Tooth Localization with Coarse-to-Fine Auto-Encoders

Ayşe Betül OKTAY

Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Istanbul, Turkey
abetul.oktay@medeniyet.edu.tr

(Geliş/Received:31.05.2017; Kabul/Accepted:24.01.2018)
DOI: 10.17671/gazibtd.317893

Abstract— Localization of teeth is a prerequisite task for most of the computerized methods for dental images such as medical diagnosis and human identification. Classical deep learning architectures like convolutional neural networks and auto-encoders seem to work well for tooth detection, however, it is non-trivial because of the large dental image size. In this study, a coarse-to-fine stacked auto-encoder architecture is presented for detection of teeth in dental panoramic images. The proposed architecture involves cascaded stacked auto-encoders where sizes of the input patches are increased with the successive steps. Only the detected candidate tooth patches are fed into the successive layers, thus the irrelevant patches are eliminated. The proposed architecture decreases the cost of detection process while providing precise localization. The method is tested and validated on a dataset containing 210 dental panoramic images and the tooth detection accuracy of the system is 91%.

Keywords—Coarse-to-Fine auto-encoder, detection, deep learning, panoramic dental image

1. INTRODUCTION

Detection of teeth from dental images [1] is the first step of most computerized dental applications like medical diagnosis [2] and human identification [3]. However, it is a challenging task because of the imaging artifacts and noise, large variation, and dental restorations. An example of a panoramic dental image is shown in Figure 1-a.

In the literature, most of the studies for tooth detection are performed with low-level vision techniques for intra-oral dental images that include a few teeth [4], [5]. There also exist machine learning based systems with hand-crafted features for tooth identification. For example, Mahoor and Abdel-Mottaleb [6] employed a Bayesian classifier to classify the teeth in bitewing images. Then, the labels of the teeth are determined using the spatial relationships between the teeth. The system presented in [7] implements the watershed algorithm to segment the
panoramic images into small regions and runs a fitness function with a set of features of each region for tooth detection. In the study of [3], support vector machines are used with several geometrical properties of the teeth for classification of the molar and the premolar teeth. Support vector machines are also employed for tooth classification and labeled them with the Markov Chain Model [8].

Recently, deep learning has been widely used in many computer vision applications and the state of the art results have been improved with the usage of training large amount of data [9], [10]. However, it is difficult to get such large training data (like 200M training images for face recognition [11]) for medical imaging applications. First of all, it is very hard to gather huge amount of real medical data from voluntary subjects with ethical approvals. In addition, it is very time-consuming to mark/delineate the images by an expert for obtaining the ground truth data. Also, the structures in medical images have high variability according to the illnesses, pathologies, etc. Despite these difficulties, there are many successful medical imaging applications based on deep learning in literature [12], [13] that produce successful results.

For panoramic dental images, it is intractable to detect all of the teeth simultaneously with a deep architecture because of the large image size. In addition, simultaneous detection of all teeth is difficult because of the missing teeth, dental restorations, and their repetitive structure. Therefore, we first find the possible tooth area in the images with a basic pre-processing steps. After reducing the search area, the teeth are detected with the cascaded stacked auto-encoders. The proposed system is tested on a dataset containing 200 images and the detection results are promising.

2. DENTAL PANORAMIC IMAGES AND PREPROCESSING

A panoramic dental image is an X-ray image (Figure 1-a) that includes the whole mouth. Normally, there are 8 teeth in each quarter $q$ of mouth where the upper left, upper right, lower left, and lower right quarters are shown by the symbols $q \in \{ \downarrow \uparrow \downarrow \uparrow \}$, respectively. We use Palmer's dental notation system where each tooth $t_i$, $1 \leq t_i \leq 8$, is numbered beginning from the mouth center with the quarter symbol $q$. For example, the upper right central incisor has the label $\uparrow \downarrow$, while the upper left canine tooth has the label $\downarrow 3$. Figure 1-b shows the label of each tooth and the other structures floor of nasal cavity and mouth gap- in a panoramic dental X-ray image.
The panoramic images include structures like jaws, sinuses, etc. besides teeth. In order to reduce the size of the search space for detection of teeth, we first detect the possible placement of teeth according to the mouth gap. While a panoramic dental image is taken, the subject bites a plastic plate with front teeth to ensure that mouth is open and to fix the head position. Thus, the mouth gap center is close to the vertical center of the image. Consider $I$ as an image of size $m \times n$. Let $\phi(w_{i,j})$ be a function that gives the location of the minimum value of horizontal projection histogram of the window $w$ centered at the image location $(i,j)$. The mouth gap center $m_{c}(n/2)$ is found by

$$m_{c}(j = \frac{n}{2}) = \phi(w_{\frac{n}{2}}), \ j = n/2$$ (1)

where $n$ is the length of the image $I$. In order to automatically determine $i$, we use the floor of nasal cavity $f$ which is the brightest part of $I$ on the upper front teeth. The maximum value of the projection histogram of the window centered at $i=m/4$ and $j=n/2$ is detected as the floor of nasal cavity.

The mouth gap $m_{c}$ is detected for the $i > f + \tau$ where $\tau$ is a threshold. Detection of mouth gap center with respect to the $f$ prevents the false detection of the sinuses (which are also dark pixels) as mouth gap. After finding the mouth gap center $m_{c}(n/2)$, the whole mouth gap is found iteratively to left and right by Eq. 2 and Eq. 3, respectively.

$$m_{c}(j + k) = \operatorname{argmin}\{\phi(w_{i,j+k}) + \delta m_{c}(j + k - 1)\}, \text{ if } j > n/2 \ (2)$$

$$m_{c}(j - k) = \operatorname{argmin}\{\phi(w_{i,j-k}) + \delta m_{c}(j + k + 1)\}, \text{ if } j < n/2 \ (3)$$

where $k$ is the iteration step and $\delta$ is a weighting term. If the value of $\phi$ is greater than a threshold the algorithm stops. The final mouth gap is found by fitting a spline to the detected points. The dental images have different sizes and the images are resized according to the mouth gap length. This process reduces the need of using many sliding windows at different sizes. Then, the mouth quarters $Q_{i} = \{J_{i} L_{i} \ \Gamma_{i}\}$ each containing 8 teeth are automatically detected. Each quarter has different tooth appearance because of rotation, so other 3 quarters $\{J_{i} L_{i} \ \Gamma_{i}\}$ are rotated and mirrored to make the similar appearance of teeth within the down left quarter $\Gamma_{i}$. Note that, this transformation provides training and testing all quarters together by eliminating the appearance differences between them.

3. STACKED-AUTO ENCODERS

Auto-encoders are neural networks that effectively learn the hierarchical representation of data. The stacked auto-encoders are stacked versions of auto-encoders and they are trained bottom up in unsupervised manner and features learned at the top layer are used for supervised training and fine-tuning.

Let $w_{i,j}$ be an image window and $x \in \mathbb{R}^{d}$ be the pixels in the window. An auto-encoder is a neural network that has an input, a hidden $h$, and an output layer. It takes $x$ as input and tries to find the optimal parameters $\Theta = (W, b)$ by minimizing the amount of distortion between the input and output. The input $x$ is mapped to a latent representation $h = (Wx + b)$ where the hidden layer $h$ is a new feature representation of the input image $x$. Then, the output layer (decoder) is trained to reconstruct the input image patch from the hidden representation with reverse mapping training by back-propagation. After training, the weights $W$ are used as the representation of data. The stacked auto-encoders are formed by stacking multiple auto-encoders by wiring output of each layer to the input of the successive layer.

4. COARSE-TO-FINE STACKED AUTO-ENCODERS

In the proposed framework, a number of stacked auto-encoders are cascaded to detect the teeth. The main objective of the architecture is making better detections at successive auto-encoders by the increasing image resolution and decreasing the running time. Only the candidates with better scores are evaluated at the successive layers which reduces the cost of detection. The architecture of the proposed framework is shown in Figure 2.

The first stacked auto-encoder takes the image patches of as input $x$ and extracts the feature representation $h$. These randomly selected patches are down-scaled into small patch sizes where $k \times k = x$. After the representations are learned, supervised learning is performed with a neural network layer with soft-max classifier that produces a score $s_{ij}$ indicating the probability of being a tooth candidate. The candidates having the best scores are fed into the successive layers. The number of best candidates $n_{l}$ with high scores at level $l$, that will be trained at the successive layer, are decreased while the size of the image patch is doubled. This allows to learn better tooth representations at successive layers and eliminates the non-tooth windows at the former stages.

5. TOOTH LOCALIZATION

In order to find the optimal final tooth positions $T_{f} = \{t_{1}, t_{2}, ..., t_{8}\}$, basic morphological operations are used. The $n_{l}$ candidates are the final candidates and maximum 8 teeth should be detected for each quarter. First, the detected pixels are grouped with labeling connected components and pixel groups are formed. A group $g_{l}$ has the average probability score

$$S_{g_{l}}^{k} = \frac{1}{n} \sum_{i,j} s_{i,j}, \ (4)$$

that is the sum of candidate scores $s_{i,j}$ divided by number of pixels in the group. Let the location of the centroid of
the group be \(L(g_k)\). We model the tooth positions with a
chain graph with 8 nodes. The final tooth positions are
determined by

\[
T' = \max_S \sum_{t_i=1,...,8} S^{g_k} + \frac{\mu(g_k) - \mu(t_i)}{2\sigma_i}
\]

where \(\mu\) is the mean and \(\sigma\) is the standard deviation of the
Euclidean distance between tooth \(t_i\) and its neighboring
tooth \(t_{i+1}\) which are calculated through the training set.
The first term in Eq. 5 includes the tooth representation
found by the stacked auto-encoders. The second term is
the geometric term that is the difference between the
group center and the average position of
the tooth \(t_i\) according to the training set. The geometric term prevents
two neighboring teeth being very close or far.

If the difference is greater than a threshold, the tooth is
determined as a missing tooth. By solving Eq. 5 by
dynamic programming, the optimal solution is found in
polynomial time.

6. EXPERIMENTS

In order to evaluate the effectiveness of the proposed
framework, a dataset including 210 dental panoramic
images of 174 different subjects is used. The images are
taken by 3 different dental panoramic X-ray machines and
they have different image sizes which are 2871x1577,
1435x791, and 2612x1244. The ground-truth bounding
boxes for each tooth are delineated by an expert. There
are 5568 teeth and number of missing teeth is 829. There
are abnormalities like impacted tooth and many dental
works like crowns, fillings, implants, and braces in the
images.

In the pre-processing step, the window size \(w\) for
detection of the mouth gap is selected as 800x100 pixels.
The weighting term \(\lambda\) is 0.2 and the threshold \(t\) is 150
pixels. After the mouth gap is detected, all of the images
are scaled to the maximum mouth gap length by
preserving the aspect ratio.

For training the system, 10 of images of 10 different
subjects are used. 50000 patches that have image sizes
between the minimum and maximum tooth sizes are
randomly selected. All of the patches are down-scaled to
32 by 32 pixels for the first level of the auto-encoders
where \(k = 32\). For the supervised softmax classifier, 150
windows that contain a whole tooth are used as positive
samples and 300 images that contain at most 30% of a
tooth are used as negative samples. The number of
candidates \(n_1\) is 3000 windows for the first level, \(n_2 =
1500\) for the second level and \(n_3 = 500\) for the third level.
Table 1. Detection percentages for each tooth

<table>
<thead>
<tr>
<th>Tooth number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>88</td>
<td>93</td>
<td>91</td>
<td>89</td>
<td>92</td>
<td>92</td>
<td>91</td>
<td>89</td>
</tr>
<tr>
<td>Γ</td>
<td>89</td>
<td>89</td>
<td>87</td>
<td>90</td>
<td>91</td>
<td>93</td>
<td>91</td>
<td>87</td>
</tr>
<tr>
<td>J</td>
<td>91</td>
<td>92</td>
<td>89</td>
<td>91</td>
<td>89</td>
<td>90</td>
<td>93</td>
<td>89</td>
</tr>
<tr>
<td>Σ</td>
<td>90</td>
<td>93</td>
<td>86</td>
<td>92</td>
<td>91</td>
<td>91</td>
<td>88</td>
<td>84</td>
</tr>
</tbody>
</table>

In order to evaluate the effectiveness of our system, we first detect the accuracy of each level at the coarse-to-fine framework. A detection is evaluated as accurate if the window includes at least 50% of the box including the tooth delineated by the expert. The ROC curves of each level is shown in Figure 3. The accuracy rate increases at each level while the true positive rate is 0.81, 0.89, and 0.91, at the levels 1, 2 and 3. The increasing true positive rates show the effectiveness of our system.

In order to evaluate the final localization results of the post-processing step, we calculated the average detection accuracy. If the finally detected tooth \( t_0 \) is fully inside the ground truth, it is evaluated as an accurate detection. The detection percentages for each tooth is shown in Table 1. The down first teeth have the lowest detection rates. This may cause because of the brightness of the plastic plate occluding the front teeth. Molar teeth numbered 6, 7, and 8 have the highest percentages. The visual results are shown in Figure 4. Figure 4-a shows the tested panoramic dental image. Figure 4-b and Figure 4-c show the detection results of the first and second level, respectively. The detected pixels detected as teeth are refined in Figure 4-c and final localization results are shown in Figure 4-d.

7. CONCLUSIONS

In this paper, a coarse-to-fine stacked auto-encoder is presented for detection of teeth in dental panoramic images. The presented technique has three important advantages. First, the search space is reduced with the developed mouth gap based pre-processing technique. Also, transforming the mouth quarters into similar appearance provided more samples for training while eliminating the cost of training each quarter separately. Second, the cascaded auto-encoders with increasing input image sizes produced precise localization results while decreasing the search time. Third, the patches not involving a tooth are effectively eliminated at the first steps. In the future, we will use the detected teeth and their hierarchical features for human identification by comparing same subjects’ dental images.

REFERENCES


