Bones detection in the pelvic area on the basis of YOLO neural network

Zuzanna Krawczyk, Jacek Starzyński
Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland, Email: zuzanna.krawczyk@ee.pw.edu.pl

***Abstract*—The article focuses on application of You Only Look Once (YOLO) neural network in the process of individual bones recognition, in the set of CT slices of human pelvic area. The network was trained on custom data and then it was employed to detect bone structures in similar yet different CT data. The accuracy of the detection was investigated. In the final step, the bounding boxes of detected regions, created by the YOLO algorithm were employed to position the idealized models of bones with respect to CT data.**

***Keywords*—object detection; YOLO; neural network; CT.**

I. Introduction

Creation of human-specific model of skeleton out of medical data, with individual bones distinguished, is a time-consuming, not fully automated task. Traditional approach involves segmentation of image data. State of art segmentation algorithms – histogram based, region growing based or those focused on edge detection [1], although they have good performance, they do not provide the information about the class of segmented object. Thus, after preforming simple segmentation step, recognized bones data structures are often incorrectly glued together, if they lie in close proximity to each other and it is not obvious which specific bone has been detected. That is why the additional step performing the object detection on CT data is needed prior the segmentation step.

In the paper we suggest the following approach based on neural network and use of idealized surface bones model:

1) Neural network is used to detect the bounding boxes surrounding individual objects and to assign them to the proper bone category.
2) Bounding boxes surrounding detected regions, created by the YOLO algorithm are employed to create the affine transformations between the idealized surface models of bones and the CT data.
3) The idealized bones model is scaled automatically to the size of patient data and moved to the correct position.
4) Idealized surface bones model is latter adjusted with non-affine transformation in order to create patient-specific model.

The paper focuses on the first three steps of the procedure – object detection and affine transformation of the model.

The YOLO [2] neural network was chosen as the object detector due to its very good time performance and low rate of false positive detection error in comparison to another state of the art convolutional detection network – Fast R-CNN [3] as well as for its good generalization abilities.

The paper is organized as follows: in the second chapter the idea of YOLO network is described, the third chapter focuses on description of the training data, the following chapter presents the detection results and evaluates their accuracy, fifth chapter describes the positioning procedure of idealized bones model. The last chapter summarizes the obtained results.

II. YOLO neural network

YOLO (You Only Look Once) is a convolution neural network based approach to object detection created and described by J. Redmon et al. in [2], which aims in the real-time object detection. In contrast to other approaches with complex pipelines, YOLO uses single convolutional network, which is trained on full images and then it simultaneously predicts multiple bounding boxes for different objects for all specified classes as well as probabilities combined with those boxes.

In the YOLO training procedure image is divided into grid of N by N cells. Each cell is responsible for prediction of B possible bounding boxes (rectangles which surround detected object). Bounding boxes are characterized by 5 values:

- x, y – coordinates of the center of the bounding box relative to the cell bounds,
- w, h – width and hight of bounding box relative to image dimensions,
- cs – confidence – value which determines how confident the network is about the presence of the object inside the bounding box and how accurate the shape of the box is, but it does not contain information about the class to which the object belongs.

If there is no object inside the cell cs value is set to 0 otherwise it is equal to the intersection over union (IoU) between the bounding box and the ground truth box (reference boundig box sourrounding object, created in train dataset by qualified person) . For each cell, for each class cp probability is computed: which is the probability of object to belong to given class if an object is present in the box. Class specific confidence score is achieved by multiplication of the cp probabilities by the bounding box confidence cs.

The architecture of the neural network consists of 24 convolutional layers and 2 fully connected layers. The convolutional layers extract features from the image whereas fully connected layers are responsible for probabilities and bounding boxes parameters prediction. Network weights values are optimized to minimize sum-squared error. The detailed architecture of the network and training process is described in [2].

III. Training data and network configuration

The training set consists of 320 monochromatic images created out of two CT data sets of pelvic area – male and female one. Images represent complete CT slices as well as regions covering the area of only one type of bone, with different lighting intensity. All of the images represent body sections in transverse plane. 10% of images was separated as the validation set, which has not been used in training process of neural network.

The bone structures on the training set images are assigned to two classes: femur and pelvis. Bone structures present on the images were surrounded with the bounding boxes by a qualified person. Training dataset includes also negative samples without desired bone structures.

The aim of the network training was for it to recognize bone classes specifically in the sets of CT data of the pelvis area. Patient which is undergoing a CT scan is usually placed in one certain position. The spacial relation between different bones is similar for different patients, although anatomical details can be different. On the basis of the forementioned reasons we made an assumption, that low number of input data will be sufficient in order to train the network correctly.

The version 2 of YOLO [4] network, implemented in Darknet framework [5] is used for training and detection. As a start point, set of convolutional weights trained on ImageNet dataset is used. The network setup is similar to the default one, with few exceptions: number of classes is set to 2 and number of filters in the final convolutional layer is equal 35. Sets of 64 images (called batches) are being used in every training step. Batch is divided into 16 smaller sets - where each set will be performed at once. (Such constraint is necessary in order not to exceed the graphic card memory during training step. The smallest possible subdivisions number for the graphic card architecture used by authors was set).

IV. Results and evaluation of accuracy of the detection step

The IoU (1), recall (2) and precision (3) measures can be used in order to asses the quality of bounding boxes detected by the trained network.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Intersection over Union defined as the area of overlap of two bounding boxes (A – ground truth box and B – detected one) divided by the area of union of both boxes indicates how similar two bounding boxes are. The measure in its current implementation evaluates only the correctness of the position and area of the detected region, but does not take into consideration class of the detected object.

The value of recall measure defined as the fraction of all true positive (TP) object detections to the sum of all objects which should be detected (sum of true (TP) and false negative (FN) detections) is the state of art measure of completeness of the detection. Whereas, precision measure describes the exactness of the method and is obtained by computing the ratio between the number of TP detections and sum of TP and FP (false positive) detections – which can be interpreted as the sum of all (good and wrong) detections performed by the network. We define a true positive (TP) detection as the one for which the IoU value between the detection and ground truth box is bigger, than given threshold (0.5) and for which the probability of belonging to the same class as the ground truth box is bigger then given value (0.5).

The YOLO network was trained for 20.000 epochs. In the last epoch the mean average error was equal to 0.232. In order to choose the network configuration with which the detection step is performed correctly, The IoU, recall and precision measures were evaluated both for training and validation data set, for network weights obtained in every hundredth epoch up to 1.000 epochs. Subsequent values were computed with a step of 1.000 epochs.

As depicted on Fig. 1 the IoU indicator has a value of approximately 90% from 10.000th iteration. The maximal recall and precision values are achieved after 6.000 iterations (Fig. 2, Fig. 3). Network weights from 10.000th training epoch were chosen to be used in further object detection procedure.

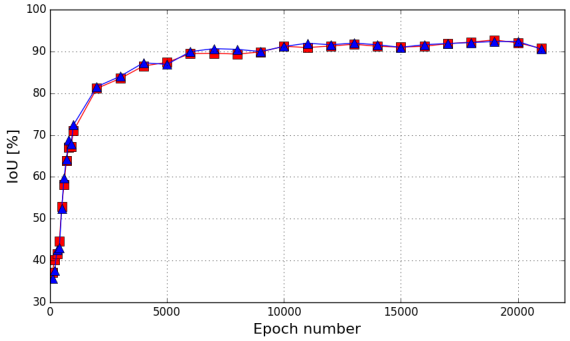

Fig. 1. The value of IoU measure with respect to the number of training epochs, calculated for training data set (red color) and validation data set (blue color)

The network was further tested on 4 various CT datasets. Two of them visualize female pelvis, whereas another two were taken of the male patients. Detailed information about data is presented in Table I. In all 4 data sets, the bone structures were categorized correctly, in most of the CT slices. In rare cases there were some inaccuracies in the shape of the bounding box. Few false positive detections (7 out of 684 total bounding boxes which were detected), in the area representing spine above the pelvis area, occurred. The problem can be probably solved by creation of a new class of data assigned to the spine bones, and taking it into consideration during learning process. Total number of 6 false negative detections

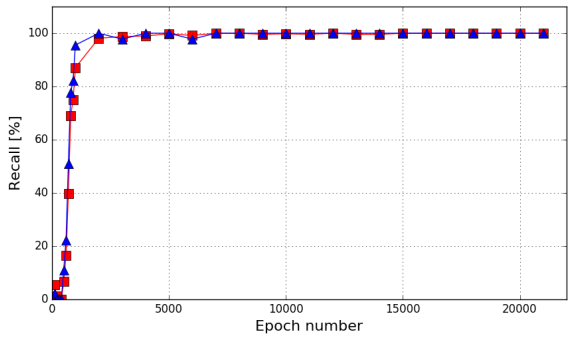

Fig. 2. The value of recall measure with respect to the number of training epochs calculated for training data set (red color) and validation data set (blue color).

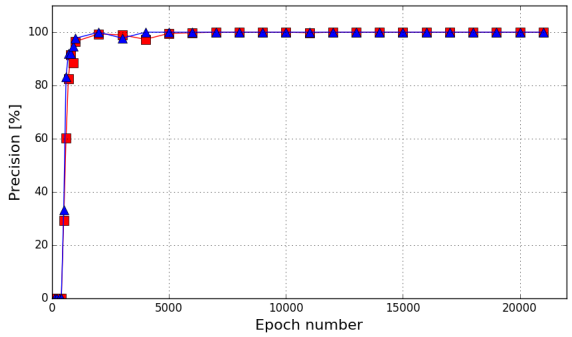

Fig. 3. The value of precision measure with respect to the number of training epochs calculated for training data set (red color) and validation data set (blue color).

occurred. Detailed statistical calculations regarding test data are presented in Table II. In case of both datasets representing female bones (sets no. 1 and no. 2.) the number of true negative (TN) detections is higher than in case of two another sets of CT. Larger area of spine structures above the pelvis was visualized in those sets in comparison to data sets number 3 and 4.

The total accuracy of pelvis and femur detection in data, calculated as shown in (4) was equal to 98%.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

TABLE I
SPECIFICATION OF TEST CT DATA SETS

Dataset	Patient sex	Slice thick-ness [mm]	Slice spacing [pixels]	Nr of slices
1	female	3.0	512x512	109
2	female	3.0	512x512	114
3	male	3.0	512x512	110
4	male	4.0	512x512	82

TABLE II
PRECISION, RECALL AND ACCURACY MEASURES FOR 4 TEST CT DATA SETS

Dataset	TP	FP	TN	FN	Precision	Recall	Accuracy
1	169	1	22	1	99%	99%	99%
2	151	4	30	1	97%	99%	97%
3	204	2	2	1	99%	99%	99%
4	153	0	0	3	100%	98%	98%
Cum.	677	7	54	6	99%	99%	98%

Please note that in calculation of precision, recall and accuracy measures for the test data the IoU value between detected bounding box and ground truth box was not taken into consideration. Detected bounding box was accepted as TP if detected bone structures were assigned to the appropriate class. Examples of detected bounding boxes are shown on Fig. 4.

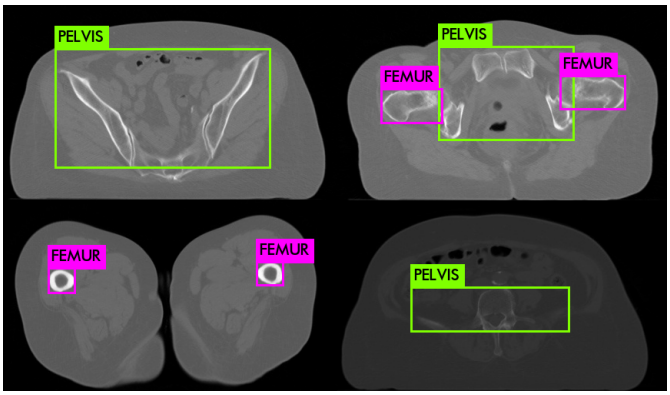

Fig. 4. The examples of detection of bones on the images from the test set: a) pelvis detected with 94% confidence,b) pelvis detected with 94% confidence both femur bones detected with 90% confidence, c) left femur bone detected with 91% and right femur bone with 89% confidence, d) pelvis detected with 66% confidence (inaccuracy in the shape of bounding box can be seen, although the class of the bone has been choosen correctly).

V. OBJECT DETECTION USED TO POSITIONING OF BONES MODEL WITH RESPECT TO CT DATA

In the second step of the procedure, the results of the object detection performed by YOLO network are used to fit the idealized surface models of separate bones, to the CT data of the patient. First, the class of the bone structure which will be transformed is chosen and on each image on which detection is found the bounding box with highest confidence score from selected class is chosen. In the following step, the rectangle bounding boxes are sorted according to the location of the corresponding CT slice on the z axis. Then, the 3-dimensional bounding box of cuboid shape (called further reference box) is created by choosing the minimal and maximal dimensions of the bounding box stack created in the previous step. Finally, the 3-dimensional bounding box is rescaled to the real world dimensions of CT data by multiplying x and y coordinates by number of rows or columns respectfully and pixel spacing parameter as well as adding the image patient position coordinates.

The bone from the idealized model is scaled in each dimension by the coefficient equal to the ratio between the size of the reference box and the size of bounding box of the model bone in a given dimension. Then, the rescaled bone model is translated to the position of reference box. The results of adjustment are depicted on Fig 5.

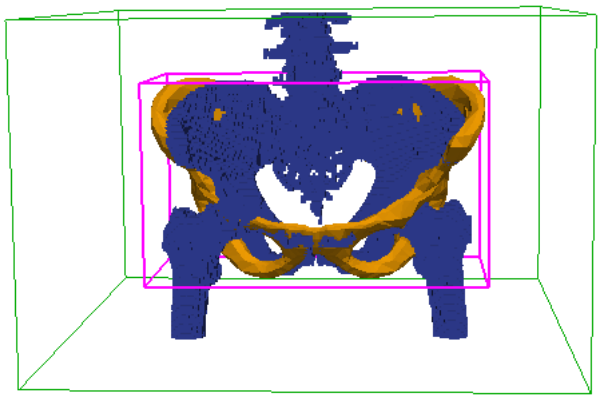

Fig. 5. The result of the positioning and scaling procedure on the example of the pelvis. Model bone depicted in the yellow color, is adjusted to the reference box (pink color). The CT bones data thresholded in range from 300 to 4000 of intensity values are marked with the blue color. The green cuboid depicts the bounds of whole CT data.

The adjusted bone model is further cut into slices along the plane of each CT slice. The result of the operation can be seen on Fig. 6 and Fig. 7. The cut is compared with the slice (blue one) created out of manual positioning of the pelvis on the basis of 4 landmark points chosen by a human expert. In the manual method 4 points are predefined and added to reference model. The expert is choosing 4 corresponding points in the CT data. Then, reference model is rescaled proportionally to the ratio between the two sets of points. In case of manual and automatic method as well the slices do not fit to the CT data but the shape of the slice created out of automatic positioning procedure resembles the pelvic cross-section visible on the slice. In the next step, it can be further adjusted by scaling the shape to fit the size of its bounding box to the two-dimensional box connected with current slice (detected by YOLO) and performing non-affine morphing transformation.

VI. Summary

Object detection performed with the use of YOLO neural network is able to predict the localization and class of different bone structures across whole CT data slices set, with satisfactory accuracy. Although the network was trained on relatively small dataset (320 images created out of 2 CT data sets) bone classes are recognized correctly in significant majority of test cases.

The series of bounding boxes obtained by YOLO framework allow to determine the boundaries of the area of individual bones in CT data, which are further used to perform affine match between model bones and real patient data. The positioning procedure is automatic and therefore more convenient

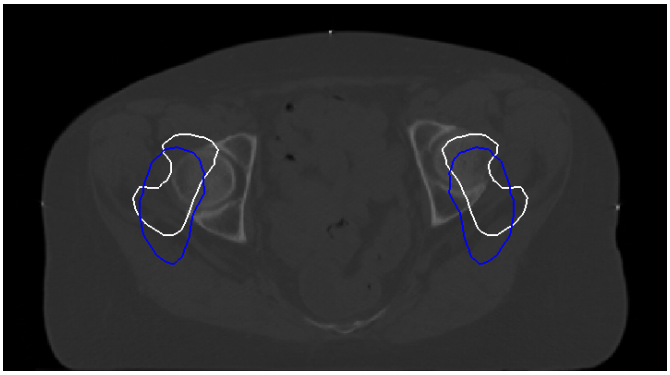

Fig. 6. Example of the slice through the manually (blue color) and automatically (white color) positioned pelvis model.

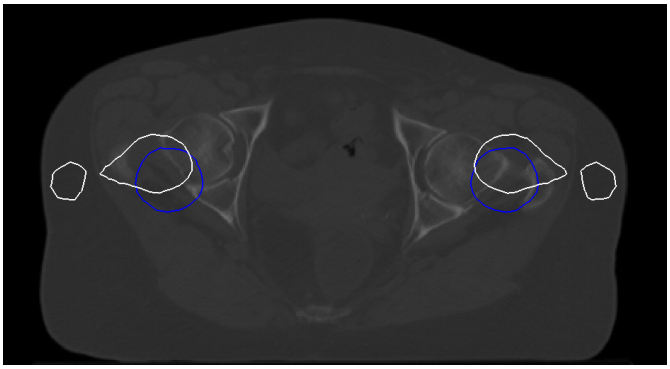

Fig. 7. Example of the slice through the manually (blue color) and automatically (white color) positioned femur model.

than manual approach, where a human expert has to indicate the landmark points on the data. The results will be further transformed by non-affine algorithm to fit them to the CT data. By using YOLO combined with the affine scaling we make this last step easier, because the source is now close to the target.

VII. Acknowledgment

The CT data obtained due to courtesy of Maria Skłodowska-Curie Memorial Cancer Center.

References

[1] W. Khan, "Images Segmentation Techniques: A Survey", Journal of Image and Graphics Vol. 1, No. 4, December 2013
[2] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", CVPR (Computer Vision and Pattern Recognition) 2016 arXiv:1506.02640
[3] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169
[4] J. Redmon, A. Farhadi, "YOLO9000: Better, Faster, Stronger", CVPR (Computer Vision and Pattern Recognition) 2017 arXiv:1612.08242
[5] J. Redmon, "Darknet: Open Source Neural Networks in C", 2013-2016, http://pjreddie.com/darknet/