Teeth and Landmarks Detection and Classification Based on Deep Neural Networks

Chapter · January 2019

DOI: 10.4018/978-1-5225-6243-6

4 authors, including:

Lyudmila N. Tuzova

10 PUBLICATIONS **423** CITATIONS

SEE PROFILE

Dmitry Tuzoff

Denti.AI Technology Inc.

10 PUBLICATIONS **428** CITATIONS

SEE PROFILE

Teeth and Landmarks Detection and Classification Based on Deep Neural Networks

Lyudmila Tuzova (ltuzova@denti.ai),
Denti.AI, Russia

Dmitry Tuzoff (tuzov@logic.pdmi.ras.ru),
Steklov Institute of Mathematics in St. Petersburg, Russia

Sergey Nikolenko (sergey@logic.pdmi.ras.ru),
Steklov Institute of Mathematics in St. Petersburg, Russia

Alexey Krasnov (alexey.krasnov@fnkc.ru),
Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Russia

ABSTRACT

In the recent decade, deep neural networks have enjoyed rapid development in various domains, including medicine. Convolutional neural networks (CNNs), deep neural network structures commonly used for image interpretation, brought the breakthrough in computer vision and became state-of-the-art techniques for various image recognition tasks, such as image classification, object detection, and semantic segmentation. In this chapter, the authors provide an overview of deep learning algorithms and review available literature for dental image analysis with methods based on CNNs. The present study is focused on the problems of landmarks and teeth detection and classification, as these tasks comprise an essential part of dental image interpretation both in clinical dentistry and in human identification systems based on the dental biometrical information.

Keywords: Radiographic Image Interpretation, Dental Image Analysis, Convolutional Neural Networks, Deep Learning, Teeth Detection, Cephalometric Landmarks Detection, Teeth Numbering, Teeth Segmentation

INTRODUCTION

Over the last decade, computer vision models and algorithms based on deep learning models have been successfully applied to various health and medicine domains in a number of medical imaging tasks such as detection and staging of cancer, lung segmentation, diagnosis of colitis, detection and classification of brain diseases, and many others (Lee, et al., 2017; Rezaei, Yang, & Meinel, 2017; Liu, et al., 2017; Litjens, et al., 2017; Shen, Wu, & Suk, 2017). In dentistry, several works have applied deep learning models and algorithms for dental radiograph analysis (Miki, et al., 2016; Lee, Park, & Kim, 2017; Wang, et al., 2016; Ö. Arik, Ibragimov, & Xing, 2017; Tuzoff, et al., 2018). However, deep learning in dentistry still remains an underdeveloped area of research, even though deep neural networks provide state-of-the-art results in many kinds of image recognition tasks (Lee, et al., 2017; Litjens, et al., 2017; LeCun, Bengio, & Hinton, 2015) and problems such as teeth and landmark detection appear to be straightforward object detection problems that could be amenable to modern computer vision approaches based on deep learning methods.

In this chapter, the relevant literature on deep learning methods, specifically convolutional neural networks (CNNs), applied for the tasks of teeth and landmarks detection and classification by type or tooth number is reviewed. These tasks comprise an important part of dental X-ray image analysis. The results can be used to automatically fill a patient's dental records for medical history and treatment

planning, preprocess an image for further pathology detection, improve speed and accuracy of postmortem human identification, perform anatomical measurements, and other problems. The deep learning methods have previously been studied for pathology detection purposes (Wang, et al., 2016; Oliveira & Proença, 2011; Imangaliyev, et al., 2016) as well; however, a review of computer-aided disease diagnostics is out of scope of the present study.

CNNs represent state-of-the-art deep learning architectures commonly applied for image recognition tasks. Modern object classification, detection and segmentation approaches based on CNNs have shown promising results, often outperforming methods based on traditional computer vision or other machine learning techniques. An important advantage of deep learning techniques compared with traditional computer vision and other machine learning approaches is that deep learning algorithms do not rely on handcrafted feature extraction and can achieve high performance working with raw input such as pixel values for X-Ray image sources. These methods allow interpreting medical images even if the images are noisy, taken with a different equipment, or in a different setting than that of the data used for training the model. Despite the increasing popularity of the CNNs, the challenges of the application of such architectures still exist. One of the most significant limitation is the amount of annotated data required for the effective model training.

A number of other deep learning models have previously been studied for medical image analysis, including stacked auto encoders (SAEs) and deep belief networks (DBNs) (Shen, Wu, & Suk, 2017; Litjens, et al., 2017). However, CNNs currently represent the state-of-the-art architectures for image recognition tasks. In (Shen, Wu, & Suk, 2017) high performance of CNNs for the images interpretation is explained by the specific properties of the CNNs architectures that better utilize spatial information of images, when most of the other deep learning models process the input in the one-dimensional vector form. Moreover, training of the models, such as DBNs and SAEs, is a complex task combining unsupervised pre-training phase followed by the fine-tuning supervised step. A number of prior studies demonstrated that CNNs models outperformed other deep learning techniques for the image interpretation tasks, when there is annotated data available and end-to-end supervised learning can be performed (Wu, 2015; Song, Zhao, Luo, & Dou, 2017).

In current literature, convolutional neural networks have been studied for the following tasks:

- landmark detection on cephalometric images (Lee, Park, & Kim, 2017);
- landmark detection and anatomical type classification on cephalometric images (Ö. Arik, Ibragimov, & Xing, 2017);
- teeth structures segmentation on bitewing images (Wang, et al., 2016);
- teeth classification by type on computed tomography (CT) (Miki, et al., 2016);
- teeth detection and numbering on panoramic views (PV) (Tuzoff, et al., 2018).

In the next sections, a brief history of deep learning and specifically CNNs is provided followed by the review of the methods listed above. Based on the reviewed works, the authors evaluate benefits and limitations of the existing approaches and provide the vision of CNN-based solutions prospects.

BACKGROUND

Deep Learning and Feedforward Neural Networks

Machine learning algorithms aim to learn important features from the set of available data and to apply this knowledge for further interpretation of previously unseen samples. In computer vision, for example, the main problem is to extract features that would be useful for high-level semantic problems such as object recognition.

Deep learning algorithms is one of the most promising classes of machine learning approaches demonstrating the growing popularity in various domains (LeCun, Bengio, & Hinton, Deep learning, 2015; Goodfellow, Bengio, & Courville, 2016; Schmidhuber, 2015). The main limitation of the

conventional machine learning algorithms is that they rely on the accuracy of feature engineering when the experts in the specific domain have to design the features by hand. On the contrary, deep learning methods allow computers to learn features from the raw data input.

Many deep learning methods are based on feed-forward neural network architectures. The neural networks have originally been inspired by the desire to imitate the work of a brain. By brain analogy, the core computational unit of the neural network is a neuron. The neurons are grouped together to form the layers of the network. The strengths of the connections (synapses) between the neurons are defined by their weights. At each layer, every neuron computes a weighted sum of inputs from the neurons of the previous layer by the weights of the connections, and passes the result through the non-linear function (activation function). The last output layer of the neural network does not typically use an activation function, but computes the final result, for example, in the form of class scores for classification task or real values for regression task.

There are two main learning approaches used in machine learning methods, both deep and not: supervised, when the annotated data is used to represent the ground truth for training, and unsupervised, when the algorithm aims to reveal clusters of data without using the ground truth annotations. In computer vision, supervised learning is more common.

The supervised learning process can be illustrated using the example of image classification task. To start the training, the model is set with the initial weights. During the training phase, the network processes input images each labeled by a particular class within the known set of possible classes. The output layer of the network produces confidence scores that estimate probability of the image to be any of these classes. The algorithm then measures the error using the special loss function that calculates the difference between the ground truth and predicted scores. To learn from the data, the stochastic gradient descent (SGD) or similar procedure is commonly used. For a batch of images, SGD consists of calculating the average gradients of the loss function, being its local derivatives with respect to the network weights, and adjusting the weights accordingly to decrease the error. The backpropagation procedure based on the chain rule is used to calculate the gradient from the top to the bottom layers. After training, the performance of the model is measured using a different set of images, called test dataset. This step is necessary to evaluate the generalization potential of the model on previously unseen images.

Deep neural networks are represented by multi-layer architectures, where non-linear modules gradually transform the input, producing more and more abstract representations of the data. In computer vision, for example, the first layer can detect simple graphical structures such as lines coming at different angles; the second layer can detect particular arrangements of edges; and so on, with higher layers recognizing complex shapes and patterns. Each next layer detects combinations of features extracted by the previous layer, and the final output layer of a neural network computes the final result. The crucial point for applications is that these layers of features are learned from data but not designed by human experts.

Convolutional Neural Networks

In computer vision, one of the most popular deep network is a convolutional neural network (CNN). The CNN architectures have their roots in the imitation of the visual cortex path organization. In 50^{th} and 60^{th}, studying the brains of the cats, Hubel and Wiesel (Hubel & Wiesel, 1962; Hubel & Wiesel, 1959) discovered that the cats' visual cortex is organized hierarchically, where the simple cells are responsive to simple stimulus such as edges orientation; the complex cells are responsive to both the orientation and the movement; and so on with the growing complexity (Hubel & Wiesel, 1962; Schmidhuber, 2015). In 1982, Fukushima et al. (Fukushima & Miyake, 1982) proposed the neural network model for visual pattern recognition task inspired by the visual cortex structure. This early model was similar to the architecture of modern CNN; however, no supervised learning was used at that moment. Finally, in 1998, Yan LeCun et. al (LeCun, Bottou, Bengio, & Haffner, 1998) introduced the convolutional neural network architecture using gradient-based learning applied for the task of document reading. Since then, a lot of new architectures were proposed; however, the core ideas stay the same. A more detailed history of deep learning is overviewed in (Schmidhuber, 2015).

Similar to traditional feedforward neural networks, CNNs consist of neurons, have weights to be learned, and use both linear and nonlinear transformations to extract features from the data. However, in contrary to the traditional feed-forward neural network, where neurons are fully connected with each other and the number of neurons at each level defines the amount of parameters to be learned, the CNN architecture is organized in a special way where the learnable weights are shared between the neurons of the same feature map. For computer vision, the CNN architectures exploit specific properties of the images allowing to process the raw pixel-level input much more efficiently using significantly lower number of parameters to be learned. In addition to the reduced number of parameters, this structure is also less prone to the overfitting problem.

The architecture of CNNs is represented by the sequence of layers of different types. First layers commonly combine convolutional and pooling layers. Convolutional layers play the key role in the input transformation. Their parameters form a set of learnable filters, where each filter have a fixed size, for example 5x5x3. At the convolutional layer, the neural network slides the filter over the input and computes dot products between the entries of the filter and the input. The results are typically passed through the non-linear function producing the feature map. The intuition behind this process is that the network learns distinctive patterns for each applied filter that are spatially correlated and can be found in different locations of the input. For the image, the network can learn filters that activate when they see some type of a visual pattern, for example a line of a particular orientation.

The pooling layers are used to reduce the spatial size of the input representation. One of the popular ways is to compute maximum value between the local units of the feature map. This procedure allows, on the one hand, to reduce the number of parameters to be learned; and, on the other hand, to control the overfitting. The output of a convolutional layer with or without pooling is a three-dimensional tensor, where the number of filters applied defines the depth. This outputting tensor can serve as an input for subsequent convolutional layers making the network deeper.

The last layers of the CNN are often represented by the fully-connected layers, where each neuron is connected to all neurons of the previous layer. The last fully-connected layer, called the output layer, calculates the output of the network. In Figure 1, examples of a regular feedforward neural network and CNN are presented.

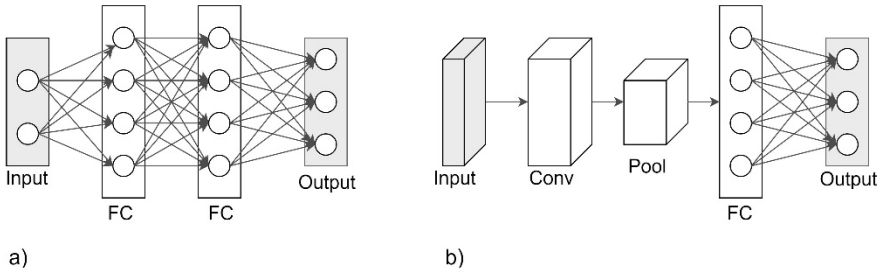


Figure 1. Regular NN and CNN. a) An example of regular feedforward neural network represented by the sequence of fully-connected layers. It consists of an input layer, two ‘hidden’ layers, and an output layer. All neurons of adjacent layers are connected with each other; however, neurons are not connected within one layer. b) Convolutional neural network. It consists of the 3D input layer, for example an image, where width and height are the dimensions, and the depth is the number of channels; convolutional layer followed by pooling layer; and two fully-connected layer including the output one.

CNNs in Computer Vision

During the last decade, efforts were made to develop architectures based on CNNs to solve various computer vision tasks, including image classification, object localization, object detection and semantic segmentation (Goodfellow, Bengio, & Courville, 2016; LeCun, Bengio, & Hinton, Deep learning, 2015; Schmidhuber, 2015; Huang, et al., 2017).

Image classification is one of the most traditional task in computer vision. The purpose of the image classification problem is to label an image with a particular class within a known set of possible classes.

In many respects, this task stimulates the current rise of the deep learning and CNNs. The convolutional neural network in its modern form was developed in 1998 by Yan LeCun (LeCun, Bottou, Bengio, & Haffner, 1998); however, for the long time, the CNNs were not actively used in practice. In 2012, the AlexNet CNN architecture (Krizhevsky, Sutskever, & Hinton, 2012) demonstrated spectacular results while classifying a large dataset of hand-annotated images (ImageNet) under the Large Scale Visual Recognition Challenge (ILSVRC). This CNN-based algorithm significantly outperformed other traditional computer vision techniques. Since that moment, the CNNs has been developed rapidly and a number of new architectures was proposed to solve the classification problem, such as ResNet-101 (He, Zhang, Ren, & Sun, 2016), VGG-16 Net (Simonyan & Zisserman, 2015), Inception v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016).

In dentistry, the object classification problem is addressed in (Ö. Arik, Ibragimov, & Xing, 2017) for landmark detection, in (Miki, et al., 2016) for teeth classification on CT, and in (Tuzoff, et al., 2018) for teeth classification on panoramic views. Ö. Arik et al. (Ö. Arik, Ibragimov, & Xing, 2017) proposed a custom CNN architecture and used multiple CNNs for binary-class classification. Miki et al. (Miki, et al., 2016) based their work on AlexNet CNN architecture (Krizhevsky, Sutskever, & Hinton, 2012). Tuzoff et al. (Tuzoff, et al., 2018) used VGG-16 (Simonyan & Zisserman, 2015) as a base CNN for teeth classification method.

The localization problem aims not only to classify the singe object on the image, but also to output the bounding box coordinates that enclose the object. This task is commonly addressed as a regression problem, when the network produces both the class scores and the real numbers that can be translated to the coordinates of bounding boxes. The CNNs used for object localization are commonly based on the same architectures as classification CNNs with the difference that the last layer of these networks branch into two heads: regressor and classifier. The joined multitask loss function is commonly used to train the model. In dentistry, the object localization problem was not addressed in prior works. However, Lee et al. (Lee, Park, & Kim, 2017) used a regression approach to define the coordinates of the landmarks on cephalometric images, without predicting class labels. The authors proposed a custom CNN architecture to solve the problem.

Object detection task aims to detect varying number of objects on a single image and mark them with bounding boxes. One of the most popular classes of object detectors uses two-phase approach: first, they find the regions of interests (RoIs); second, they use these region proposals to get class labels and bounding box coordinates for actual objects. The early architectures of this type used external non-learning algorithms for RoIs generation, such as Selective Search (Girshick, Donahue, Darrell, & Malik, 2015). In 2017, a state-of-the-art Faster R-CNN architecture was proposed that combines the region proposal with the object detection in the unified CNN (Ren, He, Girshick, & Sun, 2017). There are also architectures that follow one-phase approach to locate the objects without a separate step for RoI generation, including YOLO (Redmon & Farhadi, 2017) and SSD (Liu, et al., 2016). Huang et al. (Huang, et al., 2017) provided the extended comparison of modern object detectors and their performance. In prior works for dentistry, the object detection problem was addressed in (Tuzoff, et al., 2018) to detect the teeth on panoramic views. The authors based their method on Faster R-CNN architecture (Ren, He, Girshick, & Sun, 2017)

Finally, semantic segmentation task aims to find varying number of objects in the single image on the pixel-level basis. Semantic segmentation is one of the most challenging problem in computer vision, as it requires classifying each pixel to be an object of a particular class or background. Most of the CNN-based methods for pixel-level segmentation use the concept of the fully convolutional networks. The idea of fully convolutional networks is to replace last fully-connected layers of the typical CNN with some kind of upsampling process that allows to output the segmentation map that corresponds to the size of the initial input image. Most well-known architectures for semantic segmentation include Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017) and U-Net (Ronneberger, Fischer, & Brox, 2015). In prior works in dentistry, the U-Net (Ronneberger, Fischer, & Brox, 2015) architecture was used to segment teeth structures on the bitewing images (Wang, et al., 2016).

METHODS REVIEW

Landmark Detection and Classification with CNNs

In dentistry, one of the most common types of landmarks to be examined are the landmarks on cephalometric X-ray images. The cephalometric radiograph interpretation is a commonly used tool for human skull examination. This kind of X-rays depicts the patients' head side view and allows to observe dental and skeletal structures of a human skull for further diagnosis, treatment planning and surgery. The essential part of the cephalometric radiograph processing is an accurate localization of the landmarks, special anatomical points used for cephalometric analysis.

By measuring distances and angles between the landmarks, it becomes possible to detect abnormalities of the skull structure. Manual localization of landmarks and an accurate estimation of spatial relationships among them is a time-consuming and subjective procedure (Wang, et al., 2016). An automatic solution of cephalometric landmark detection could eliminate these problems and help improve the quality of the diagnosis and treatment.

Over the recent years, efforts have been made to automate the task of cephalometric landmark detection and classification. In 2015, the Cephalometric X-ray Image Analysis Challenges were performed under the support of IEEE International Symposium on Biomedical Imaging (IEEE ISBI) (Wang, et al., 2016). Wang et al. (Wang, et al., 2016) presented a comparison of different algorithms for automatic cephalometric analysis. In this challenge, the task of cephalometric landmark detection was addressed mostly using the random forests algorithms (RFs). Results of the 2015 challenge demonstrated the progress in this area made over the last years; however, accurate interpretation of cephalometric radiography remains to be a challenging task.

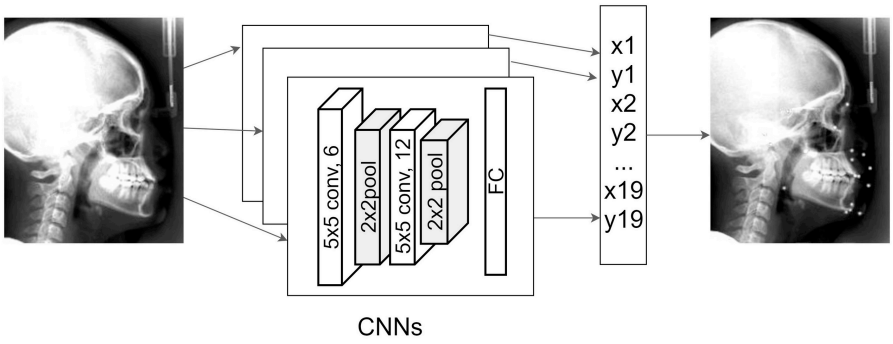
Landmark Detection on Cephalometric Images

In 2017, Lee H. et al. (Lee, Park, & Kim, 2017) proposed a novel method for landmark detection on cephalometric images based on CNNs. Lee H. et al. (Lee, Park, & Kim, 2017) used a dataset of 300 cephalometric images that was randomly divided into 150 training samples and 150 test samples. For each image, two experts provided ground truth annotations for all 19 landmarks.

To train the model, Lee H. et al. (Lee, Park, & Kim, 2017) formulated the task of landmark detection as a regression problem. For each of 19 landmarks, the two target variables were defined: X and Y-coordinates, each normalized by the length of the coordinate axis. Lee H. et al. (Lee, Park, & Kim, 2017) then constructed CNN regression systems for all 38 coordinates with the same underlying architecture and trained CNNs to predict all these coordinate variables from an input X-Ray image. In the proposed system, the authors used a very simple CNN structure with two convolutional layers each followed by the pooling layer, and one fully-connected layer. The output layer of the system produces values that correspond to the normalized coordinate variables of the landmarks (Lee, Park, & Kim, 2017). The schematic representation of the proposed solution is shown on Figure 2.

Lee H. et al. (Lee, Park, & Kim, 2017) did not provide benchmark comparisons with other methods providing the results in form of the box plot of average Euclidean distances between the landmark and ground truths. Lee H. et al. (Lee, Park, & Kim, 2017) reported that the proposed solution demonstrated promising performance locating landmarks within the reasonable distances from the correct annotations; however, the detection accuracy was relatively low. The authors made a conclusion that results of applying CNNs for this new problem is promising considering the simplicity of the method; however, the model performance can be improved with applying of additional techniques: training of deeper networks, using larger input images, and implementing more advanced techniques like data augmentation.

Figure 2. Lee H. et al. (Lee, Park, & Kim, 2017) method architecture. The input cephalometric image is processed by multiple CNNs each defining one coordinate of a landmark. The resulting set of coordinates is used to place the landmarks on the image.



Landmark Detection and Anatomical Type Classification on Cephalometric Images

In 2017, Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) proposed an automatic solution for landmark detection and anatomical type classification on cephalometric images. The authors implemented a framework based on CNNs to locate the landmarks, and used estimated landmark locations to perform a further pathology analysis. The authors used a public dataset of 400 cephalometric images from the IEEE ISBI Challenge (Wang, et al., 2016). In their work, Arik S. et al (Ö. Arik, Ibragimov, & Xing, 2017) provided a comparison with the results of Cephalometric X-ray Image Analysis Challenges 2015 (Wang, et al., 2016). To be consistent with this challenge, Arik S. et al used 150 images for training and 250 for testing. The authors also divided the test set into the two subsets to follow the approach used in this challenge, where Test1 was provided to train and evaluate the models and Test2 was used for on-site competition.

In the proposed method, Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) followed the classification approach to detect landmarks using CNNs. Multiple CNNs were trained to estimate the probability of each pixel being a particular anatomic landmark. A statistical shape model was then used to identify the optimal candidate points. The architecture of the proposed CNN consists of four stages of convolutions each followed by ReLU non-linearity. The max-pooling layers are used to reduce the number of parameters. The last layer uses the sigmoid function to produce the scalar output in the range of (0, 1) estimating the probability of pixel to be a landmark (Ö. Arik, Ibragimov, & Xing, 2017). The simplified architecture is shown on Figure 3.

To evaluate the performance of the algorithms the authors used a success detection rate (SDR) for landmark detection and success classification rate for classification of anatomical types. The success detection rate p_z with precision less then z mm was formulated as:

$$P_z = \frac{\#\{j:||L_d(j)-L_r(j)||<z\}}{\#\Omega} \times 100\%, \quad (1)$$

where L_d, L_r are the location of the detected landmark and the ground truth landmark respectively; z is a precision range; $j \in \Omega$, and $\#\Omega$ is the number of detections made. The success classification rate was calculated as:

$$P_{succ} = \frac{\#\ of\ accurate\ classification}{\#of\ classification} \times 100\% \quad (2)$$

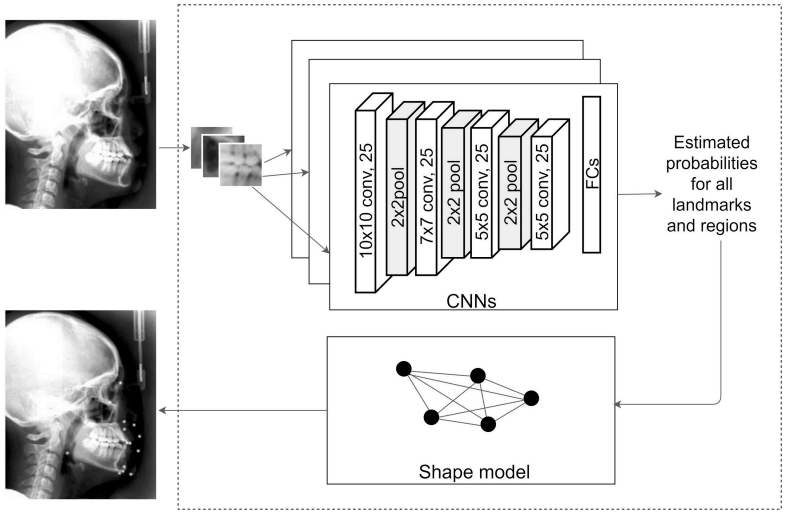
Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) reported the following results for their method: SDR of 75.37%, 80.91%, 84.32%, 88.25% in Test1 and 67.8%, 74.16%, 79.11%, 84.63% in Test2 using 2 mm, 2.5 mm, 3 mm, and 4 mm precision ranges; success classification rate of 75.92% and 76.75% for Test1 and Test2 respectively. Wang et al. (Wang, et al., 2016) reported the following results for the winner 2015 method based on RFs: the SDR of 73.68%, 80.21%, 85.19% and 91.47% in Test1 and 66.11%, 72%, 77.63% and 87.43% in Test2 using 2 mm, 2.5 mm, 3 mm, and 4 mm precision ranges; success classification rate of 76.12% and 80.99% for Test1 and Test2 respectively.

Based on these results achieved, the authors concluded that the proposed solution outperformed RF-based methods in the detection accuracy for the 2-, 2.5- and 3-mm ranges. Their method achieved a lower SDR

for 4 mm compared with the Lindner et al. method, the best solution of 2015 Challenge (Wang, et al., 2016; Lindner, 2015). The results of classification are slightly worse than the Lindner et al. technique due to the existence of outliers. Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) emphasized the fact that their CNN-based solution demonstrated better results for 2mm range, a clinically accepted detection range.

Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) also reported that while the proposed framework is a synergy of CNNs and a statistical shape model, the overall performance was mostly produced by the CNNs part. The naïve technique such as spatially averaging the highest CNNs output gave the SDR of 67.22%, 79.4%, and 83.20% for the ranges of 2-, 2.5- and 3-mm. The authors also expected that the performance of the proposed method can be improved with the increased size and diversity of the training dataset.

Figure 3. Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) method architecture. The CNNs are used to locate the landmarks on cephalometric images. The resulting classification scores are combined together to be refined by statistical shape model.



Teeth Detection and Classification with CNNs

The task of teeth detection and classification is an essential part of dental radiography interpretation applied for various types of radiographic images. Results of dental radiography interpretation can be used both in clinical practice and in human forensic identification.

In clinical dentistry, results of teeth detection and classification can be used to automatically fill dental charts for monitoring of dental health and treatment planning. Results of teeth detection can also be applied to improve further automated pathology detection. Digitalization of dental X-rays also stimulated the interest in using dental radiographs as a source of biometric information. Automatic dental identification systems (ADIS) use radiographic images to match X-rays for human identification. Accurate detection and classification of teeth on radiographic images can improve the speed and robustness of the ADIS systems (Hosntalab, Zoroofi, Tehrani-Fard, & Shirani, 2010).

During the past decade, a number of methods were proposed for teeth detection and classification: isolation with separation lines or curves (Wanat, 2011; Lin P. L., Huang, Cho, & Kuo, 2013; Zak, et al., 2018; Said, Nassar, Fahmy, & Ammar, 2006), pixel-level teeth segmentation (Hosntalab, Zoroofi, Tehrani-Fard, & Shirani, 2010; Lin, Lai, & Huang, 2010; Lin P. L., Huang, Huang, Hsu, & Chen, 2014; Oliveira & Proença, 2011; Shah, Baza, Ross, & Ammar, 2006; Tom & Thomas, 2015; Wang, et al., 2016), teeth detection using bounding boxes (Tuzoff, et al., 2018), and teeth classification by numbers or types (Hosntalab, Zoroofi, Tehrani-Fard, & Shirani, 2010; Lin, Lai, & Huang, 2010; Miki, et al., 2016; Tuzoff, et al., 2018).

Most of the methods for teeth detection were based on traditional computer vision techniques, such as thresholding, histogram-based methods, or contour models. The methods for teeth classification (Lin, Lai,

& Huang, 2010; Hosntalab, Zoroofi, Tehrani-Fard, & Shirani, 2010) used machine learning approaches, such as SVMs and neural networks; however they were still based on hand-crafted feature extraction algorithms.

Teeth Structures Segmentation on Bitewing Images

In 2015, the method of pixel-level segmentation based on CNN, was introduced under the Computer-Automated Detection of Caries in Bitewing Radiography Challenge (Wang, et al., 2016). Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) proposed a fully-automated solution for segmentation of teeth structures, such as caries, enamel, dentin, pulp, crown, restoration, and root cannel treatment. In this competition, a dataset of 120 bitewing images was used. As with the Cephalometric X-ray Image Analysis Challenges, the bitewing dataset was split into three parts: training set of 40 images, Test1 set of 40 images, and Test2 set of 40 images.

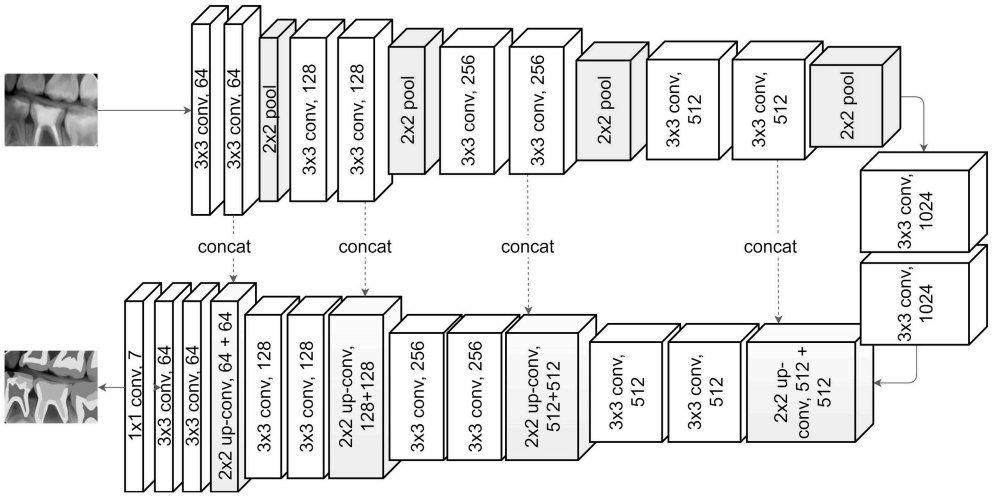
Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) implemented a method based on U-Net (Ronneberger, Fischer, & Brox, 2015), a CNN architecture for image segmentation. The U-Net was initially developed for biomedical image segmentation. This architecture of U-Net consists of two parts: contracting path and expansive path. The standard layers of typical CNN represent the contracting path: the combination of convolutional and pooling layers that compress the original image into a lower-dimensional representation. The expansive path performs upsampling process. At each level, the convolution is applied to halve the number of feature channels. The resulting feature channels are concatenated with the corresponding feature map from the contracting path to add high-resolution features. At the final layer, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015) the seven-class model was trained to segment each of teeth structure types. The architecture of the method is shown on Figure 4.

The following metrics were used to evaluate the methods proposed under the challenge: $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, $Fscore = \frac{2TP}{2TP+FP+FN}$, where TP, FP, FN represent the true positives, false positives, and false negatives.

Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) reported the following average results for their method: Sensitivity of 0.578, Specificity of 0.983, and F-score of 0.567 in Test1; and Sensitivity of 0.531, Specificity of 0.982, and F-score of 0.525 in Test 2. The second-best results for the method based on random forests achieved Sensitivity of 0.548, Specificity of 0.912, and F-score of 0.322 in Test1; and Sensitivity of 0.497, Specificity of 0.913, and F-score of 0.287 in Test 2.

Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) method achieved greater than 0.7 F-score for the segmentation of the three fundamental dental structures: enamel, dentin, and pulp. The authors relied on data augmentation because of little available training data. The augmentation techniques allowed them to extend training dataset by transforming the available images in specific ways, such as rotation, zooming, distortion, etc. (Jung, 2015). Wang et al. (Wang, et al., 2016) reported that data augmentation helped to create additional training images for enamel, denting, and pulp; however, the augmentation process was less successful for other types of teeth structures.

Figure 4. Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) method architecture. The method is based on the U-Net CNN architecture (Ronneberger, Fischer, & Brox, 2015). The contracting path uses convolutional and pooling layers to extract the features of the image; expansive path upsample the feature maps, concatenate the result with the corresponding feature map of contracting path, and perform convolutions to reduce the number of channels.



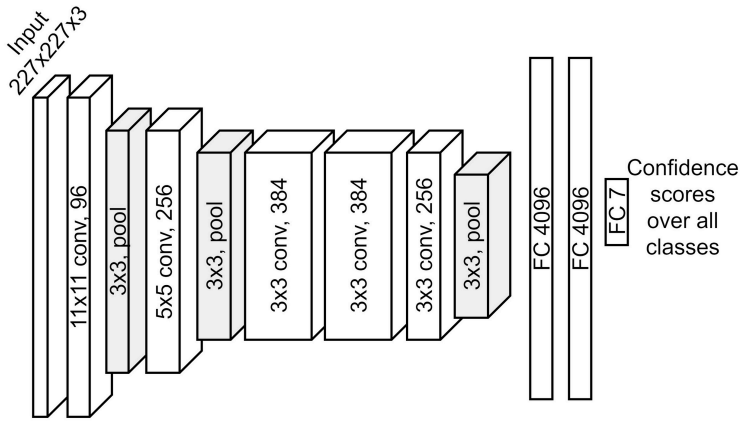
Teeth Classification by Type on Computed Tomography

In 2016, the method for teeth classification on CT based on CNN was proposed by Miki et al. (Miki, et al., 2016). This method investigated the application of deep learning for the purpose of teeth classification on CT by seven types: central incisor, lateral incisor, canine, 1st premolar, 2nd premolar, 1st molar, 2nd molar. Miki et al. (Miki, et al., 2016) used dataset of 52 CT images split into two parts: 42 images for training and 10 images for model evaluation.

To classify teeth on CT, Miki et al. (Miki, et al., 2016) first manually placed bounding boxes enclosing each tooth on the images. To classify each region into seven teeth types, the authors used a CNN based on classical AlexNet architecture (Krizhevsky, Sutskever, & Hinton, 2012). To increase the variety of the dataset Miki et al. (Miki, et al., 2016) also applied augmentation techniques, such as image rotation and intensity transformation. The architecture of AlexNet consists of five convolutional and three max pooling layers. Each convolutional layer is followed by the rectified linear unit (ReLU) activation function. The network outputs the probabilities for seven teeth types using the softmax function. The architecture of the method is shown on the Figure 5.

To evaluate the model performance, Miki et al. (Miki, et al., 2016) used the accuracy metric. Miki et al. (Miki, et al., 2016) reported 88.8% average accuracy for teeth classification, where the augmentation resulted in an approximately 5% improvement in classification accuracy. Miki et al. (Miki, et al., 2016) also reported that their study has some limitations: the third molars were excluded from the study due to the small number of samples, and teeth effected by metal artifacts were also excluded. Based on the achieved results and a comparison with prior works in the area, the authors of the study concluded that the CNN-based approach can achieve similar classification accuracy based on bounding boxes approach without the precise teeth segmentation. Miki et al. (Miki, et al., 2016) emphasized that the number of samples used to train the network was relatively small and the bigger amount of image available can improve the method performance. Miki et al. (Miki, et al., 2016) also reported that slice images were evaluated independently in this study, and discussed the possible application of 3D convolutions or combining the results for the tooth to improve the accuracy of teeth classification on CT.

Figure 5. AlexNet CNN architecture (Krizhevsky, Sutskever, & Hinton, 2012) used in Miki et al. (Miki, et al., 2016)



Teeth Detection and Numbering on Panoramic Views

In 2018, the method for teeth detection and numbering on panoramic views (PV) was introduced (Tuzoff, et al., 2018). Tuzoff et. al. proposed an end-to-end solution for teeth detection and classification based on CNNs. The teeth were detected on images and numbered with accordance to the two-digit FDI standard (ISO, 2016). The dataset of 1574 images was used to train and test the model. The dataset was split into two groups: training set of 1352 images and test set of 222 images. Five experts provided the ground truth annotations for the images.

To detect and classify teeth on PV, Tuzoff et. al. (Tuzoff, et al., 2018) proposed a system consisting of two modules. The detection module processes the image to detect the teeth in the form of bounding boxes; and the classification module classifies each detected tooth to assign a number among 32 possible teeth numbers. The overall architecture is shown on the Figure 6.

The authors used state-of-the-art Faster R-CNN architecture for the teeth detection module (Ren, He, Girshick, & Sun, 2017). The Faster R-CNN architecture evolved from Fast R-CNN (Girshick, Fast R-CNN, 2015), which, in turn, was based on R-CNN (Girshick, Donahue, Darrell, & Malik, 2015). The basic R-CNN model proposed a solution where RoI proposals were generated separately with the Selective Search algorithm (Uijlings, Sande, Gevers, & Smeulders, 2013), and then each region was evaluated with a CNN, also modifying bounding boxes using special regressors. The Fast R-CNN approach improved performance by simplifying the pipeline and sharing computation between the RoI extraction and object localization. Finally, the Faster R-CNN method introduced a unified network for object detection fully based on convolutional networks. Instead of Selective Search, it added a special region proposal network (RPN) to the CNN. Tuzoff et. al. (Tuzoff, et al., 2018) used the implementation of the Faster R-CNN (Hosang, 2016) with the Tensorflow backend (Abadi, et al., 2015). To initialize the model's parameters, the authors also used the weights pretrained on the conventional ImageNet data source (Deng, et al., 2009).

To generate region proposals the system slides the windows over the convolutional feature map. In (Tuzoff, et al., 2018) the VGG-16 Net was used as a base CNN for both modules. At each window location, Faster R-CNN outputs k potential bounding boxes named "anchors". Features extracted for each anchor are fed to a ROI box-regression layer and a softmax layer. The softmax layer produces objectiveness scores that estimate the probability of a box to be an object or a background. Region proposals generated by RPN were used as input for the Fast R-CNN detector. For each proposal, a region of interest (RoI) pooling layer extracts a feature vector from the feature map. Each feature vector was fed to a sequence of fully connected layers that branch into output layers: softmax layer to produce the class score to be a tooth or a background and a box regression layer to generate the final bounding box coordinates. The network used a combined loss function that consists of RPN and Fast R-CNN parts, jointly training for the classification problem and bounding-box regression. The Faster R-CNN architecture is shown on Figure 7.

The classification module of proposed system was based on VGG-16 Net CNN architecture (Simonyan & Zisserman, 2015). This CNN architecture consists of thirteen convolutional layers, five pooling layers, and three fully-connected layers. The output layer produces confidence scores over all 32-classes corresponding with 32 possible teeth numbers. To classify each tooth, the system cropped the images based on the results of detection module producing the image containing a classified tooth. Additional context was added to improve the classification results. Tuzoff et. al. (Tuzoff, et al., 2018) also implemented a heuristic method to improve classification results based on the teeth arrangement rules. The authors also used simple augmentation techniques, such as zooming and rotating, to increase the variety of the training dataset. Tuzoff et. al. (Tuzoff, et al., 2018) implemented classification method using Keras framework (Chollet, 2015). The VGG-16 Net architecture is shown on the Figure 8.

Tuzoff et. al. (Tuzoff, et al., 2018) used the following metrics to evaluate their method: precision $= \frac{TP}{TP+FP}$, recall $= \frac{TP}{TP+FN}$, and F2-score $= \frac{5*precision*recall}{4*precision+recall}$, where TP, FP, FN represent true positives, false positives and false negatives, respectively. For teeth classification, the accuracy metric was used.

The authors also evaluated the human expert performance based on the error analysis using the same metrics. Tuzoff et. al. (Tuzoff, et al., 2018) reported the following results for the system: precision of 0.9944, recall of 0.9941, F2-score of 0.9941 for teeth detection, and 0.98 accuracy for teeth classification. The human experts detected and classified teeth with the following results: precision of 0.9998, recall of 0.9980, F2-score of 0.9994 for teeth detection, and 0.99 accuracy for teeth classification. Based on these results, the authors concluded that their method achieved performance that is comparable with the human level. Tuzoff et. al. (Tuzoff, et al., 2018) also reported that additional context and augmentation allowed improving classification results on 6pp and 2pp, while the heuristic added 0.5pp improvement. The authors emphasized that heuristics had a very modest influence on system performance, and overall results were mostly based on the pure deep-learning techniques.

Figure 6. Tuzoff et. al. (Tuzoff, et al., 2018) *method architecture. Teeth detection module is based on Faster-RCNN* (Ren, He, Girshick, & Sun, 2017) *architecture. Teeth classification module utilizes VGG-16 Net CNN architecture to classify each tooth followed by heuristics to ensure correct teeth arrangement.*

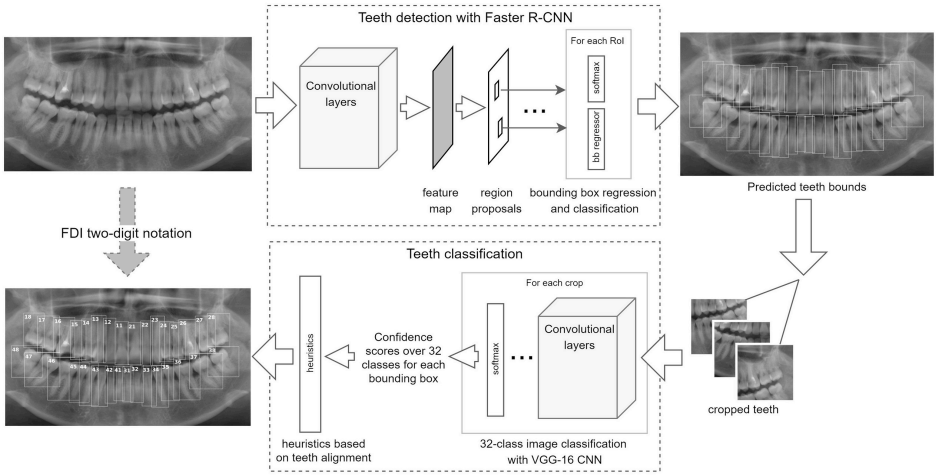


Figure 7. Faster R-CNN architecture used in (Tuzoff, et al., 2018)*. The RPN generates the region proposals; the object detector use the region proposals to classify each RoI and refine the bounding box coordinates.*

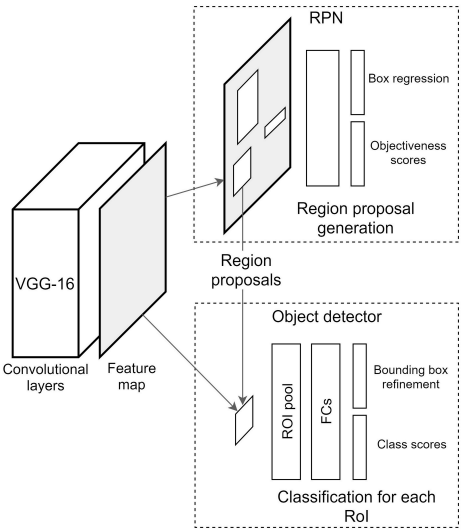
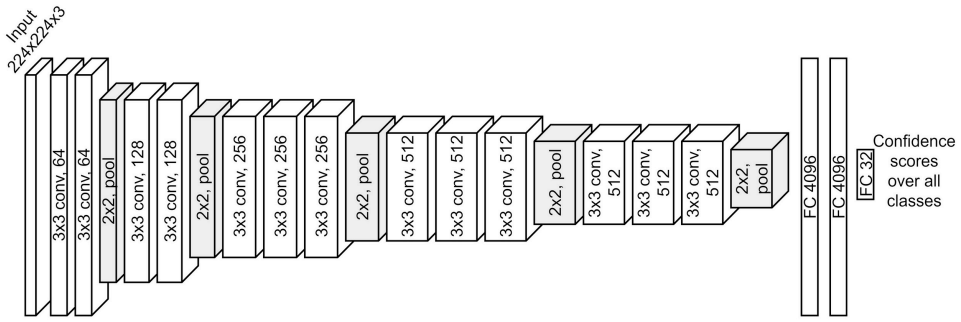


Figure 8. VGG-16 Net architecture used in (Tuzoff, et al., 2018)



CONCLUSION

In this chapter, the review of deep learning methods in dental radiography interpretation was performed. The present review was mainly focused on the tasks of landmark and teeth detection and classification, as these tasks is an important part of X-ray interpretation process.

Lee H. et al. (Lee, Park, & Kim, 2017) and Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) methods demonstrated the potential of application of CNNs for landmark detection on cephalometric X-Ray images. Lee H. et al. (Lee, Park, & Kim, 2017) used a straightforward approach to experiment with the CNNs. The authors reported that the proposed solution achieved promising results and their method allowed detecting landmarks within reasonable distances from ground truth. Lee H. et al. expected that the results can be improved by implementing more advanced techniques; including using deeper networks, applying image augmentation, training the model on larger training images dataset.

Arik S. et al. (Ö. Arik, Ibragimov, & Xing, 2017) in their work proposed a CNN-based architecture combined with the statistical shape model for cephalometric landmark detection. The authors compared the method performance with the winner solution from the Challenge 2015 (Wang, et al., 2016) that was based on RFs. Arik S. et al. CNN-based method outperformed the RF-based solution for 2-, 2.5- and 3-mm ranges. At the same moment, the method produced slightly lower results for larger ranges. Arik S. et al. concluded that the deep learning approach demonstrated very promising results, especially for the clinically accepted range of 2mm. Arik S. et al. also expected that their method could be further improved using extended training dataset (Ö. Arik, Ibragimov, & Xing, 2017).

Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) introduced a CNN-based solution for teeth structures segmentation on bitewing images under the Challenge 2015 (Wang, et al., 2016). Despite the fact, that the accurate teeth segmentation stayed to be a challenging task, Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) succeeded in demonstrating the high

potential of CNN-based architecture for image segmentation. Ronneberger et al. (Ronneberger, Fischer, & Brox, 2015; Wang, et al., 2016) significantly outperformed other methods mostly based on the RFs.

Miki et al. (Miki, et al., 2016) method for teeth classification on CT also showed promising results. The authors emphasized that high accuracy scores were achieved without the precise segmentation of teeth. Miki et al. (Miki, et al., 2016) also expected to improve the accuracy of classification by using more advanced techniques for 3D image processing and extending the training dataset.

Tuzoff et. al. (Tuzoff, et al., 2018) demonstrated the potential of deep learning techniques for the task of teeth detection and classification on panoramic views. The authors reported that the achieved performance results are comparable with the human level. Tuzoff et. al. (Tuzoff, et al., 2018) expected also to improve the results by implementing of more advanced augmentation techniques, using modern CNN architectures, and extending the dataset of the panoramic X-ray images used to train the models.

Based on the reviewed works, it can be concluded that deep learning algorithms, specifically the CNN architectures, have high potential to be applied for the tasks of dental radiography interpretation. At the same time, it can be seen that simple and straightforward approaches cannot compete with the traditional techniques. Effective application of CNN requires both using complex architectures optimized for particular computer vision task and supplementing techniques, such as advanced augmentation. CNN-based methods also significantly benefit from large amounts of annotated data available, which is not always the case.

The important advantage of CNN-based approaches is that they allow to process raw input without by-hand feature engineering. Since these architectures do not rely on hand-crafted features, the CNNs can be studied for a wide range of tasks: various dental and oral structures classification, pathologies detection, post-treatment evaluation. For dental and oral structures, it can be useful to locate implants, bridges, and crowns. For post-treatment procedures, CNNs can be trained to find treatment mistakes, such as root canal overfilling. For pathologies, the CNNs can be studied to detect different diseases, such as caries, periodontitis, and cysts.

Over recent years, the CNNs have become state-of-the-art architectures in all computer vision tasks independent of particular domains. Large community of researches believe in the future of deep learning methods and continue to develop more and more advanced architectures and techniques to meet the growing demand in automation of computer vision problems in various domains, including dentistry.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/

Chollet, F. (2015). keras. GitHub. Retrieved from https://github.com/fchollet/keras

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *The Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 248-255).

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition, 15*(6), 455-469.

Girshick, R. (2015). Fast R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, (pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*, 142-158.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 6). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 770-778).

Hosang, J. (2016). Faster RCNN TF. GitHub. Retrieved from https://github.com/smallcorgi/Faster-RCNN_TF

Hosntalab, M., Zoroofi, R. A., Tehrani-Fard, A. A., & Shirani, G. (2010). Classification and numbering of teeth in multi-slice CT images using wavelet-Fourier descriptor. *International journal of computer assisted radiology and surgery, 5*(3), 237-249.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., . . . Murphy, K. (2017, 7). Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive Fields of Single Neurons in the Cat's Striate Cortex. *Journal of Physiology, 148*(3), 574-591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106-154.

Imangaliyev, S., van der Veen, M., Volgenant, C., Keijser, B., Crielaard, W., & Levin, E. (2016). Deep Learning for Classification of Dental Plaque Images. In P. M. Pardalos, P. Conca, G. Giuffrida, & G. Nicosia, *Machine Learning, Optimization, and Big Data. MOD 2016. Lecture Notes in Computer Science* (Vol. 10122, pp. 407-410).

ISO. (2016). *ISO 3950:2016 Dentistry--Designation system for teeth and areas of the oral cavity.*

Jung, A. (2015). Image augmentation for machine learning experiments. GitHub. Retrieved from https://github.com/aleju/imgaug

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Annual Conference on Neural Information Processing Systems (NIPS)*, *2*, pp. 1097-1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2323.

Lee, H., Park, M., & Kim, J. (2017). Cephalometric Landmark Detection in Dental X-ray Images Using Convolutional Neural Networks. In S. G. Armato, & N. A. Petrick (Ed.), *SPIE Medical Imaging*, *10134*, pp. 1-6.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Bae Kim, G., Beom Seo, J., & Kim, N. (2017, 7). Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology, 18*, 570.

Lin, P. L., Huang, P. W., Cho, Y. S., & Kuo, C. H. (2013). An automatic and effective tooth isolation method for dental radiographs. *Opto-Electronics Review, 21*(1), 126-136.

Lin, P. L., Huang, P. Y., Huang, P. W., Hsu, H. C., & Chen, C. C. (2014). Teeth segmentation of dental periapical radiographs based on local singularity analysis. *Computer Methods and Programs in Biomedicine, 113*(2), 433-445.

Lin, P. L., Lai, Y. H., & Huang, P. W. (2010). An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information. *Pattern Recognition, 43*(4), 1380-1392.

Lindner, C. a. (2015). Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1862-1874.

Litjens, G., Kooi, T., Ehteshami Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., . . . I. Sánchez, C. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis, 42*.

Liu, J., Wang, D., Lu, L., Wei, Z., Kim, L., Turkbey, E. B., . . . Summers, R. M. (2017). Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Medical Physics, 44*, 4630-4642.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C.-Y. &. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, & N. &. Sebe, *Computer Vision -- ECCV 2016* (pp. 21-37).

Miki, Y., Muramatsu, C., Hayashi, T., Zhou, X., Hara, T., Katsumata, A., & Fujita, H. (2016). Classification of teeth in cone-beam CT using deep convolutional neural network. *Computers in Biology and Medicine, 80*, 24-29.

Ö. Arik, S., Ibragimov, B., & Xing, L. (2017). Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging, 4*, 014501.

Oliveira, J., & Proença, H. (2011). Caries Detection in Panoramic Dental X-ray Images. In *Computational Methods in Applied Sciences* (Vol. 19, pp. 175-190). Springer, Dordrecht.

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517-6525.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149.

Rezaei, M., Yang, H., & Meinel, C. (2017). Deep Neural Network with l2-norm Unit for Brain Lesions Detection. In L. D., X. S., L. Y., Z. D., & E.-A. ES., *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science* (Vol. 10637, pp. 798-807).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer Assisted Interventions Conference (MICCAI)*, *9351*, pp. 234-241.

Said, E. H., Nassar, D. E., Fahmy, G., & Ammar, H. H. (2006). Teeth segmentation in digitized dental x-ray films using mathematical morphology. *IEEE Transactions on Information Forensics and Security, 1*(2), 178-189.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85-117.

Shah, S., Baza, A. A., Ross, A., & Ammar, H. (2006). Automatic tooth segmentation using active contour without edges. *Biometrics Symposium*, (pp. 1-6).

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering, 19*, 221-248.

Simonyan, K., & Zisserman, A. (2015, 5). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*.

Song, Q., Zhao, L., Luo, X., & Dou, X. (2017). Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images. *Journal of Healthcare Engineering, 2017*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2818-2826).

Tom, C. E., & Thomas, J. (2015). Segmentation of Tooth and Pulp from Dental Radiographs. *International Journal of Scientific & Engineering Research, 6*.

Tuzoff, D., Tuzova, L., Bornstein, M. M., Krasnov, A., Kharchenko, M., Nikolenko, S., . . . Bednenko, G. (2018). Teeth detection and numbering in panoramic radiographs using convolutional neural networks. *DentoMaxilloFacial Radiology*.

Uijlings, J. R., Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision, 104*(2), 154-171.

Wanat, R. (2011). A Problem of Automatic Segmentation of Digital Dental Panoramic X-Ray Images for Forensic Human Identification. *Central European Seminar on Computer Graphics*.

Wang, C. W., Huang, C. T., Lee, J. H., Li, C. H., Chang, S. W., Siao, M. J., . . . Lindner, C. (2016). A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis, 31*, 63-76.

Wu, M. a. (2015). Image recognition based on deep learning. *Chinese Automation Congress (CAC)*.

Zak, J., Korzynska, A., Roszkowiak, L., Siemion, K., Walerzak, S., Walerzak, M., & Walerzak, K. (2018). The method of teeth region detection in panoramic dental radiographs. In *Advances in Intelligent Systems and Computing* (Vol. 578, pp. 298-307). Springer, Cham.

KEY TERMS AND DEFINITIONS

Backpropagation: The algorithm used in in artificial neural networks to calculate the gradient, the vector of partial derivatives, for further update of model parameters. The algorithm is based on the chain rule of derivation.

Convolutional Neural Network (ConvNet or CNN): A special type of feed-forward neural network optimized for image data processing. The key features of CNN architecture include sharing weights, using pooling layers, implementing deep structures with multiple hidden layers.

Feed-Forward Neural Network: The artificial neural network wherein the input data is processed in one direction without any cycle possible. The network gets its input and transforms it with multiple layers of neurons to produce the final result depending on the specific task, e.g. class scores for classification problem or real-values for regression problem.

Loss Function: A function used in supervised learning to measure the difference between the prediction and the ground truth.

Image Classification: A classical computer vision problem where the task is to label an image with the particular class within a known set of possible classes.

Object Detection: A computer vision problem that aims to locate a varying number of objects of different classes on a single image.

Semantic Segmentation: A computer vision problem where the task is to identify various objects on a single image on a pixel-level basis.

Stochastic Gradient Descent (SGD): The algorithm that aims to minimize the loss function for the supervised learning algorithms, including neural networks. It calculates the gradient, e.g. using backpropagation, and then changes the parameters of the models in the negative gradient direction. It works iteratively and typically uses a mini-batch of training samples at a moment.