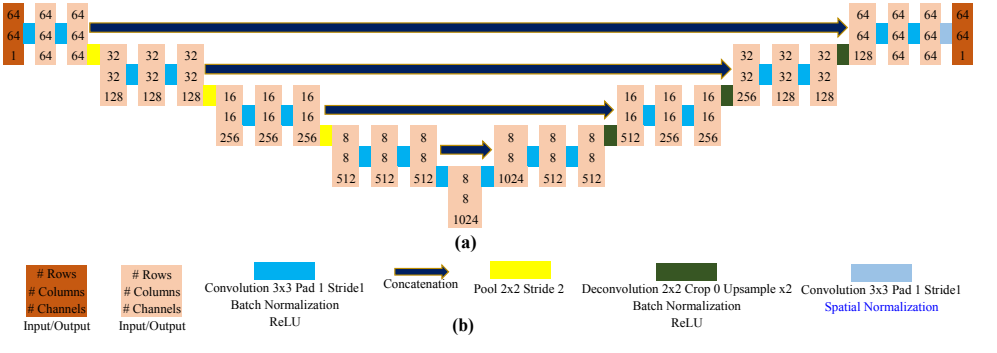

Figure 3: (a) Network architecture (b) legends.

3.2 Training

Given a dataset of training patch (x)- ground truth probability distribution (y) pairs, training a deep neural network means finding a set of parameters $\hat{W}$ that minimizes a loss function, L,

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

where N is the number of training examples and $\{x^{(n)}, y^{(n)}\}$ represents n-th example in the training set with corresponding ground truth corner probability distribution. The last layer of the network, spatial normalization layer, generates a valid probability distribution. Let $P(x)$ be the output of the network for the input x. We define a differentiable loss function that measures the similarity between the ground truth and prediction distributions. The Bhattacharyya coefficient (BC) has the capability of measuring overlap between two probability distributions. BC is zero if there is no overlap and increases to a maximum of unity as the overlap increases. Based this knowledge, we define the loss function per input as following:

$$L(\{x, y\}; W) = -2BC(y, P(x)) \tag{2}$$

$$BC(y, P(x)) = \sum_{i \epsilon \Omega_p} \sqrt{y_i P_i(x)} \tag{3}$$

where Ω_p represents the pixel space. Eqn. 2 is easily differentiable with respect to the input of the loss layer, $P(x)$. The pixel-wise derivative of Eqn. 2 with respect to $P(x)$ is used for the back propagation of the loss during training.

$$\frac{\partial}{\partial P_i(x)} L_i(\{x, y\}; W) = -\sqrt{\frac{y_i}{P_i(x)}} \tag{4}$$

3.3 Post-processing

The network is trained on 73,620 image patches generated from the training images. At the test time, given a test image and corresponding manually clicked vertebrae centers, we create test patches following the same procedure described in Sec. 2. Each patch is then resized to 64×64 pixel and passed forward through the trained network which generates a patch level spatial probability distribution. These probability distributions often have noise and residual probabilities in the background. The residual probabilities are a result of the combined effects of the padding operations of the convolutional layers of the network and the introduced spatial normalization layer. Throughout the network, we have used zero padding in the convolutional layers to keep the output size similar to the input. This zero padding results in a lower value in the border of the output. As our spatial normalization layer simply forces the minimum to be zero, the border area of the final activation becomes zero and rest of the background assume small residual values. The effect can be seen in Fig. 4c, where the patch borders are visible and have probability values near zero. The range of values for the residual probability in each patch can be found by analysing its histogram. In the next step, we remove these residual probabilities from the background and re-normalize the distributions to have a range between 0 and 1. These patch level predictions are then resized to their original size and transformed back on the original image space using their known location, orientation and size. These values are known from the patch generation process.

The resultant distribution for each vertebra is then weighted by a prior corner distribution for that vertebrae learned from the training examples. Finally, on the original image space, the vertebral corners are localized by finding the maximum in each of four quadrants of each vertebra. The quadrants are defined using the manually clicked center points. The process is similar to the Harris based naive Bayes corner detector in the state-of-the-art work on cervical vertebrae corner detection [1]. In case the algorithm does not find any probability distribution for a corner, which may be a result of occlusion, surgical implant and/or low contrast, it uses this prior distribution of corners determine a possible corner location. In the example of Fig. 4e, we show that the bottom-left corner is missing on the original image space because of very low contrast. The complete process of corner localization starting from a test image including the post-processing steps is summarized in Fig. 4.

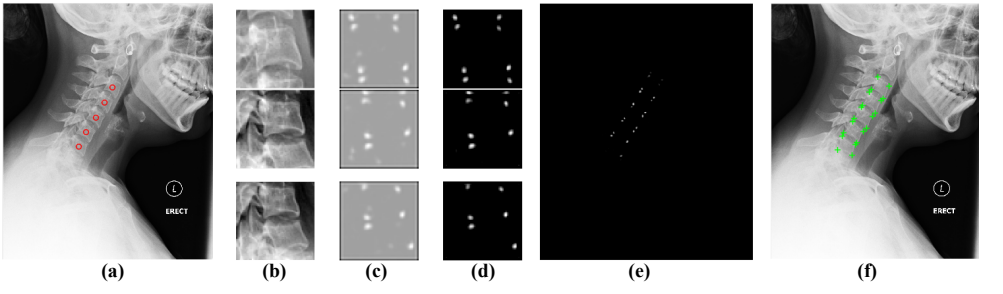

Figure 4: Post-processing (a) Input image with manually clicked vertebrae centers (b) Extracted image patches to be forwarded through the network (c) Patch level prediction results from the network (d) Patch level predictions after removing residual probabilities (e) Image level prediction (f) Localized corners.

4 Results and Discussion

We first evaluate the performance of the trained network at the patch level by reporting Bhattacharyya coefficient (BC) between each predicted spatial probability map and its corresponding ground truth probability for the 90,480 image patches generated from our 172 test images. The BC between two probability distribution is defined in Eqn. 3. An average BC of 0.9794 has been achieved over the test patches. A Bhattacharyya coefficient of 1 indicates a perfect match between two probability distributions. The histogram plot of the BC metrics is shown in Fig. 5a. It can be noted that the BC is always in the high range of 0.96 to 0.99 for all the test patches. However, the BC has limitations in measuring the similarity between two distributions. Since the majority of pixels on the ground truth probability distribution have zero values, it doesn't penalize if a small prediction probability is present in those places thus BC can be high even if the prediction looks different. This is why BC stays high even when there are border effects and small residual probabilities in the background (Fig. 4c). Although these results look different from the patch-level ground truth (Fig. 1b), the BC between them can be high as long as the locations of the maximum probabilities match. As our loss function is based on this metric, the trained network failed to remove the residual probabilities in the background (Fig. 4c). However, despite this limitation, the network robustly learns to predict high probabilities at the corner locations. After the post-

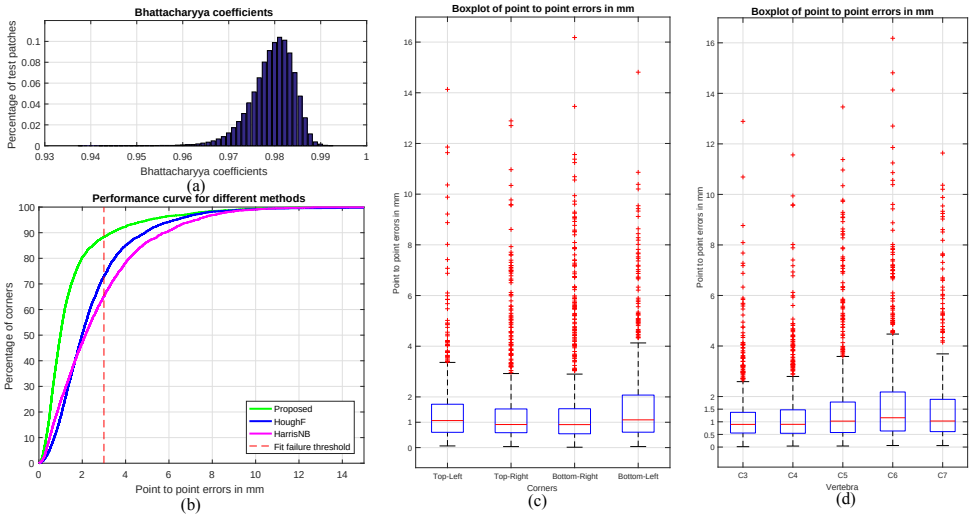

Figure 5: Result analysis (a) Histogram plot of Bhattacharyya coefficients for patch-level predictions (b) Cumulative error curve for different methods (c) Boxplot of errors for different corners (d) Boxplot of errors for different vertebrae.

processing phase, the corners are localized on the original image. The ground truth corners and the vertebral boundary curves are known from the manual annotations. We report two metrics: 1. P2P (point to point): Euclidean distance between the predicted corner to manually annotated corner in millimeters (mm) and 2. P2C (point to curve): distance between the predicted corner and splined vertebral boundary (green lines in Fig. 1a). The second metric is more appropriate when the corner area is smooth and determining a corner depends on human interpretation. We also report a third metric called fit failure [5]. We define fit failure as the percentage of corners with a P2P error greater than 3 mm. The median, mean and standard deviation of these metrics over the 3,188 corners of the test dataset are reported in Table 1. In order to compare the performance of the proposed corner detection framework, two methods from the state-of-the-art have been trained and tested on our dataset: 1. HarrisNB: Harris based naive Bayes corner detector with rectangular ROI and 2. HoughF: Hough forest-based method with their best-performing feature (Haar-Mixed) and prediction type (KC+KDE1). In terms of P2P error, the HoughF performs better than HarrisNB. Our proposed method achieved a 38% relative improvement in terms of the mean error compared to the best performing state-of-the-art method, HoughF. The median error for the proposed

Table 1: Euclidean distance between predicted and manually annotated corners.

	Point to point (P2P) (mm)					Point to curve (P2C) (mm)		
	Median	Mean	Std	Fit failure (%)		Median	Mean	Std
HarrisNB	2.15	2.70	2.20	34.91		0.53	0.95	1.10
HoughF	1.99	2.48	1.98	27.13		0.88	1.12	1.07
Proposed	**0.99**	**1.54**	**1.74**	**11.70**		**0.35**	**0.58**	**0.76**

method achieved a large drop of 1 mm from the HoughF method. The number of vertebrae with fit failure also decreased by more than 15%. The cumulative P2P errors for the compared methods are shown in Fig. 5b. It can be seen our proposed method outperforms the state-of-the-art methods by a large margin. In terms of P2C error, the HarrisNB outperforms the HoughF method. We believe this is because of the edge detection process utilized in the HarrisNB method, which forces the detected corners to be near vertebral boundaries. But our proposed method still outperforms the HarrisNB method with a relative improvement of 39% in terms of the mean error. However, it can be noted that the standard deviation of the proposed method is still somewhat high. This is because of the complexity in our test dataset. As we mentioned our data is not collected under a controlled environment, thus it contains challenging images full of clinical conditions, bone implants, image artefacts and contrast variations. Some of these challenging cases are shown in Fig. 6c,d. The boxplots of Fig. 5c,d also reveal that there are many outliers, most of which belong to the corners from these challenging cases. In Fig. 5c, we show a boxplot of the P2P errors for different corners. It can be noted corners on the right (or anterior side) have comparatively lower errors than the left side (posterior). This is due to the fact that anterior side of the cervical spine often has better image contrast than the posterior side which contains posterior spinal arches and processes. The vertebrae corners are also closer in between two vertebrae on the posterior side. The boxplot of corners for different vertebrae reveals that C3 and C4 have a lower error from the rest of the spine. As we go down the spine (from C3 to C7) the variation of the vertebrae increases as well as the image quality and contrast decrease to some extent, making it harder for the algorithm to predict corners. Some vertebrae level results for all the compared methods are shown in Fig. 6. In the first row, Fig. 6a, we show some relatively

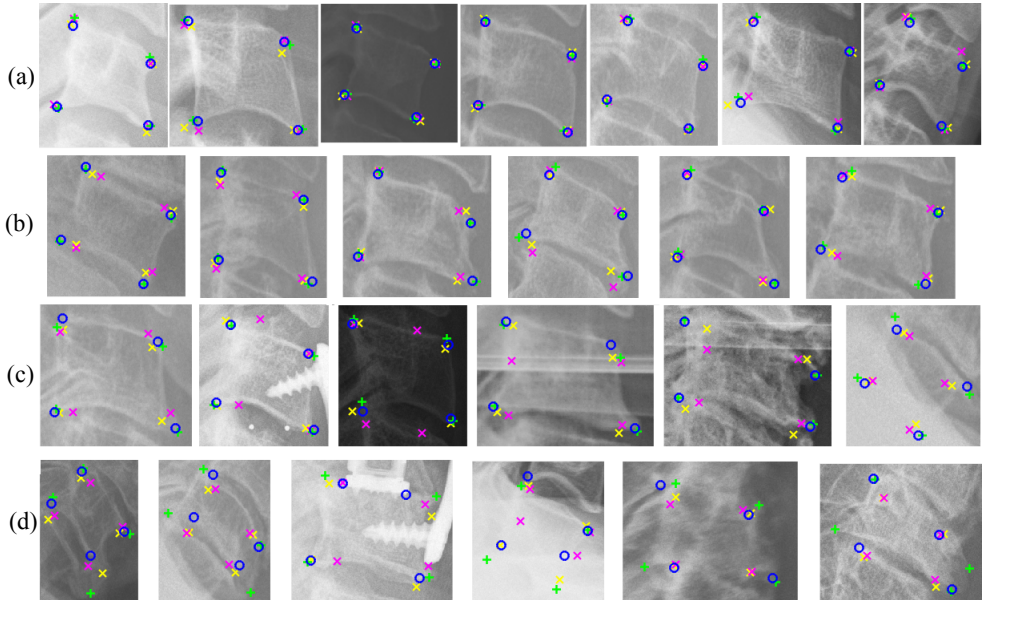

Figure 6: Vertebra level corner predictions: ground truth (+), proposed prediction (o), HarrisNB (x) and HoughF (x).

easy cases where predictions of all the methods are comparatively good. In Fig. 6b, we show some more easy cases, where our proposed method outperformed the state-of-the-art methods. Some challenging cases with bone implants, low contrast, image artefacts and clinical condition are shown in Fig. 6c, where the proposed method has outperformed the state-of-the-art methods. Finally, in Fig. 6d we show some more challenging cases where most of the methods including the proposed method have failed.

A few qualitative results with the full cervical spine with the predictions from the proposed deep probabilistic spatial regressor based corner localization framework are shown Fig. 7. Fig. 7a,b show two examples of healthy spines where the prediction results are near perfect for almost all the corners. A severe case of bone loss, osteoporosis and low image contrast is shown in Fig. 7c. It can be seen even with such severe conditions, the prediction results are considerably correct. Fig. 7d shows an example with surgical bone implants, which affected some of the prediction results, especially at the C5-C6 area. However, because of the patch based framework, other corners are well detected. A few results for vertebral misalignment (spondylolisthesis) are reported in the rest of the Fig. 7. Fig. 7e shows misalignment between C4-C5, Fig. 7f C3-C4 and Fig. 7h C5-C6. The predicted corners can be used to determine these misalignments automatically.

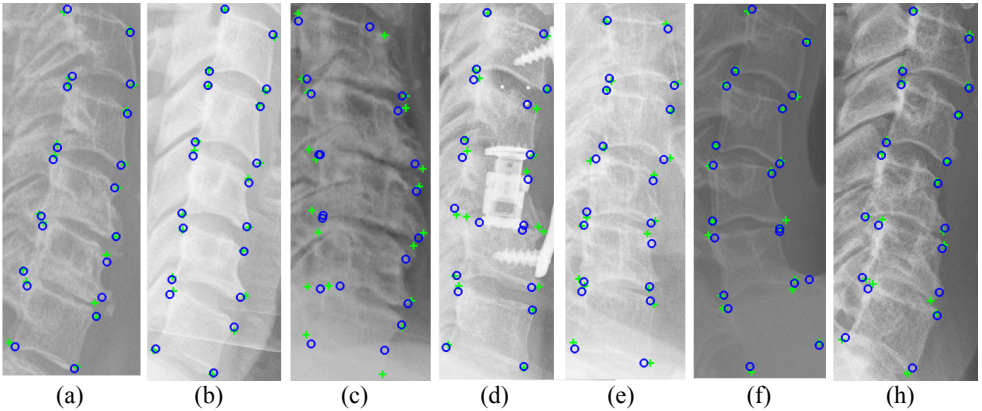

Figure 7: Vertebral corner prediction using fully convolutional network based probabilistic spatial regressor : ground truth (+), proposed prediction (o).

5 Conclusion

Classification and segmentation output of the convolutional neural networks are probabilistic but regression is often deterministic. In this work, we have introduced a novel fully convolutional network capable of predicting probabilistic output over the image space for image landmark localization. In the process, we have introduced a new spatial normalization layer and a novel Bhattacharyya coefficient based loss function. The proposed network has been adapted in a semi-automatic vertebral corner localization framework and evaluated on challenging dataset of 172 real life emergency room cervical X-ray images. The proposed method has outperformed the previous state-of-the-art by a large margin. However, the are still several limitations to overcome. The simple normalization layer can be further improved

to resolve the border effect coming from the convolutional layers. The loss function can be further modified so that it also penalizes the residual background probabilities. This work is a part of our overarching goal of building a fully automatic injury detection system for lateral cervical X-images for the emergency room physicians. To this end, we have been able to produce a semi-automatic system for localization of vertebral corners. These corners can be used to detect spinal misalignment or spondylolisthesis of the vertebrae. In future work, we plan to use these corners to initialize active shape models and/or level-set models to achieve segmentation of the vertebrae.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] S M Masudur Rahman Al-Arif, Muhammad Asad, Michael Gundry, Karen Knapp, and Greg Slabaugh. Patch-based corner detection for cervical vertebrae in x-ray images. *Signal Processing: Image Communication*, 2017.

[2] S M Masudur Rahman Al-Arif, Michael Gundry, Karen Knapp, and Greg Slabaugh. Improving an active shape model with random classification forest for segmentation of cervical vertebrae. In *Computational Methods and Clinical Applications for Spine Imaging: 4th International Workshop and Challenge, CSI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers*, volume 10182, page 3. Springer, 2017.

[3] Mohammed Benjelloun, Saïd Mahmoudi, and Fabian Lecron. A framework of vertebra segmentation using the active shape model-based approach. *Journal of Biomedical Imaging*, 2011:9, 2011.

[4] Aïcha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 460–468. Springer, 2016.

[5] PA Bromiley, JE Adams, and TF Cootes. Localization of vertebrae on dxa vfa images using constrained local models with random forest regression voting. *Journal of Orthopaedic Translation*, 4(2):227–228, 2014.

[6] Hao Chen, Xiaojuan Qi, Jie-Zhi Cheng, and Pheng-Ann Heng. Deep contextual networks for neuronal structure segmentation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1167–1173. AAAI Press, 2016.

[7] Timothy F Cootes. Fully automatic localisation of vertebrae in ct images using random forest regression voting. In *Computational Methods and Clinical Applications for Spine Imaging: 4th International Workshop and Challenge, CSI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers*, volume 10182, page 51. Springer, 2017.

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Juying Huang, Fengzeng Jian, Hao Wu, and Haiyun Li. An improved level set method for vertebra ct image segmentation. *Biomedical Engineering Online*, 12(1):48, 2013.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[12] Sidi Ahmed Mahmoudi, Fabian Lecron, Pierre Manneback, Mohammed Benjelloun, and Säid Mahmoudi. GPU-based segmentation of cervical vertebra in X-ray images. In *Cluster Computing Workshops and Posters (CLUSTER WORKSHOPS), 2010 IEEE International Conference on*, pages 1–8. IEEE, 2010.

[13] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2992–2992, 2015.

[14] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[15] Deepak Pathak, Philipp Krähenbühl, Stella X Yu, and Trevor Darrell. Constrained structured regression with convolutional neural networks. *arXiv preprint arXiv:1511.07497*, 2015.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[17] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR), 2015*. http://arxiv.org/abs/1409.1556.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[20] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.