A Novel Framework for 3D-2D Vertebra Matching

Hanchao Yu[1], Yang Fu[1], Haichao Yu[1], Jianbo Jiao[1],
Yunchao Wei[1], Xinchao Wang[2], Bihan Wen[3], Zhangyang Wang[4],
Matthew Bramlet[5], Thenkurussi Kesavadas[6], Honghui Shi[7,1], Thomas Huang[1]
[1]IFP, Beckman Institute, UIUC, [2]Stevens Institute of Technology, [3]NTU, [4]CSE, TAMU,
[5]University of Illinois College of Medicine Peoria, [6]HCESC, UIUC, [7]IBM Research

Abstract

3D-2D medical image matching is a crucial task in image-guided surgery, image-guided radiation therapy and minimally invasive surgery. The task relies on identifying the correspondence between a 2D reference image and the 2D projection of the 3D target image. In this paper, we propose a novel image matching framework between 3D CT projection and 2D X-ray image, tailored for vertebra images. The main idea is to learn a vertebra detector by means of the deep neural network. The detected vertebra is represented by a bounding box in the 3D CT projection. Next, the bounding box annotated by the doctor on the X-ray image is matched to the corresponding box in the 3D projection. We evaluate our proposed method on our own-collected 3D-2D registration dataset. The experimental results show that our framework outperforms the state-of-the-art neural network-based keypoint matching methods.

1. Introduction

Computer vision and Multimedia technologies are making significant impact on medical imaging fields. In this work, our target is to adopt the state-of-the-art object detection techniques to address one of the inportant medical imaging issues, i.e., 3D-2D registration. 3D-2D registration is pivotal for the image-guided surgery, image-guided radiation therapy (IGRT) and other image-guided medical tasks [19]. During an image-guided surgery, doctors need to compare the images taken before the surgery with the one acquired during the surgery [7]. The image taken before the surgery, such as CT and MRI, is often of good quality. During the surgery, extra images such as X-ray will be taken. Since the images need to be compared with what are taken at a different time and by different devices, registration between them is necessary.

The most common strategy for 3D-2D registration is to get the 2D projection of 3D images first. Then the problem is formulated as 2D image registration [19]. To find the best projection, some prior knowledge about the relationship between coordinate systems of different imaging devices is needed. The main challenges of 3D-2D registration are: (1) in some cases, prior knowledge of projection parameters is unavailable and projection parameter estimation itself is a difficult problem. (2) There are some artificial implanted items present in images taken during and after the surgery, which does not exist in the pre-interventional 3D image. Current mutual information based registration approaches cannot compute a global mapping from one image to another without initial pose estimation. Motivated by the current advancements of deep learning techniques in computer vision, we alternatively provide a more promising framework to tackle such a challenging issue. Concretely, we propose a detection-based, end-to-end multimodal 3D CT-X-ray vertebra matching system under the following two assumptions: (1) the best 3D-2D projection parameter is given. (2) a bounding box which contains a vertebra from X-ray is given by a doctor who wants to know the corresponding region in CT projection.

The first step of our proposed method is building a deep detector for vertebra detection in 3D CT projection image. Recently, deep-learning based object detection method has achieved great success in medical image analysis [16]. We choose the state-of-the-art faster region proposal network (Faster-RCNN) [21] as our detection framework. To the best of our knowledge, this is the first attempt to use deep detector in 3D/2D matching task. Although the size of our dataset is limited, the object patterns are also limited, which enables us to train a good detector. With vertebra localized in CT projection image, the next step would be find a matching between the bounding box given by doctor and one of these vertebra bounding boxes provided by the detector. Since vertebra can be considered as rigid, we use Generalized Hough Transform (GHT) [1], which is widely used in detecting arbitrary shapes given a good binary template. The idea is detecting the edge of the vertebra from X-ray image, which would be the template for matching and performing Generalized Hough Transform to store the

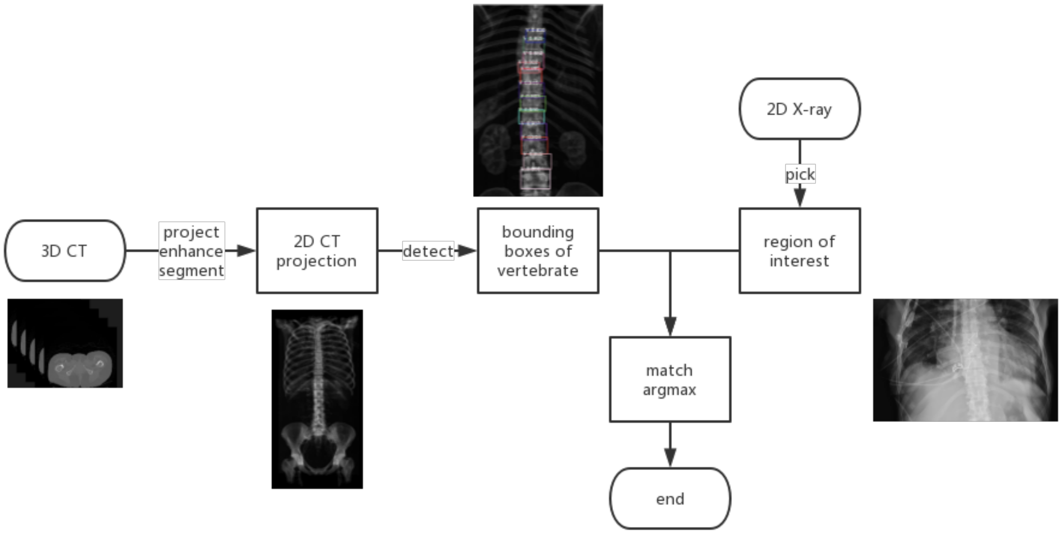


Figure 1. The framework of this work. The input is a pre-operation 3D CT volume. With projection and enhancement the 3D CT volume is transformed into a 2D CT projection image. Deep neural network based vertebra detection is performed on this 2D projection. On the other side, the doctor will pick a region of interest from the post-operation 2D X-ray image that contains one vertebra. Generalized Hough based matching is performed betweens detected candidate regions from 2D CT projection and the region of interset from 2D X-ray image.

shape information in a R-table. Then we can compare the R-table with every possible bounding box and find the best match. Here the position information is used for reducing the search range. Since images from multi-modality might vary in scale and rotation, we implement a modified GHT algorithm with scale transformation, rotation and translation. To the best of our knowledge, this is the first work to formulate 3D-2D registration as a region-to-region matching problem.

The main contributions of our method are as follows: (1) We propose a novel framework for end-to-end multimodal vertebra matching. (2) We introduce deep learning based detector to improve the performance of matching. (3) We propose a modified Digitally Reconstructed Radiography (DRR) generation algorithm with data augmentation for better detection. (4) We introduce Generalized Hough Transform for the multimodal image matching task.

The rest of this paper is organized as follows. Related work is reviewed in Section 2, including 3D-2D image registration, deep-learning based object detection and generalized Hough Transform. In Section 3 there is detailed description of our proposed method, including modified DRR generation, Faster-RCNN detection with data augmentation and GHT matching. Experiments and results are presented in Section 4. Section 5 gives the conclusion and discusses the future work.

2 Related Work

2.1 3D-2D Image Registration

The main goal of 3D-2D image registration is to find the correspondences between the 3D images and 2D ones. Image registration finds its crucial applications in various computer vision tasks include low-levels ones like image rectification [34] and super-resolution [17], as well as higher-level ones like detection [29, 32] and tracking [18, 31, 33, 11]. Since 3D and 2D data differs in the dimension , dimensional correspondence should be built before the alignment process, i.e., the data to be registered should have the same dimension. Obviously, there are two directions, from 3D data to 2D or from 2D to 3D. Three strategies have been proposed to achieve this dimensional correspondence, i.e., projection, back-projection and reconstruction [19].

By projection, a series of 2D images are produced with different projection parameters, and the problem is now a 2D-2D image registration problem. Projection parameters can be determined once the best 2D match pairs are find. The projected 2D images are called DRR, which will be discussed in the following subsection. By back-projection, imaginary virtual ray is projected to the ray source using a back-projection matrix, and the comparison is in

3D space. By reconstruction method, multiple 2D intra-interventional images are used to reconstruct the 3D object. Basically, three main registration methods have been explored in 3D-2D registration: feature-based, intensity-based and gradient-based.

2.2 Digitally Reconstructed Radiography generation

Digitally Reconstructed Radiography, or DRR, has been studied for decades. It is the X-ray like image projected from the 3D CT data, which is widely used in projection strategy of 3D CT and 2D X-ray registration. One fast ray-tracing algorithm is proposed by Siddon in 1985 [26]. Instead of voxel-wisely computing intersections between each ray and voxels, this algorithm considers voxel planes as equally spaced and computes the intersections incrementally, which yielded a significant speedup. Following Siddon's work, some improved versions of the ray-tracking algorithms are proposed [10] [8]. These algorithms improved Siddon's method by avoiding unnecessary array index calculation. In this paper, we employ a similar ray-tracing method to DRR calculation.

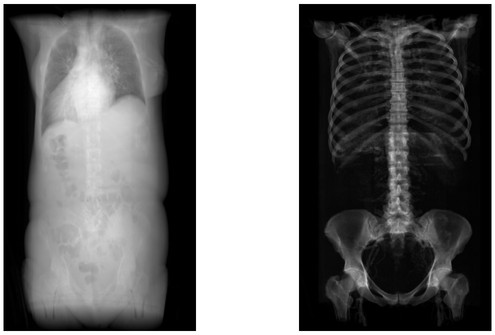
2.3 Deep Network for Object Detection

With the development of modern deep convolution neural network, some object detectors [6, 21, 4, 5, 30, 14, 15, 9] show dramatic improvements in accuracy compared with early methods based on hand-engineered features. The R-CNN method adopted selective search to obtained object region proposals[28] and trained CNNs to classify the proposal regions into object categories or background. And Faster-RCNN utilized Region Proposal Network instead of selective search to generate object region proposals faster and more accurately. The whole architecture can be trained end-to-end.

There are some attempts on vertebra localization with deep neural network[25][13]. However, there is no attempt to formalize it as a object detection problem.

2.4 Generalized Hough Transform

Generalized Hough Transform (GHT) is a useful method for template matching[1]. The main idea is that given a template image, the gradients of the edge map can be found and saved in a R-table. When matching, every possible position in the image is evaluated using the R-table. For every possible position, a matching score will be computed by voting. The more points fall in the shape, higher the matching score will be. However, the original GHT cannot deal with the variants in scale and rotation but only find the shape in the image which is strictly identical to the template. Besides, previous methods[27][12] only use GHT for general vertebra detection and matching. In those methods, the vertebra template comes from a modal which is an average of many patients.



(a) raw DRR (b) remove heart

Figure 2. Enhancement of DRR image. Figure(a) shows the raw DRR image that comes from direct projection of 3D image. Figure(b) shows the enhanced DRR image with slice-level enhancement

3 Proposed Methods

The pipeline of our framework is shown in figure 1. The input is a pre-operation 3D CT volume. With projection and enhancement the 3D CT volume is transformed into a 2D CT projection image. Deep neural network based vertebra detection is performed on this 2D projection. On the other side, the doctor will pick a region of interest from the post-operation 2D X-ray image that contains one vertebra. Generalized Hough based matching is performed between detected candidate regions from 2D CT projection and the region of interest from 2D X-ray image. The following parts of this section will introduce the 3 major parts: Enhancement and generation of DRR image, vertebra Detection in 3D CT Projection, region to region matching with Generalized Hough Transform.

3.1 Enhanced DRR Generation

The enhancement includes semi-automatic segmentation of the heart and histogram adjustment. Removal of heart is necessary since vertebra, the target to be detected, is partially blocked by heart. We project the CT volume in the direction orthogonal to sagittal plane. In the resulted 2D projection image, we draw a curve to separate the heart and spine. According to the 2D curve in sagittal plane, heart can be removed in CT volume.

To generate high-quality DRR for registration, we enhance CT slices by histogram equalization before DRR generation. First, CT volume values are clipped such that HU values are in the range [-1024, 500]. Then piecewise histogram equalization is employed to enhance each CT slice. Since most vertebra voxels are in a particular range of grey levels, we enhance vertebra regions by mapping the range to a wider one. In our experiments, we map [80, 300] to [-800, 400]. Histogram equalization can be described as a

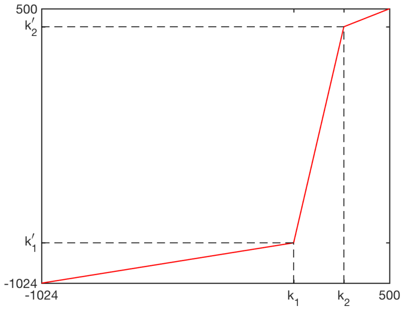


Figure 3. Linear mapping function of piecewise histogram equalization. Unit: HU

mapping function as shown in Figure 3. In our experiments, $k_1 = 80, k_2 = 300, k_1' = -800, k_2' = 400$ (Unit: HU).

To calculate DRR of a CT image, we make use of a fast ray-tracing algorithm implemented in Plastimatch[24]. We simulate the chest X-ray imaging process and generate DRRs in the directions of the axial plane. The projection angles are equally spaced by 6 degrees. Thus, a total number of 60 simulated chest X-ray images are created. Figure 2 shows the importance of enhancement.

3.2 Vertebra Detection in 3D CT Projection

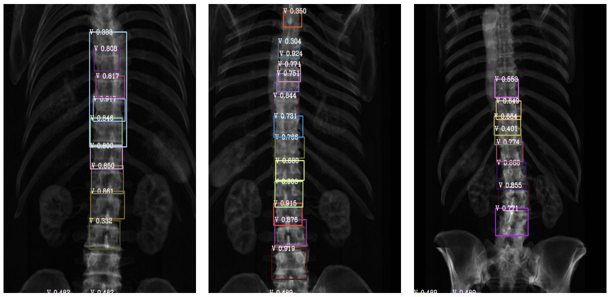
We adapt the state-of-the-art faster-rcnn object detection model to detect the vertebra in our task. Since our dataset is relatively small, we perform the data augmentation to make the model more robust and prevent overfit. Then a modified contextual faster-rcnn is used to detect the vertebra in the image.

3.2.1 Data Augmentation

Since our dataset is small, the data augmentation is necessary. Generally, data augmentation contains image rotation, translation, center-crop, etc. It is widely used in many deep learning based medical image analysis tasks[22][23]. In our task, there is some noise after the semi-auto segmentation of the CT projection and we sample random Gamma transform and spatial Gaussian-distributed noise, then we add the sampled transform and noise to image. With these data augmentation methods, we can generate any number of training samples.

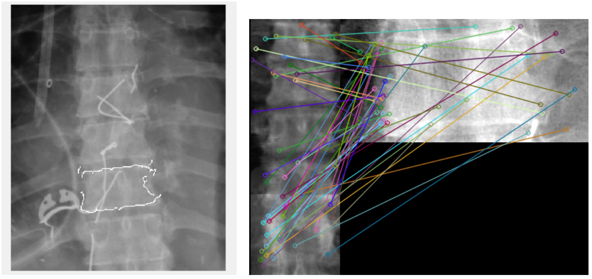
3.2.2 Contextual Faster-RCNN

In order to achieve the region-based matching, we need to detect each vertebra from the original image firstly. Since each vertebra looks very similar, it's difficult to just use a small bounding box as a input to train a detector. We proposed the contextual Faster-RCNN introduced by [3]. Specifically, we used two branch network based on Faster-RCNN: the first branch is the normal Faster-RCNN, and for the other branch, we enlarged the object region proposal obtained by first branch and used this as new region proposal to do the bounding box regression and classification. By doing this, the model will receive more background information, like costae and achieve better detection results.



(a) (b) (c)

Figure 4. Example results of detection. Each bounding box is a candidate region to be matched with 2D X-ray region



(a) Proposed methods (b) keypoint match results

Figure 5. Comparison of matching results. Figure(a) shows the overlapped(matched) images from proposed method. Figure(b) shows the results of keypoint match method

3.3 Region to Region Matching

After the detection process, now we have 1 bounding box that contains at least 1 complete vertebra given by doctor and several candidates bounding boxes given by the detector. This is a 1 vs N matching problem. Since the number of candidates is very limited, we choose brute force match strategy. For every candidate vertebra, looping over a range of scale and a range of rotation degree, find the rotation degree and scale factor that maximize the matching score in GHT. Then find the best candidates by compare matching score between different candidates.

	direct GHT	Proposed
Matching accuracy	0.865	0.912

Table 1. compare of direct GHT and detect-GHT. Direct GHT is matching two images directly.

4 Experiments and Results

4.1 Dataset

We use our 3D-2D matching data for training and testing. Our dataset is about image-guided heart surgery, in which the main focus is the area near the heart. Therefore, the rigid vertebra near the heart is a good landmark. Raw data is a private and cleaned dataset collected from our partner hospital. The dataset contains 12 3D CT scans with around 300 slices each. The resolution of each CT slice is 512×512. The ground truth annotation comes from experienced doctors. For every 3D CT data, we have corresponding 2D X-ray images available. Though the dataset is private, we are considering refine the dataset and releasing it to public.

4.2 Training of deep neural nets based detector

We train our model on mxnet [2], K40 GPU for 200 epochs. The performance is MAP 0.9090@0.5. The model we use is Contextual Faster-RCNN which is introduced in the related work. Mean Average Precision(mAP) is the most popular metric in the object detection. Some examples of detection results are shown in figure 4.

4.3 Results and Evaluations

In 3D-2D registration area, to the best of our knowledge, there is still no public dataset, so we mainly compare the results with our own dataset. We use 2 baselines, one is the state-of-the-art deep learning based keypoint match method [20] and another is directly applying the GHT without detection.

The best metric for 3D-2D registration is Target Registration Error (TRE)[19], which is used to compare difference between the error between computed transform and ground truth transform. In our dataset, ground truth transform is unavailable, so we simply use matching accuracy as our metric. Matching accuracy is the ratio between correct matched examples and the number of all examples.

Surprisingly, the state-of-art key points method does not work here, there are rarely matched keypoints, which is shown in figure 5. We also tried other methods popular in 3D-2D matching such as mutual information, which did not produce good results. The main reason is probably that the illumination changes greatly between these multimodal images and both keypoint-based and intensity-based methods are not rubost to such changes. However, since the vertebra can be viewed as rigid, the shape of the vertebra does not change much in different modalities. Our proposed method is faster and more accurate than the direct GHT, since the match is restricted to limited numbers of bounding boxes. Table 1 shows the results of direct GHT and GHT with detection. The search area is reduced and the result is more robust to noise.

5 Conclusion and Future Work

In this paper, we propose a new framework for 3D CT/2D X-ray image matching. We introduce the deep learning based detection methods to detect the vertebra. Combined with the Generalized Hough Transform, we can reduce the computation time and improve the matching accuracy. The state-of-art keypoint based matching methods seems does not work in our task. Compared with the direct GHT method, our method is faster and more accurate.

Possible future work includes: (1) Combine the projection parameter searching to make a end to end system. (2) Using deep learning to predict the scale and rotation parameter.

Acknowledgement We thank the generous support from JUMP ARCHES that enabled this work.

References

[1] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.

[2] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[3] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015.

[4] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Decoupled classification refinement: Hard false positive suppression for object detection. *arXiv preprint arXiv:1810.04002*, 2018.

[5] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] W. E. L. Grimson, R. Kikinis, F. A. Jolesz, and P. M. Black. Image-guided surgery. *Scientific American*, 280(6):62–69, 1999.

[8] G. Han, Z. Liang, and J. You. A fast ray-tracing technique for tct and ect studies. In *Nuclear Science Symposium, 1999. Conference Record. 1999 IEEE*, volume 3, pages 1515–1518. IEEE, 1999.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988, 2017.

[10] F. Jacobs, E. Sundermann, B. De Sutter, M. Christiaens, and I. Lemahieu. A fast algorithm to calculate the exact radiological path through a pixel or voxel space. *Journal of computing and information technology*, 6(1):89–94, 1998.

[11] L. Lan, X. Wang, S. Zhang, D. Tao, W. Gao, and T. Huang. Interacting Tracklets for Multi-object Tracking. *IEEE Transactions on Image Processing*, 27:4585–4597, 2018.

[12] M. A. Larhmam, M. Benjelloun, and S. Mahmoudi. Vertebra identification using template matching modelmp and k -means clustering. *International journal of computer assisted radiology and surgery*, 9(2):177–187, 2014.

[13] F. Lecron, M. Benjelloun, and S. Mahmoudi. Fully automatic vertebra detection in x-ray images based on multi-class svm. In *Medical Imaging 2012: Image Processing*, volume 8314, page 83142D. International Society for Optics and Photonics, 2012.

[14] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *IEEE CVPR*, 2017.

[15] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2017.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[17] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, and T. Huang. Learning Temporal Dynamics for Video Super-Resolution: A Deep Learning Approach. *IEEE Transactions on Image Processing*, 27:3432–3445, 2018.

[18] A. Maksai, X. Wang, F. Fleuret, and P. Fua. Globally Consistent Multi-People Tracking with Motion Patterns. In *International Conference on Computer Vision* , 2017.

[19] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš. A review of 3d/2d registration methods for image-guided interventions. *Medical image analysis*, 16(3):642–661, 2012.

[20] D. Mishkin, F. Radenovic, and J. Matas. Learning discriminative affine regions via discriminability. *arXiv preprint arXiv:1711.06704*, 2017.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] 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*, pages 234–241. Springer, 2015.

[23] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015.

[24] G. C. Sharp, R. Li, J. Wolfgang, G. Chen, M. Peroni, M. F. Spadea, S. Mori, J. Zhang, J. Shackleford, and N. Kandasamy. Plastimatch-an open source software suite for radiotherapy image processing. In *Proceedings of the XVIth International Conference on the use of Computers in Radiotherapy (ICCR), Amsterdam, Netherlands*, 2010.

[25] W. Shen, F. Yang, W. Mu, C. Yang, X. Yang, and J. Tian. Automatic localization of vertebrae based on convolutional neural networks. In *Medical Imaging 2015: Image Processing*, volume 9413, page 94132E. International Society for Optics and Photonics, 2015.

[26] R. L. Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12(2):252–255, 1985.

[27] 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*, pages 224–228. IEEE, 2002.

[28] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[29] F. Wang, L. Zhao, X. Li, X. Wang, and D. Tao. Geometry-Aware Scene Text Detection with Instance Transformation Network. In *Computer Vision and Pattern Recognition* , June 2018.

[30] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *The European Conference on Computer Vision (ECCV)*, pages 434–450, 2018.

[31] X. Wang, B. Fan, S. Chang, Z. Wang, X. Liu, D. Tao, and T. Huang. Greedy Batch-based Minimum-cost Flows for Tracking Multiple Objects. *IEEE Transactions on Image Processing*, 26:4765–4776, 2017.

[32] X. Wang, Z. Li, and D. Tao. Subspaces indexing model on Grassmann manifold for image search. *IEEE Transactions on Image Processing*, 20:2627–2635, 2011.

[33] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking Interacting Objects Using Intertwined Flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:2312–2326, 2016.

[34] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao. FishEyeRecNet: A Multi-Context Collaborative Deep Network for Fisheye Image Rectifica- tion. In *European Conference on Computer Vision*, September 2018.