Texture-Map Based Branch-Collaborative Network for Oral Cancer Detection

Chih-Hung Chan, Pau-Choo Chung, Chih-Yang Chen, Chein-Chen Lee, Man-Yee Chan, Tze-Ta Huang

***Abstract*—The paper proposes an innovative deep convolutional neural network (DCNN) combined with texture map for detecting cancerous regions and marking the ROI in a single model automatically. The proposed DCNN model contains two collaborative branches, namely an upper branch to perform oral cancer detection, and a lower branch to perform semantic segmentation and ROI marking. With the upper branch the network model extracts the cancerous regions, and the lower branch makes the cancerous regions more precision. To make the features in the cancerous more regular, the network model extracts the texture images from the input image. A sliding window is then applied to compute the standard deviation values of the texture image. Finally, the standard deviation values are used to construct a texture map, which is partitioned into multiple patches and used as the input data to the deep convolutional network model. The method proposed by this paper is called texture-map based branch-collaborative network. In the experimental result, the average sensitivity and specificity of detection are up to 0.9687 and 0.7129 respectively based on wavelet transform. And the average sensitivity and specificity of detection are up to 0.9314 and 0.9475 respectively based on Gabor filter.**

***Index Terms*—oral cancer, auto-fluorescence image, texture, wavelet transform, Gabor filter, convolutional network, texture-map based branch-collaborative network**

I. Introduction

In 2012, the estimated numbers of the oral cancer case are more than 300000 and the numbers of the oral cancer deaths are more than 140000, globally [1]. Several studies [2-3] have shown that the diagnosis and treatment of oral cancer at the early stage results in a much higher survival rate than at a later stage. However, recognizing the symptoms of early cancer manually is extremely challenging, and depends heavily on the skills and experience of the particular doctor or health worker involved. As a result, almost half of all oral cancers are not currently diagnosed until stage III or stage IV. To address this problem, more effective screening programs for oral cancer are urgently required.

Various techniques are available for oral cancer screening, including incisional / excisional biopsy [4], brush biopsy [5] and salivary markers [6]. However, these methods all require a long analysis time. Methylene blue staining [7] provides a faster analysis result. However, in addition to an extremely unpleasant taste, methylene blue is a biologically active substance and can thus give rise to a number of health complications if not properly administered.

Accordingly, several authors have investigated the feasibility for using an auto-fluorescence imaging technique, in which a different response is observed between normal regions and lesion regions, respectively, for oral cancer recognition [8-10]. In [11], two auto-fluorescence images were mainly produced using different fluorophores (i.e., NADH and FAD). A biomarker referred to as the redox ratio based on the fluorescence values of corresponding pixels in the NADH and FAD images was used to perform tumor detection. Various auto-fluorescence-based devices have been proposed for non-invasive diagnosis in oral pathology. For example, the Velscope proposed in [12] uses an excitation blue light source to determine the auto-fluorescence response spectrum of oral tissue over a wavelength range of 400 ~ 460 nm. However, the Velscope only produces fluorescent images. In other words, a doctor or professional healthcare worker is still required to perform the actual ROI marking and diagnosis task.

In a previous study [13-14], the present group developed an oral cavity imaging system for performing the classification of oral cancer from normal tissue on selected ROI regions. However, in practical applications, the ROI must still be marked by a doctor. In this paper a novel deep learning architecture, texture-map based branch-collaborative network is developed for marking ROIs.

The texture-map based branch-collaborative network model comprises two branches, namely an upper branch consisting of fully connected layers for the detection, and a lower branch to mark the ROI. In the network architecture proposed for oral cancer detection, both the residual network (resNet) [15] approach and the inception modules [16] approach are explored. And the fully convolutional network (FCN) [17] and the feature pyramid network (FPN) [18] are explored for semantic segmentation of the cancerous regions which are detected by

This work was partly supported by the Ministry of Science and Technology, under Grant MOST-107-2634-F-006-004, and partly supported by the Delta-NCKU joint project.

Chih-Hung Chan is with the NVIDIA Corporation, Taipei, Taiwan (e-mail: junglec@nvidia.com).

Pau-Choo Chung, is with the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan (e-mail: pcchung@ee.ncku.edu.tw).

Chih-Yang Chen is with the Delta Electronics, Inc., Tainan, Taiwan(e-mail: rayan.chen@deltaww.com).

Chein-Chen Lee is with the Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan (e-mail: cclee@saturn.yzu.edu.tw).

Man-Yee Chan is with the Department of Stomatology at Taichung Veterans General Hospital. Taichung, Taiwan. (e-mail: c3883@vghtc.gov.tw)

Tze-Ta Huang is with the Department of Stomatology and Institute of Oral Medicine, College of Medicine, National Cheng Kung University, Taiwan. (e-mail: tzetahuang@gmail.com).

the upper branch. The FCN replaces the fully connected layers in traditional convolution network architectures with convolutional networks to achieve pixel-to-pixel prediction. Compared to FCN the FPN further connects the feature map in the front encoder stage to achieve a higher performance. The two paths in the model are integrated using the mask R-CNN method [19], thereby enabling oral cancer detection and ROI marking to be performed using a single model without scanning the whole image.

Another issue which worth attention is that the oral auto-fluorescence images have little visual features and the characteristics (feature representations) and shapes of oral cancer regions are of high variation. As such it is difficult to train the network to detect directly from the original image automatically, even with a deep convolutional network model. To address this problem, the texture feature maps are used as supports for building the branch-collaborative network model. Texture representations, which characterize lesions of the morphological variations caused by cancer, are powerful in diagnosis. As such, it has been extensively applied in the analysis of medical images [20]. In the texture-map based branch-collaborative network in this paper, two different texture features are used to facilitate the detection of cancerous regions in oral auto-fluorescence images. The first feature is the wavelet-based texture feature, which filters and decomposes the original signal into subbands representing the high-frequency components distributed over the vertical, horizontal or diagonal directions under different scales. The second feature is the 2D Gabor filter [21], which simulates the visual cortex of humans. The feature image is extracted via the convolution of the original image with a Gabor filter kernel consisting of a product of Gaussian and cosine functions referring to various orientations and spatial frequencies. Having extracted the feature image using either the wavelet transformation method or the Gabor filtering method, the standard deviations of the local areas in the image are computed to construct a feature map. The feature map is then partitioned into multiple patches and taken as the input to the oral cancer detection and segmentation framework. The feasibility of the proposed model is demonstrated experimentally. The results confirm that the model has high sensitivity and specificity, and allows the ROI to be marked automatically with a high degree of precision.

The paper is organized as follows: Section II introduces the related works. Section III describes the detection and the ROI marking methods. The texture extraction methods are also described in Section III. Several related network architectures such as the residual network, the inception module, fully convolutional network (FCN) and feature pyramid network (FPN) are also reviewed in this section. Then, a series of experimental results are presented in Section IV. Finally, the conclusions are drawn in Section V.

II. Related Work

Deep convolution neural networks (CNNs) are used for many applications nowadays, including object detection [22], segmentation [23] and image enhancement [24]. Due to powerful computation capability of modern computers and the graphic processor unit (GPU), it is enable to train the deep CNNs (convolutional neural networks) on large datasets to provide a result better than traditional machine learning methods. Some examples are the residual network (ResNet) [15] published by Microsoft research and GoogleNet [16] published by Google research. The residual network model is based on the concept of residual blocks. The residual blocks contain elements addition of identity and residual mapping for building the convolutional network. By contrast, the inception module computes various convolution kernel sizes in the same network layer and then concatenates all the results together. And with the fully connected layers, both of two methods provide a powerful classification for the object detection. Girshick et al. [25] presented a system referred to as region CNN (R-CNN) for performing object detection and semantic segmentation using a hierarchy of features. A selective search algorithm was first employed to find thousands of bounding box candidates with different sizes. And the input feature images were then cropped or warped into the same size and fed into a deep convolutional network architecture in order to learn the features. Finally, a support vector machine (SVM) algorithm was used to match the input features with appropriate bounding box candidates. However, the need to warp or crop the input feature maps to the same size represents a major disadvantage of R-CNN since it is not significantly flexible. Accordingly, in SPP Net [26], the need to warp or crop the input feature maps was eliminated by using a spatial pyramid pooling approach. In [27], the speed of R-CNN was further improved by replacing the SVM with a fully connected layer in order to integrate the feature learning step and classification step into a single model. This so-called fast R-CNN scheme was later extended to a faster R-CNN [28] scheme by using a region proposal network (RPN) to select suitable bounding box candidates. In a recent study [19], a mask R-CNN model was proposed incorporating three separate branches to perform classification, bounding box regression and segmentation, respectively, on a single platform.

Image segmentation has been widely used for ROI delineation. Typically, deep learning models for image segmentation have the form of fully convolutional networks (FCNs) [17, 29-30], in which semantic segmentation is performed following a process of end-to-end and pixel-to-pixel training. The feature pyramid network (FPN) model [18] can also be used for semantic segmentation and object detection. In FPN, the network model comprises two architectures, namely a bottom-up architecture and a top-down architecture. The bottom-up is an encoder. By contrast, the top-down architecture with lateral connections presents the high-level semantic feature maps at all scales for segmentation. In practice, either model (i.e., FCN or FPN) enables image segmentation to be performed using a deep learning method.

Various deep convolutional network architectures are available for constructing deep learning models [31-32]. In addition, several techniques exist for improving the training efficiency of deep learning models. One of techniques is dropout [33], which provides an efficient means of averaging many large neural nets. The other technique is Adam optimization [34], which combines the respective advantages of

AdaGrad and RMSProp, two commonly used methods based on the estimated first and second moments of the gradients.

However, these deep learning models are applied mostly on video images, which normally have obvious visual features. As for the medical images especially the auto-fluorescence images, the characteristics of the lesions are much more subtle and with high variation from patient to patient. The texture features used in the analysis of bio-medical images are generally extracted using some form of statistical over spatial domain, such as gray level co-occurrence matrix (GLCM), run length matrix and local binary pattern (LBP) [35-38]. However, the oral cavity images in different orientations and spatial frequency present different detail of textures. The texture extraction method of statistical over spatial domain are not enough to extract the useful information from oral cavity images. In this paper, the texture-maps are extracted from multi-resolution and multi-orientation to obtain different information of textures in spatial frequency domain. Generally speaking, the Gabor filter [21] and wavelet transformation [39-40] features are both robust rotation-invariant and multi-resolution texture features, which can be used either directly or combined with statistical methods (such as calculating the entropy, standard deviation or contrast) for analysis, classification or segmentation purposes.

III. METHOD

A. Texture-Map Extraction

According to [11] the redox ratio is computed on a pixel-by-pixel basis as

$$I_{Redox\ Ratio}(x,y) = \frac{I_{FAD}(x,y)}{I_{FAD}(x,y)+ I_{NADH}(x,y)} \quad (1)$$

where $I_{NADH}(x,y)$ and $I_{FAD}(x,y)$ denote the intensity values of corresponding pixels in the images associated with NADH and FAD and $I_{RedoxRatio}(x,y)$ is the intensity value of the same pixel in the redox ratio image. It should be noted that the specific frequency bands of light for activating the NADH and FAD, would also trigger the emittion of fluorescent lights of other tissue components, such as collagen. However, this would not affect the use in the discrimination between tumor and normal tissues. Even though collagen can also emit the fluorescence, [8-11] showed that the change of collagen is also one sign for differentiating tumor from normal. On the overall, tumor causes the tissues, including the NADH, FAD, Collagen, and etc, to change. The change of the tissues can be observed from the fluorescence emitted after activated by certain band of light. However, this would not affect the use in the discrimination between tumor and normal tissues.

Fig. 1 shows the feature map generation process. The image $I_{RedoxRatio}$ is computed by Gabor filter or wavelet feature computation, to obtain the texture image. The texture image is then scanned by a sliding window and standard value computed from the slide window to construct a corresponding feature map. Finally, the feature map is partitioned into patches, which are used as input to the deep convolutional network model to perform cancer detection and ROI segmentation, as illustrated.

To extract features in different spatial frequencies and directions, two multi-resolution transformations, namely

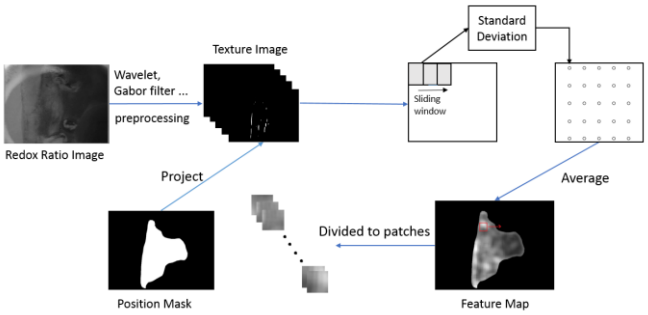

Fig. 1. The Feature map generation process.

wavelet transformation and Gabor filtering are explored. Brief descriptions of the two methods are shown in the following.

1) 2D Discrete Wavelet Transformation

Discrete wavelet decomposition is computed by convoluting the signal with a low-pass filter and a high-pass filter, denoted as $g(n)$ and $h(n)$ in the horizontal and vertical orientations, respectively, so as to obtain the approximation and detail coefficients. In practice, the filters can be acquired from various wavelet bases. The present study chooses the Daubechies basis with a vanishing moment of 2 (DB2).

For 2D wavelet decomposition, the convolution is calculated along both the horizontal direction and the vertical direction of the given image $I(x,y)$. Hence, four channels (subbands) are generated; denoted as LL, HL, LH and HH, respectively. Mathematically, the procedure used to generate each subband can be formulated as follows:

Decomposition along vertical direction –

$$I_L(m,n) = \sum_{k=0}^{K-1} I(m,2n-k)g(k), \quad (2)$$

$$I_H(m,n) = \sum_{k=0}^{K-1} I(m,2n-k)h(k), \quad (3)$$

Decomposition along horizontal direction –

$$I_{LL}(m,n) = \sum_{k=0}^{K-1} I_L(2m-k,n)g(k), \quad (4)$$

$$I_{LH}(m,n) = \sum_{k=0}^{K-1} I_L(2m-k,n)h(k), \quad (5)$$

$$I_{HL}(m,n) = \sum_{k=0}^{K-1} I_H(2m-k,n)g(k), \quad (6)$$

$$I_{HH}(m,n) = \sum_{k=0}^{K-1} I_H(2m-k,n)h(k), \quad (7)$$

The present paper adopts a three-level decomposition approach, which yields ten subbands for each image. The images of the vertical and horizontal subband images in each level are then processed using a sliding window to compute the corresponding standard deviation values and then averaged. As described above, these averaged standard deviation values are then used to construct a feature map as an input to the deep convolutional network.

2) 2D Gabor Filter

The Gabor feature image $r(x,y)$ is obtained by convoluting the input image $I(x,y)$ with a 2D Gabor function $g(x,y)$, i.e.

$$r(x,y) = \iint_{\Omega} I(i,j)g(x-i,y-j)didj, \quad (8)$$

where $(x, y) \in \Omega$, in which Ω is the set of image pixel coordinates.

The Gabor function in Eq. (8) has the form

$$g_{\lambda,\theta,\varphi}(x,y) = e^{-\left(\frac{x^2+\gamma^2 y'^2}{2\sigma^2}\right)} \cos\left(2\pi\frac{x'}{\lambda}+\varphi\right), \quad (9)$$

where

$x' = x\cos\theta + y\sin\theta,\ y' = -x\sin\theta + y\cos\theta$, $\sigma = 0.56\lambda$ and $\gamma = 1$

Here, σ is the standard deviation of the Gaussian function and determines the size of the receptive field, and γ is the spatial aspect ratio and determines the eccentricity of the receptive field ellipse. In addition, λ determines the spatial frequency $(1/\lambda)$, θ ($\theta \in (0, \pi]$) is the orientation of the normal-to-parallel stripes of the Gaussian function, and φ ($\varphi \in (-\pi, \pi]$) is the phase offset and determines the symmetry of the Gabor filter. In particular, the set $\{0, \pi\}$ represents a symmetric filter, while $\{\frac{\pi}{2}, -\frac{\pi}{2}\}$ represents an anti-symmetric filter [21].

As indicated above, the Gabor filter parameters were set as $\sigma = 0.56\lambda$ and $\gamma = 1$ in the present study. Moreover, a symmetric Gabor function ($\varphi = 0$) was employed. The Gabor feature images were generated by filtering the input redox ratio image with frequencies that are determined by $\lambda = 22.2, \lambda = 16.6$ and $\lambda = 10$. For each frequency, features were extracted in eight different orientations ($\theta = \frac{k}{8}\pi, 0 \le k \le 7$). For each frequency, the eight feature images were processed using a sliding window to obtain a set of standard deviation values and then averaged with which to generate the corresponding feature map.

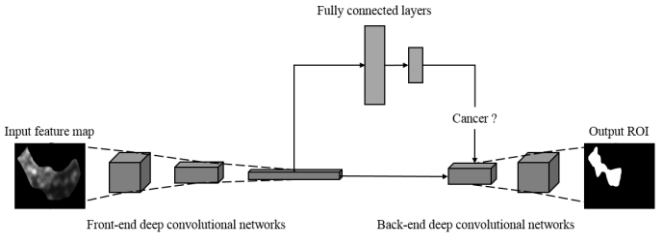

Fig. 2. Texture-map based branch-collaborative network model used for oral cancer detection and ROI marking.

B. Oral Cancer Detection

The present paper develops the texture-map based branch-collaborative network model which enables object detection and semantic segmentation to be performed on a single platform by exploiting the mask R-CNN architecture proposed in [19]. Fig. 2 illustrates the proposed model. As shown, the model contains two branches, namely an upper branch which is fully connected and is used to perform the detection of cancerous regions, and a lower branch which has the form of a deep convolutional network and is used for marking the ROI via a process of semantic segmentation if the current region is considered by the upper branch to be cancerous.

The feature maps obtained by the sliding window are divided into 28 x 28 patches which are then provided as inputs to the deep convolutional network model, as shown in Fig. 2. Each patch represents one particular region of the feature map. The upper branch in Fig. 2 uses the features in each patch to classify the corresponding region to be either cancerous or normal. For each region judged to be cancerous, the deep convolutional network in the lower branch of Fig. 2 is used to perform semantic segmentation and ROI marking.

Formally, during training, the multi-task loss (L_{tatol}) function is defined as

$$L_{tatol} = (1-\alpha)L_{clf} + \alpha L_{seg}, (0 \le \alpha \le 1) \quad (10)$$

where L_{clf} is the classification loss function of the fully connected branch in Fig. 2, L_{seg} is the segmentation loss function of the segmentation branch in Fig. 2, and α is the segmentation rate parameter, which can be considered as the relative weight of the two loss functions. The larger the α is, the higher accuracy the segmentation branch will achieve, but the higher miss rate the detection branch will encounter. Thus, an extreme large α value may result in very bad detection. Then there will be no regions proposed from the detection module for further segmentation. A very small α even though will give very good detection, the poor segmentation would also ruin the result. In the present study, the parameter is assigned a value of $\alpha = 0.5$ since the detection and segmentation processes are judged to be of equal importance. The classification loss function (L_{clf}) and segmentation loss function (L_{seg}) both represent the average of the cross entropy function. And the output function of the upper branch of fully connected layers, which is for binary detection, is a softmax function, while the probability output function of the lower segmentation branch is a sigmoid function.

For comparison purposes, the classification branch in the integrated model shown in Fig. 2 was implemented using two different deep convolutional network architectures, namely the residual network architecture [15], and the inception module [16]. The details of the residual network and inception module architectures are presented in the following sub-sections.

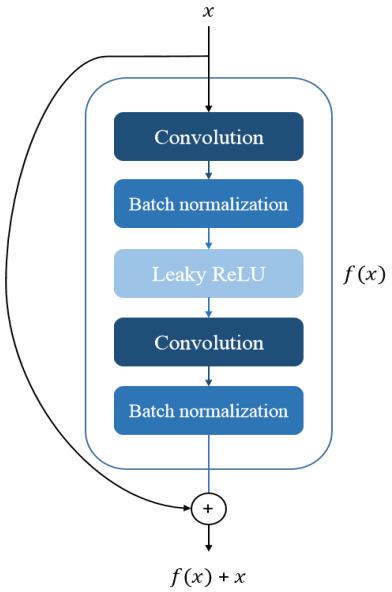

Fig. 3. Residual block

1) Residual Network Architecture

Residual networks [15] are designed to address the problem inherent in conventional convolutional networks of a loss in accuracy as the number of layers increases. In other words, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBCAS.2019.2918244, IEEE Transactions on Biomedical Circuits and Systems

residual network architecture is ideally suited to the implementation of deep convolutional networks.

The fundamental block of residual networks is known as the residual block (or residual unit), which contains the convolution layer, activation function and batch normalization [41]. In the present paper, the residual block has the structure shown in Fig. 3.

We tested on the patch sizes of 25, 28, 32, and 50, and found that patch size of 28 is the best choice. Therefore, in this paper, the input (patch) size is chosen as 28×28. Moreover, the convolution kernel size is set as 3×3 for feature maps of size $\{28, 14, 7\}$. Finally, $2n$ convolution layers are used for each feature map size, where n denotes the number of residual blocks. The total number of output filters is equal to $\{16, 32, 64\}$. Down-sampling is performed by convolution with a stride size of 2. Before the first residual block, a single convolution layer is computed. Finally, the layer before the 2-way fully connected layers is an average pooling layer with a 3×3 kernel size. In evaluating the performance of the proposed model, experiments are performed using various number of layers, where the number of layers is determined by the value assigned to the number of residual blocks, i.e. $n = \{3, 5, 9\}$.

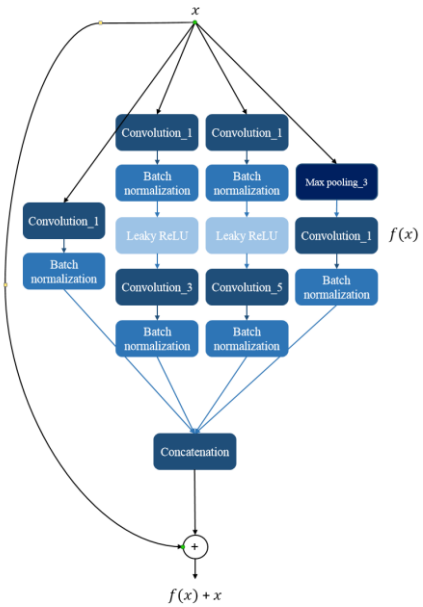

Fig. 4. Inception module.

2) Inception Module Architecture

The segmentation branch in the proposed model was also implemented using an inception module [16]. The basic architecture of the inception module is shown in Fig. 4.

As shown, the inception module combines multiple residual networks to extend the depth of the convolutional network. In particular, the output of the inception module, $f(x)$ is the channel-wise concatenation of four convolution layers. Note that *convolution_ x* and *Max pooling_x* represent convolution and max pooling operations, respectively, with a $x \times x$ kernel size in each case.

The basic principle of the inception module is to determine the optimal convolution structure and construct a deep convolutional network model by increasing the width of the network rather than the depth. Referring to Fig. 4, if convolution kernel sizes $\{1 \times 1, 3 \times 3, 5 \times 5\}$ were used in the same layer, many clusters would be concentrated in each layer and they would be covered by a layer of 1×1 convolutions in the next layer. Thus, in the inception module, 1×1 convolutions before the convolutions with 3×3 and 5×5 kernel sizes are used to reduce the dimensions of the output filter and limit the computational cost.

In this paper, the input to the inception network has the form of 28×28 patches. Moreover, the convolution kernel size is set as $\{1 \times 1, 3 \times 3, 5 \times 5\}$ for feature maps of size $\{28, 14, 7\}$. $2n$ convolution layers are used for each feature map size, where n is the number of inception modules. The total number of output filters is equal to $\{16, 32, 64\}$. As for the residual network architecture, down-sampling is performed by convolution with a stride of 2. Similarly, a convolution layer is computed before the first inception module. Finally, the layer before the 2-way fully connected layers has the form of an average pooling layer with a 3×3 kernel size.

C. Semantic Segmentation

Deep convolutional networks for semantic segmentation contain two processing stages, designated as the encoder and the decoder, respectively. In the deep convolutional network model proposed in the present paper (see Fig. 2), the front-end convolutional network serves as the encoder and learns the features of the input feature map, while the back-end convolutional network serves as the decoder and performs pixel-wise prediction of the ROI. In this paper, two different deep convolutional architectures are employed to perform semantic segmentation, namely a fully convolutional network (FCN) [17] and a feature pyramid network (FPN) [18]. The details of the two architectures are presented in the following sub-sections.

1) Fully Convolutional Network (FCN)

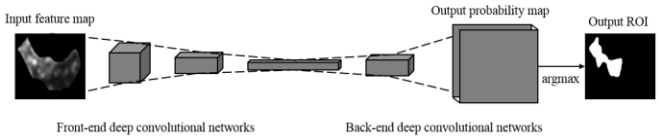

Fig. 5. FCN architecture.

Fully convolutional networks (FCNs) are one of typical deep convolutional network architectures for semantic segmentation. It replaces the fully connected layers in traditional convolution network architectures with convolutional networks to achieve pixel-to-pixel prediction. In other words, all of the layers in the FCN are convolution layers and the output is in image format rather than classification labelling.

In the encoder stage, sub-sampling is performed to learn the useful features for improving the performance. In the decoder stage, up-sampling is performed to reversing the output sizes to corresponding input sizes. The up-sampling process is performed using a deconvolution (i.e., backward convolution) operation. Fig. 5 illustrates the FCN architecture employed in the present paper.

In Fig. 5, the front-end deep convolutional network has the form of the residual block or inception module introduced in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBCAS.2019.2918244, IEEE Transactions on Biomedical Circuits and Systems

Section 3.3, and the back-end deep convolutional network consists of two deconvolution layers with output filters {32,2} and kernel sizes of {3×3,5×5} on feature maps of size {14,28} for each deconvolution layer. The deconvolution operation performed in this paper consists of convolution following up-sampling with nearest bilinear interpolation. In addition, batch normalization is performed between each deconvolution layer and the activation function, where the activation function has the form of a leaky ReLU function.

The output feature map produced by sigmoid function in the last deconvolution layer is a probability feature map with two channels. The first channel shows the probability of the input region belonging to the normal region, while the second channel shows the probability of the input region belonging to the cancerous region. After the sigmoid function, a pixel is labeled to the class of higher output value. Given this probability feature map, the present paper determines the ROI using an argmax function. In particular, the ROI $I(x,y)$ is defined in the form of binary mask obtained from the output probability feature map $P(x,y,z)$ of the FCN as follows:

$$I(x,y) = \arg\max_{z\in\{0,1\}} P(x,y,z), \tag{11}$$

where z has a value of 0 for pixels judged to be normal and a value of 1 for pixels judged to be cancerous.

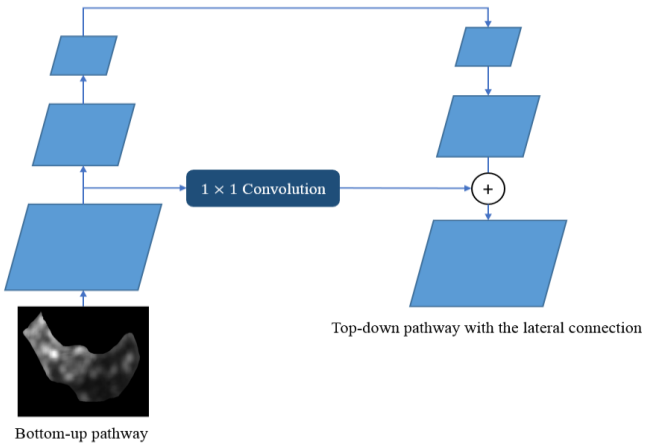

Fig. 6. Bottom-up pathway and top-down pathway with lateral connections.

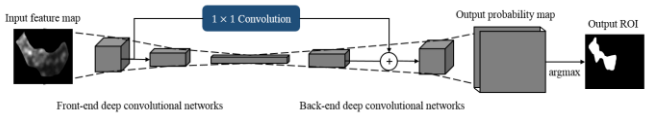

Fig. 7. FPN architecture.

2) Feature Pyramid Network (FPN)

The feature pyramid network was proposed by Lin, Tsung-Yi, et al in [18] and exploits the inherent multi-scale pyramid hierarchy of deep convolutional networks to construct a feature pyramid with strong semantics at all scales. As shown in Fig. 6, the feature pyramid comprises a bottom-up pathway and a top-down pathway with lateral connections.

The bottom-up pathway in the FPN performs a feed forward convolution computation to obtain a hierarchy of feature maps with different scales. By contrast, the top-down pathway performs up-sampling from the highest resolution feature map to obtain a strong semantic feature map [18]. The lateral connections are employed to enhance the quality of the feature maps obtained by the top-down pathway. In particular, the lateral connections facilitate element-wise addition, in which the feature maps obtained following up-sampling in the top-down pathway are merged with the feature maps of the corresponding size in the bottom-up pathway. (Note that a 1×1 convolution computation is performed prior to element-wise addition.) Fig. 7 shows the FPN architecture implemented in the present paper.

The front-end deep convolutional network is constructed by the residual blocks or inception modules and can be regarded as the bottom-up pathway in the FPN. Meanwhile, the back-end deep convolutional network has a symmetric architecture to that of the front-end network and can be considered as the top-down pathway in the FPN. The up-sampling performed in the back-end deep convolutional network is achieved through a process of deconvolution, i.e., convolution following up-sampling with bilinear interpolation. Layers producing feature maps of the same size are said to belong to the same stage. The proposed FPN architecture comprises three stages $\{S_{28}, S_{14}, S_7\}$ with feature map sizes {28,14,7} in the front-end deep convolutional network. The equivalent stages in the back-end deep convolution network have the same feature sizes and parameters and are denoted as $\{S'_{28}, S'_{14}, S'_7\}$. Moreover, the number of output filters in each stage is {16,32,64}. Lateral connections are implemented at stages S'_7 to S'_{14}. As for the case of the FCN network, the ROI is marked using the argmax function.

Table I
BUCCAL MUCOSA IMAGE DATASET USED FOR EVALUATION PURPOSES.

	Training/Validation	Testing	Total
Cancer	25	5	30
Normal	45	5	50

IV. Experimental Results

A. Oral Cavity Auto-fluorescence Image dataset

In the present paper, the evaluation experiments were performed using images in the buccal mucosa sub-set since oral cancer has a higher incidence rate in this particular region of the cavity. As shown in Table I, the experiments thus involved a total of 80 images, consisting of 30 cancerous images and 50 normal images. For each type of image, testing was performed 5 times, and using 5 randomly choosing images in each time, while the remaining images were used for training and validation purposes. In performing the experiments, the training / validation images were partitioned into 180 K patches, of which 160 K patches were chosen for training purposes (80 K for cancer images and 80 K for normal images), and 20 K patches were used for validation purposes (10 K for cancer images and 10 K for normal images). Furthermore, in our experiment, 10 (5 for cancer cases and 5 for normal cases) unseen data are separated out for testing.

B. Evaluation Criterion

The accuracy of the proposed model in performing oral cancer detection was evaluated using two criteria, namely the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBCAS.2019.2918244, IEEE Transactions on Biomedical Circuits and Systems

Table II
SENSITIVITY AND SPECIFICITY WHEN USING RESIDUAL NETWORK (RESNET) AND WAVELET-BASED FEATURE TEXTURE MAPS.

No. of residual blocks		$n=3$		$n=5$		$n=9$		Average	
	Subband	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
resNet	Level 1	0.7274	0.6844	0.6875	0.7835	0.8961	0.6022	0.7703	0.69
+	Level 2	0.7549	0.4765	0.8325	0.4639	0.7677	0.5432	0.785	0.4945
FCN	Level 3	0.9506	0.7247	0.9273	0.607	0.9217	0.7059	0.9332	0.6792
resNet	Level 1	0.6547	0.6781	0.847	0.6161	0.9091	0.6024	0.8063	0.6322
+	Level 2	0.7761	0.6933	0.7632	0.6813	0.7329	0.6913	0.7574	0.6886
FPN	Level 3	0.9742	0.7036	0.9563	0.7171	0.9757	0.7071	0.9687	0.7093

Table III
SENSITIVITY AND SPECIFICITY WHEN USING RESIDUAL NETWORK (RESNET) FRAMEWORKS AND GABOR FILTER-BASED TEXTURE MAPS.

No. of residual blocks		$n=3$		$n=5$		$n=9$		Average	
	Frequency	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
resNet	$\lambda=22.2$	0.9194	0.9358	0.9147	0.9504	0.901	0.9562	0.9106	0.9475
+	$\lambda=16.6$	0.93	0.9361	0.8985	0.9081	0.9078	0.9475	0.9121	0.9306
FCN	$\lambda=10.0$	0.7874	0.8158	0.8551	0.7741	0.7878	0.8043	0.8101	0.8328
resNet	$\lambda=22.2$	0.9223	0.9604	0.9345	0.9017	0.9376	0.9584	0.9314	0.9401
+	$\lambda=16.6$	0.85	0.917	0.8819	0.931	0.8233	0.9473	0.8517	0.9317
FPN	$\lambda=10.0$	0.7757	0.9021	0.7883	0.9038	0.8166	0.8858	0.7935	0.8972

sensitivity (Eq. (12) and the specificity (Eq. (13). Meanwhile, the accuracy of the ROI segmentation results was evaluated using the intersection over union (IOU) criterion.

The sensitivity and specificity criteria were defined respectively as

$$sensitivity(Se.)=\frac{TP}{TP+FN}, \tag{12}$$

$$speficity(Sp.)=\frac{TN}{TN+FP}, \tag{13}$$

where TP denotes true positive, TN denotes true negative, and FP and FN denote false positive and false negative, respectively.

The IOU criterion used to evaluate the accuracy of the ROI segmentation process was defined as where the "segmentation" refer to the binary ROI mask produced by the network model and the "truth" refer to the ROI mask provided by the doctor.

$$IOU=\frac{segmentation\cap truth}{segmentation\cup truth}, \tag{14}$$

C. Results of Oral Cancer Detection

The feature map was partitioned into multiple patches (regions), and these patches were then taken as the input data for the proposed oral cancer detection and ROI segmentation framework. For the considered experimental dataset, the smallest area of the cancerous regions in the cancer images was around 2500 (50×50). Therefore, to enable the patches to be down sampled two times, both the sliding window size and the patch size were set as 28×28. Furthermore, the network model was trained using the Adam optimization method [34] with a learning rate of 10^{-4} and decay rates of $\{0.9, 0.99\}$.

1) Accuracy of Residual Networks

The performance of the residual network model used to perform cancer detection was evaluated for various numbers of layers, where the number of layers is determined by the number of residual blocks, i.e., n={3,5,9}. Furthermore, as described above, the input data had the form of 28×28 patches (regions) of the feature maps constructed using the standard deviation

Table IV
SENSITIVITY AND SPECIFICITY WHEN USING INCEPTION MODULE FRAMEWORKS AND WAVELET-BASED TEXTURE MAPS.

No. of inception modules		$n=3$		$n=5$		$n=9$		Average	
	Subband	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
inception	Level 1	0.7979	0.6534	0.7886	0.6468	0.812	0.6337	0.7995	0.6446
+	Level 2	0.8112	0.6746	0.7602	0.7071	0.7675	0.7344	0.7796	0.7054
FCN	Level 3	0.9568	0.7124	0.9383	0.7386	0.9325	0.7425	0.9426	0.7312
inception	Level 1	0.8334	0.6461	0.8333	0.6559	0.8725	0.6886	0.8464	0.6653
+	Level 2	0.7425	0.7531	0.8355	0.6057	0.8616	0.5692	0.8132	0.6427
FPN	Level 3	0.9451	0.735	0.9747	0.7098	0.9703	0.7108	0.9633	0.7158

Table V
SENSITIVITY AND SPECIFICITY WHEN USING INCEPTION MODULE FRAMEWORKS AND GABOR FILTER-BASED TEXTURE MAPS.

No. of inception modules		$n=3$		$n=5$		$n=9$		Average	
	Frequency	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
inception	$\lambda=22.2$	0.9132	0.967	0.8893	0.9675	0.9072	0.9528	0.9033	0.9624
+	$\lambda=16.6$	0.8492	0.9179	0.8881	0.9615	0.9126	0.9559	0.8833	0.9451
FCN	$\lambda=10.0$	0.7599	0.9063	0.7872	0.8603	0.7862	0.9157	0.7778	0.8941
inception	$\lambda=22.2$	0.9009	0.9517	0.9014	0.9348	0.8815	0.9559	0.8946	0.9475
+	$\lambda=16.6$	0.8445	0.9134	0.8805	0.9091	0.8533	0.9187	0.8594	0.9085
FPN	$\lambda=10.0$	0.7065	0.912	0.7031	0.9187	0.7721	0.9141	0.7272	0.9149

values obtained by the sliding window. And the feature images to which the sliding window was applied were constructed using two different feature extraction methods, namely wavelet transformation and the Gabor filter. The experiments commenced by evaluating the performance of the model when using the features extracted by the wavelet transformation method. The feature extraction process was performed using three different levels of wavelet transformation, where for each level, the vertical and horizontal (two direction) subbands were averaged in order to generate the feature map.

Table II shows the corresponding results obtained for the sensitivity and specificity of the all-residual networks (i.e., residual network + FCN and residual network + FPN) when the wavelet subbands are used to extract features. It is seen that the highest sensitivity and specificity pair resides on about 0.97 and 0.71, respectively, using residual networks + FPN along with subband Level 3.

Table III shows the sensitivity and specificity performance of the all-residual network models when using the Gabor filter with three different frequencies, i.e., λ=22.2,λ=16.6 and λ=10, to extract the texture features. For each frequency, eight different orientations ($\theta=k/8\pi$, $0\leq k\leq 7$) were employed to construct the feature map. It is seen that the highest sensitivity and specificity (both higher than 0.9) are obtained with Gabor filter frequencies of λ=22.2. In other words, the effectiveness of the Gabor filter in extracting useful feature information for classification purposes is enhanced at lower spatial frequencies.

2) Accuracy of Inception Module

This section evaluates the classification performance of the all-inception module frameworks with different numbers of layers ($n=\{3,5,9\}$. As in the previous section, results are presented for feature maps generated using the wavelet transformation method and the Gabor filter method, respectively. Table IV shows the sensitivity and specificity

Table VI

IOU SCORES OBTAINED BY FCN ARCHITECTURES USING WAVELET-BASED TEXTURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Subband	$n=3$	$n=5$	$n=9$	Average
resNet	Level 1	0.687	0.7379	0.5897	0.6715
+	Level 2	0.6939	0.6736	0.7103	0.6926
FCN	Level 3	0.6352	0.6755	0.6196	0.6434
inception	Level 1	0.6574	0.6635	0.6254	0.6488
+	Level 2	0.6592	0.7006	0.7051	0.6883
FCN	Level 3	0.6391	0.6649	0.6547	0.6529

Table VII

IOU SCORES OBTAINED BY FCN ARCHITECTURES USING GABOR FILTER TEXTURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Frequency	$n=3$	$n=5$	$n=9$	Average
resNet	$\lambda=22.2$	0.8061	0.7947	0.8065	0.8024
+	$\lambda=16.6$	0.8059	0.8039	0.8106	0.8068
FCN	$\lambda=10.0$	0.7841	0.7083	0.7756	0.756
inception	$\lambda=22.2$	0.8301	0.8204	0.8191	0.8232
+	$\lambda=16.6$	0.8042	0.8123	0.811	0.8127
FCN	$\lambda=10.0$	0.8088	0.7549	0.8125	0.7921

Table VIII

IOU SCORES OBTAINED BY FPN ARCHITECTURES USING WAVELET-BASED TEXTURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Subband	$n=3$	$n=5$	$n=9$	Average
resNet	Level 1	0.669	0.6275	0.6199	0.6388
+	Level 2	0.6837	0.6816	0.6982	0.6878
FPN	Level 3	0.6405	0.644	0.6354	0.64
inception	Level 1	0.6339	0.6385	0.642	0.6381
+	Level 2	0.6921	0.6541	0.6739	0.6733
FPN	Level 3	0.6537	0.6386	0.645	0.6457

Table IX

IOU SCORES OBTAINED BY FPN ARCHITECTURES USING GABOR FILTER FEATURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Frequency	$n=3$	$n=5$	$n=9$	Average
resNet	$\lambda=22.2$	0.8261	0.8061	0.8171	0.8164
+	$\lambda=16.6$	0.8072	0.8105	0.8156	0.8111
FPN	$\lambda=10.0$	0.8016	0.7878	0.7982	0.7959
inception	$\lambda=22.2$	0.8099	0.802	0.8175	0.8098
+	$\lambda=16.6$	0.8108	0.8135	0.8173	0.8139
FPN	$\lambda=10.0$	0.7652	0.7792	0.8051	0.7832

results obtained when using the wavelet-based feature maps. As for the all-inception network models, the highest average sensitivity and specificity pair are about 0.97 and 0.71, respectively, that are obtained when using the third level wavelet subband features.

Table V shows the sensitivity and specificity of the two inception module-based networks when using the Gabor filter to extract the texture features. It is again observed that the Gabor filter feature maps with low spatial frequency ($\lambda = 22.2$) result in the highest average sensitivity and specificity (0.9033 and 0.9624).

Comparing the classification results obtained using the wavelet transform and Gabor filter feature maps, respectively, it is found that the Gabor filter feature maps result in a much higher maximum average specificity than the wavelet transform feature maps. However, the two feature maps yield a similar sensitivity. In other words, the texture features extracted by the Gabor filter at different spatial frequencies and multiple

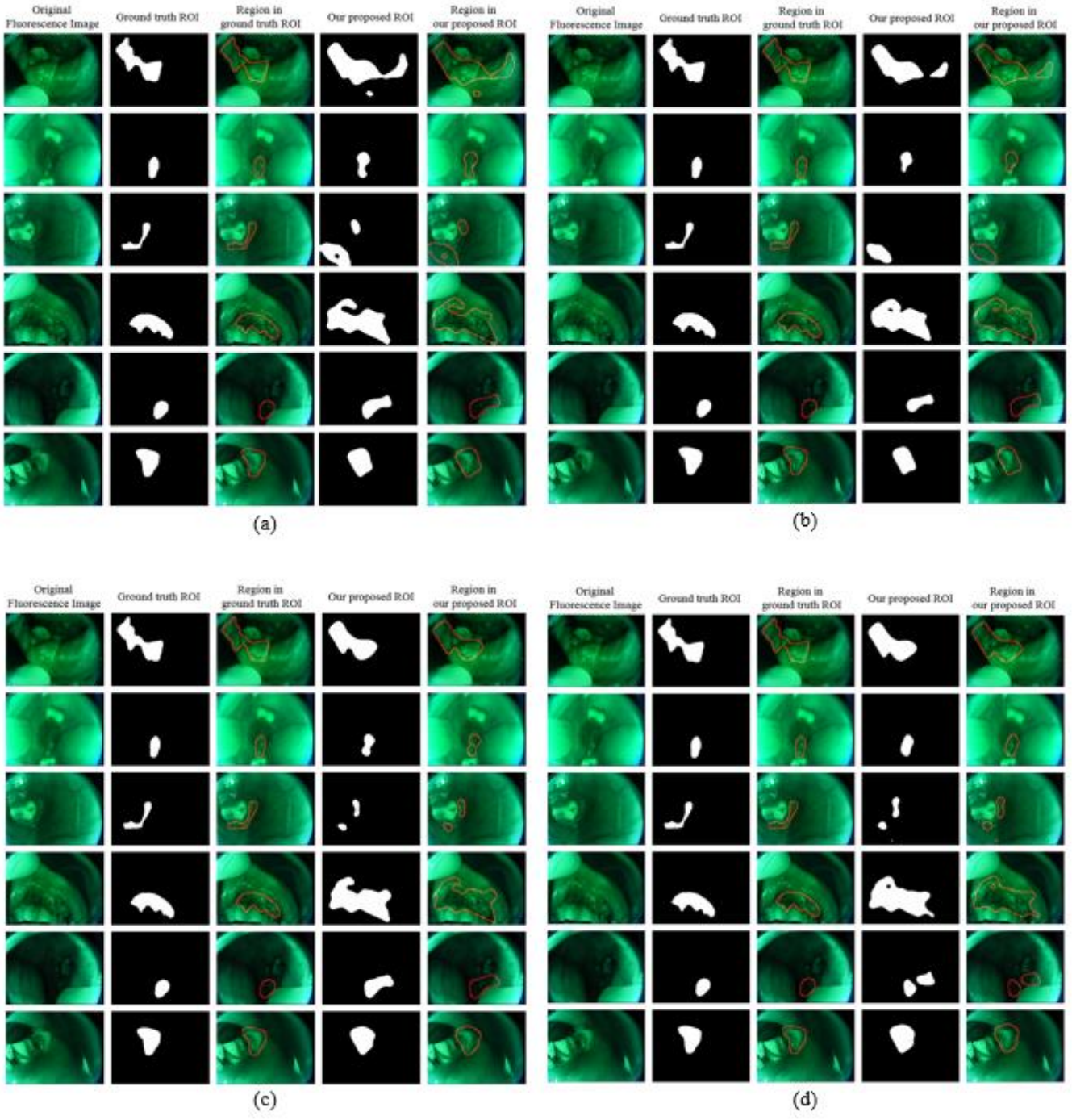

Fig. 8. Cancerous ROIs marked by the resNet + FCN and inception module + FCN architectures. (a) is ROIs marked by FCN architecture based on third level wavelet subband features and resNet ($n = 5$). (b) is ROIs marked by FCN architecture based on third level wavelet subband features and inception module ($n = 5$). (c) is ROIs marked by FCN architecture based on Gabor filter features obtained at frequency of $\lambda = 22.2$ and resNet ($n = 3$). (d) is ROIs marked by FCN architecture based on Gabor filter features obtained at frequency of $\lambda = 22.2$ and inception module ($n = 5$).

orientations are more useful for oral cancer detection purposes than those obtained using the wavelet transform method using high-pass and low-pass filters with only two directions. Moreover, comparing the performances of the residual network and inception module networks, it is observed that the two architectures achieve a similar sensitivity. Finally, the results suggest that when constructing the back-end deep convolutional network, the choice of architecture (i.e., FCN or FPN) has only a minor effect on the detection accuracy.

D. Results of Marking ROI

This section evaluates the ROI marking performance of the proposed framework when using the FCN and FPN architectures, respectively. As in the previous section, the results are compared for both the wavelet transformation feature extraction method and the Gabor filter feature extraction method.

1) Accuracy of FCN

Table VI and Table VII show the IOU scores when the

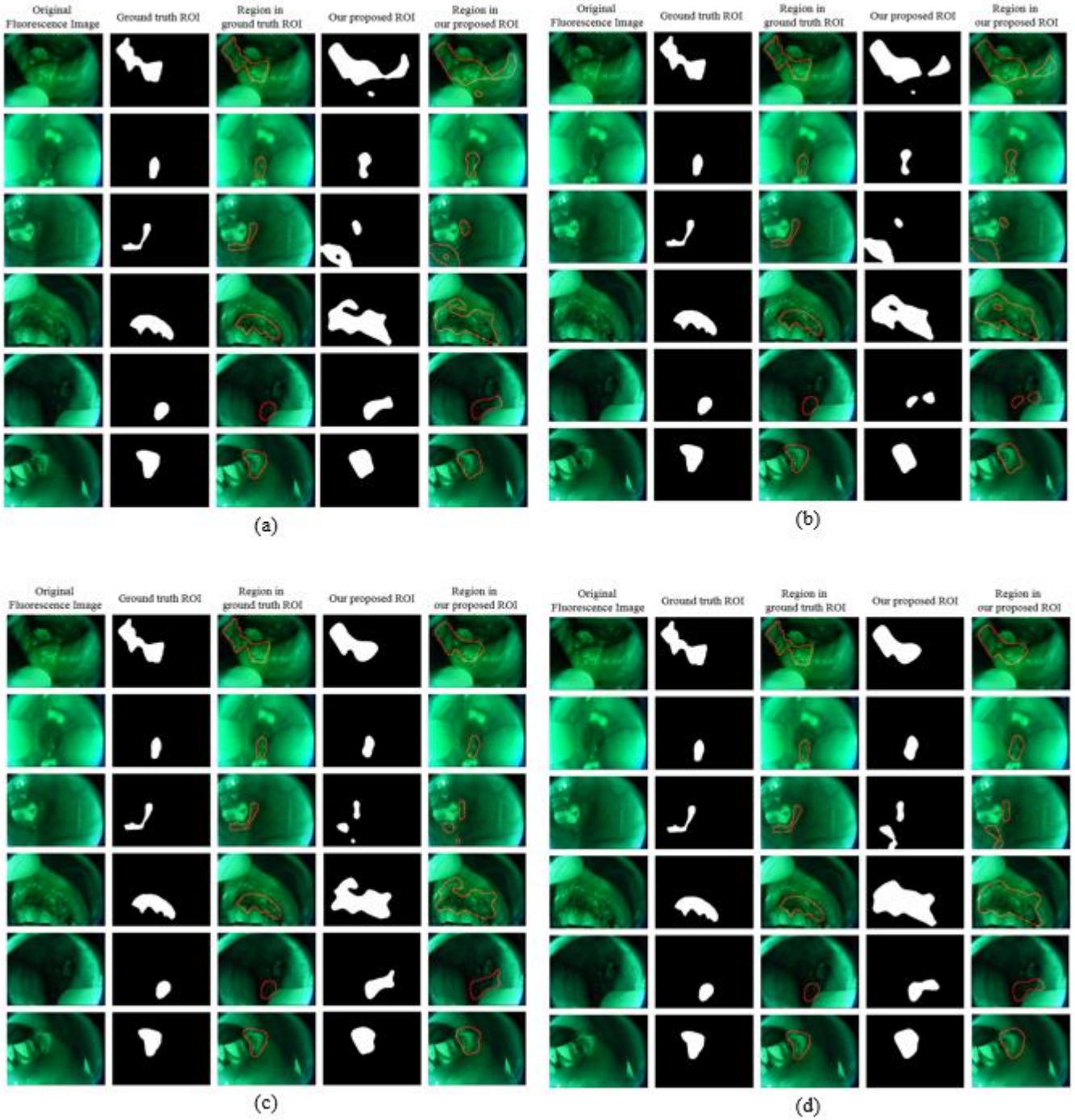

Fig. 9. Cancerous ROIs marked by the resNet + FPN and inception module + FPN architectures. (a) is ROIs marked by FPN architecture based on third level wavelet subband features and resNet ($n = 3$). (b) is ROIs marked by FPN architecture based on third level wavelet subband features and inception module ($n = 3$). (c) is ROIs marked by FPN architecture based on Gabor filter features with frequency of $\lambda = 22.2$ and resNet ($n = 9$). (d) is ROIs marked by FPN architecture based on Gabor filter features with frequency of $\lambda = 22.2$ and inception module ($n = 3$).

wavelet texture map and the Gabor texture map, respectively, are used and the front-end network is implemented using a resNet + FCN or an inception module + FCN. The results indicate that Gabor texture map reveals dramatically higher performance than the wavelet texture map. The results presented in Table VII for the case where the feature maps are generated using the Gabor filter method, show that the highest average IOUs of the resNet and inception module architectures are 0.8068 and 0.8232, respectively, and are obtained using a filter frequency of $\lambda = 16.6$ and $\lambda = 22.2$. However, it is worth to mention that the two frequencies ($\lambda = 16.6$ and $\lambda = 22.2$) reveal comparable performances in terms of IOU marking.

Fig. 8 shows the cancerous ROIs marked by the resNet + FCN and inception module + FCN architectures following median filter processing when using the two different feature extraction methods. In each figure, the first column shows the original auto-fluorescence images, the second column shows the ground truth ROIs marked by a doctor, the third column shows the outline of the ground truth ROIs overlaid on the image, the fourth column shows the ROIs marked by the proposed framework, and the fifth column shows the outline of the ROIs marked by the proposed framework overlaid on the image. Fig. 8 (a) and (b) show the results obtained when using the third level wavelet subband features and the resNet and inception module architectures with $n = 5$ in both cases. Fig. 8 (c) and (d) show the corresponding results obtained when using the Gabor filter features extracted with a frequency of $\lambda = 22.2$ and the resNet and inception module architectures with $n = 3$ and $n = 5$, respectively. Comparing the four figures, it is seen that the ROIs obtained using the Gabor filter features are in better agreement with the ground truth ROIs than those obtained using the wavelet transform features. In other words, the texture features extracted with different spatial frequencies and in multiple orientations provide a more reliable basis for ROI segmentation and marking when using the FCN architecture.

2) Accuracy of FPN

This section evaluates the ROI marking accuracy of the proposed framework based on an FPN architecture. Table VIII

Table X

IOU SCORES OBTAINED BY PROPOSED FRAMEWORK WITHOUT SEGMENTATION BRANCH USING WAVELET-BASED TEXTURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Subband	$n = 3$	$n = 5$	$n = 9$	Average
resNet	Level 1	0.5619	0.6309	0.476	0.5562
	Level 2	0.5371	0.5261	0.5564	0.5398
	Level 3	0.5671	0.5675	0.5812	0.5719
inception	Level 1	0.5299	0.5298	0.5155	0.5251
	Level 2	0.516	0.5406	0.5596	0.5387
	Level 3	0.5974	0.6556	0.6052	0.6194

Table XI

IOU SCORES OBTAINED BY PROPOSED FRAMEWORK WITHOUT SEGMENTATION BRANCH USING GABOR FILTER TEXTURE MAPS WHEN DIFFERENT OF NUMBER OF RESIDUAL BLOCKS OR INCEPTION MODULES (n ARE USED).

	Frequency	$n = 3$	$n = 5$	$n = 9$	Average
resNet	$\lambda = 22.2$	0.8109	0.7935	0.7825	0.7956
	$\lambda = 16.6$	0.7830	0.7784	0.793	0.7848
	$\lambda = 10.0$	0.7329	0.7273	0.7105	0.7236
inception	$\lambda = 22.2$	0.7948	0.8206	0.8104	0.8086
	$\lambda = 16.6$	0.789	0.7795	0.7748	0.7811
	$\lambda = 10.0$	0.7201	0.7589	0.7631	0.7474

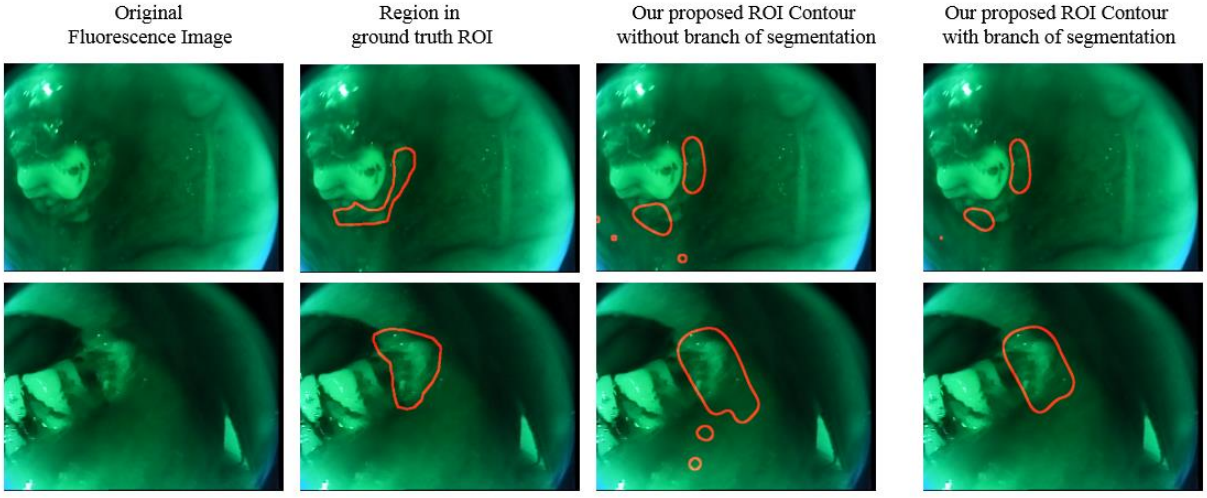

Fig. 10. ROI marking performance of proposed framework with and without segmentation branch.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBCAS.2019.2918244, IEEE Transactions on Biomedical Circuits and Systems

and Table IX show the IOU results obtained when using feature maps constructed using the wavelet transformation and Gabor filter methods, respectively. Similar to a conclusion in FCN, Gabor filter extraction method demonstrates superior performance to the wavelet feature extraction method in ROI marking. Comparing Table VII and Table IX, it is additionally concluded that no significant difference in IOU scores is observed using either FCN or FPN architecture.

Fig. 9 shows the ROI marking results obtained by the resNet + FPN and inception module + FPN architectures based on the wavelet transform features and Gabor filter features, respectively. In particular Fig. 9 (a) and (b) show the results obtained using the third level wavelet subband features with $n = 3$ in the resNet network and $n = 3$ in the inception module network, respectively. Similarly, Fig. 9 (c) and (d) show the results obtained using the features extracted by the Gabor filter with a frequency of $\lambda = 22.2$ with $n = 9$ and $n = 3$ in the resNet and inception module architectures, respectively.

E. Comparison and Discussion

This section commences by comparing the ROI marking results if the segmentation branch is not used. Table X and Table XI show the IOU scores obtained by the proposed deep convolutional network framework using the wavelet transform features and Gabor filter features, respectively, when the segmentation branch is not implemented. The results confirm that for both feature extraction methods, the omission of the segmentation branch results in a poor IOU performance (compared with Table VI to Table IX). In other words, the segmentation branch greatly improves the ROI marking ability of the proposed deep convolutional model. Fig. 10 shows the ROIs marked by the proposed network model with and without the segmentation branch, respectively. It is seen that while some normal regions are regarded as cancerous regions by the

Table XII
SENSITIVITY AND SPECIFICITY WHEN USING RESNET NETWORK MODELS AND WAVELET TRANSFORM TEXTURES MAPS. (PROBABILITY OF CANCER GREATER THAN 0.4 IS USED AS CRITERION IN DETERMINING CANCER.)

No. of residual blocks		$n = 3$		$n = 5$		$n = 9$		Average	
	Subband	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
resNet	Level 1	0.7761	0.533	0.7498	0.6514	0.955	0.4152	0.827	0.5332
+	Level 2	0.9519	0.4388	0.9339	0.4369	0.9576	0.4715	0.9478	0.4491
FCN	Level 3	0.9833	0.5264	0.9881	0.4876	0.9413	0.5318	0.9709	0.5153
resNet	Level 1	0.7964	0.515	0.9652	0.3839	0.9637	0.3891	0.9084	0.4293
+	Level 2	0.9585	0.477	0.9688	0.4199	0.9462	0.4516	0.9578	0.4495
FPN	Level 3	0.9774	0.4885	0.9901	0.5024	0.9836	0.5013	0.9837	0.4947

Table XIII
SENSITIVITY AND SPECIFICITY WHEN USING RESNET NETWORK MODELS AND GABOR FILTER TEXTURES MAPS. (PROBABILITY OF CANCER GREATER THAN 0.4 IS USED AS CRITERION IN DETERMINING CANCER.)

No. of residual blocks		$n = 3$		$n = 5$		$n = 9$		Average	
	Frequency	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
resNet	$\lambda = 22.2$	0.9892	0.8457	0.9839	0.8732	0.9905	0.8504	0.9879	0.8564
+	$\lambda = 16.6$	0.997	0.7939	0.9837	0.8582	0.9984	0.8503	0.993	0.8341
FCN	$\lambda = 10.0$	0.9392	0.8188	0.9527	0.773	0.9751	0.7441	0.9557	0.7786
resNet	$\lambda = 22.2$	0.9895	0.8866	0.9861	0.8586	0.9862	0.8744	0.9872	0.8732
+	$\lambda = 16.6$	0.9874	0.8604	0.9857	0.8567	0.9999	0.7449	0.991	0.8207
FPN	$\lambda = 10.0$	0.9636	0.785	0.9756	0.7382	0.9516	0.7689	0.9636	0.764

Table XIV
SENSITIVITY AND SPECIFICITY WHEN USING INCEPTION MODULE NETWORKS USING WAVELET TRANSFORM TEXTURES. (PROBABILITY OF CANCER GREATER THAN 0.4 IS USED AS CRITERION IN DETERMINING CANCER.)

No. of inception modules		$n = 3$		$n = 5$		$n = 9$		Average	
	Subband	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
inception	Level 1	0.9169	0.4637	0.877	0.4644	0.8795	0.4543	0.8911	0.4609
+	Level 2	0.9744	0.4088	0.9481	0.4487	0.9014	0.4827	0.9413	0.4467
FCN	Level 3	0.9513	0.5197	0.9689	0.5698	0.9891	0.5933	0.9697	0.5609
inception	Level 1	0.9643	0.4085	0.9186	0.4586	0.9016	0.5747	0.9282	0.4806
+	Level 2	0.9319	0.5326	0.9818	0.3935	0.9827	0.3864	0.9655	0.4375
FPN	Level 3	0.9668	0.577	0.9793	0.5194	0.9917	0.5091	0.9792	0.5352

Table XV
SENSITIVITY AND SPECIFICITY WHEN USING INCEPTION MODULE NETWORKS AND GABOR FILTER TEXTURES. (PROBABILITY OF CANCER GREATER THAN 0.4 IS USED AS CRITERION IN DETERMINING CANCER.)

No. of inception modules		$n = 3$		$n = 5$		$n = 9$		Average	
	Frequency	Se.	Sp.	Se.	Sp.	Se.	Sp.	Se.	Sp.
inception	$\lambda = 22.2$	0.9884	0.8864	0.9755	0.9122	0.9942	0.8394	0.986	0.8793
+	$\lambda = 16.6$	0.9652	0.8981	0.9906	0.879	0.9895	0.8731	0.9817	0.8662
FCN	$\lambda = 10.0$	0.9674	0.7699	0.9462	0.8131	0.9711	0.7599	0.9616	0.781
inception	$\lambda = 22.2$	0.9819	0.8716	0.9856	0.9063	0.9779	0.8942	0.9818	0.8907
+	$\lambda = 16.6$	0.9721	0.861	0.991	0.8385	0.979	0.8908	0.9807	0.8634
FPN	$\lambda = 10.0$	0.9698	0.7411	0.9626	0.7928	0.948	0.8579	0.9601	0.7973

detection branch, these regions are not denoted as the ROI by the segmentation branch.

In clinical diagnosis, it is generally necessary to make a tradeoff between sensitivity and specificity since, in the case of erroneous diagnosis, it is preferable to identify normal regions as cancerous regions, rather than cancerous regions as normal regions. Accordingly, a series of experiments are performed in which the criterion for cancer diagnosis used in the detection process was specified as an output probability of cancer greater than 0.4 (the default setting used in the previous experiments is 0.5).

Table XII and Table XIII show the resulting sensitivity and specificity of the residual networks given the use of the wavelet transformation features and Gabor filter features, respectively. It can be shown that the maximum sensitivity can achieve around 0.99, for both feature maps used in the detection process. Furthermore, the Gabor filter features present superior performance, where the specificity maintains a value above 0.85. Table XIV and Table XV present the sensitivity and specificity of the inception networks given the use of the wavelet transformation features and Gabor filter features, respectively. Comparing Table XII and Table XIV, and Table XIII and Table XV, it is obviously observed that the inception networks present similar performances as the residual networks. It was also observed that given the less crucial criterion for the threshold set on 0.4, the relatively high frequency Gabor filter of $\lambda = 16.6$ achieves comparable performances as the $\lambda = 22.2$. Also, the use of Gabor filter features again results in a superior performance compared to wavelet features achieving both good sensitivity and high specificity.

V. Conclusions

This paper has presented a model for the detection and ROI marking of cancer in the buccal region of the oral cavity using a texture-map based branch-collaborative network based on texture feature images. In the proposed model, wavelet transformation and the Gabor filtering method have been used to extract the texture feature images. A sliding window has then been applied to compute the corresponding standard deviation values. Finally, the standard deviation values have been used to construct a feature map, which is partitioned into multiple patches and taken as the input to the deep convolutional network model. The network model comprises two integrated branches, namely an oral cancer detection branch and a ROI marking branch. The oral cancer detection branch has been implemented using two different architectures, namely a residual network model and an inception module. Similarly, the ROI segmentation and marking branch has been implemented using both a fully convolutional network (FCN) architecture and a feature pyramid network (FPN) architecture. The experimental results have shown that the features extracted by the Gabor filter provide more useful information for the cancer detection and ROI marking tasks than those obtained using the wavelet transformation method. Furthermore, the FCN and FPN architectures result in a similar ROI marking performance. The results have confirmed that the segmentation branch greatly improves the ROI marking ability of the proposed model. The test set in each experiment is randomly selected. From our experiments, an average of 70% images have IOU score (which is one way to measure the boundary match accuracy) greater than 0.75. It has been shown that an acceptable tradeoff can be achieved between the sensitivity and specificity of the proposed model by easing the probability criterion employed in the detection branch to detect cancer. The results have confirmed that the model provides a good ability to mark the high-risk regions automatically, and hence provides a useful tool for oral cancer screening.

References

[1] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359-E386, Mar. 2015.

[2] C.S. Farah *et al.*, "Advances in Early Detection and Diagnostic Adjuncts in Oral Cavity Cancer," in *Contemporary Oral Oncology*, Springer, Cham, 2017, pp. 355-421.

[3] C.-J. Chiang *et al.*, "Incidence and survival of adult cancer patients in Taiwan, 2002-2012," *Journal of the Formosan Medical Association*, vol. 115, no. 12, pp. 1076–1088, Dec. 2016.

[4] K. Thankappan and M.A. Kuriakose, "Sentinel Node Biopsy in Oral Cancer," in *Contemporary Oral Oncology*, Springer, Cham, 2017, pp. 211-233.

[5] T.W. Remmerbach *et al*., "Liquid-based versus conventional cytology of oral brush biopsies: a split-sample pilot study," *Clinical Oral Investigations*, vol. 21, no. 8, pp. 2493-2498, Nov. 2017.

[6] X. Wang, K.E. Kaczor-Urbanowicz and D.T.W. Wong, "Salivary biomarkers in cancer detection," *Medical Oncology*, vol. 34, no.1, pp. 7, Jan. 2017.

[7] Y.-W. Chen *et al.*, "Use of methylene blue as a diagnostic aid in early detection of oral cancer and precancerous lesions," *British Journal of Oral and Maxillofacial Surgery*, vol. 45, no.7, pp. 590-591, Oct. 2007.

[8] R.A. Schwarz *et al.*, "Autofluorescence and diffuse reflectance spectroscopy of oral epithelial tissue using a depth-sensitive fiber-optic probe," *Applied optics*, vol. 47, no. 6, pp. 825-834, Feb. 2008.

[9] D.C.G. De Veld *et al.*, "Autofluorescence and diffuse reflectance spectroscopy for oral oncology," *Lasers in Surgery and Medicine: The Official Journal of the American Society for Laser Medicine and Surgery*, vol. 36, no. 5, pp. 356-364, Jun. 2005.

[10] D.C.G. de Veld *et al.*, "Clinical study for classification of benign, dysplastic, and malignant oral lesions using autofluorescence spectroscopy," *Journal of biomedical optics*, vol. 9, no. 5, pp. 940-951, Sep. 2004.

[11] K. Alhallak *et al.*, "Optical redox ratio identifies metastatic potential-dependent changes in breast cancer cell metabolism," *Biomedical Optics Express*, vol. 7, no. 11, pp. 4364-4374, Oct. 2016.

[12] M. Ciccìù *et al.*, "Tissue Fluorescence Imaging (VELscope) for Quick Non-Invasive Diagnosis in Oral Pathology," *Journal of Craniofacial Surgery*, vol. 28, no. 2, pp. e112-e115, Mar. 2017.

[13] T.-T. Huang *et al.*, "Two-channel autofluorescence analysis for oral cancer," *Journal of Biomedical Optics*, vol. 24, no. 5, pp. 051402, Nov. 2018.

[14] T.-T. Huang *et al.*, "Novel quantitative analysis of autofluorescence images for oral cancer screening," *Oral oncology*, vol. 68, pp. 20-26, May. 2017.

[15] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[16] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[17] J. Long, E. Shelhamer and T. Darrell "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[18] T.-Y. Lin *et al.*, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117-2125.

[19] K. He *et al.*, "Mask r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2961-2969.
[20] D.K. Iakovidis, E.G. Keramidas and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," in *International conference image analysis and recognition*, 2008, pp. 750-759.
[21] P. Kruizinga, N. Petkov, "Nonlinear operator for oriented texture," *IEEE Transactions on image processing*, vol. 8, no. 10, pp. 1395-1407, Oct. 1999.
[22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767 [cs.CV]*, Apr. 2018.
[23] H. Noh, S. Hong and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision.*, 2015, pp. 1520-1528.
[24] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.
[25] R Girshick *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.
[26] K. He *et al.*, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904-1916, Sep. 2015.
[27] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.
[28] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.
[29] O. Ronneberger, P. Fischer and T. Brox "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241.
[30] S. Iizuka, E. Simo-Serra and H. Ishikawa, "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 110, Jul. 2016.
[31] A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
[32] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs.CV]*, Apr. 2015.
[33] G.E. Hinton *et al.*, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580 [cs.NE]*, Jul. 2012.
[34] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs.LG]*, Jan. 2017.
[35] M.M. Galloway, "Texture analysis using grey level run lengths," *NASA STI/Recon Technical Report N*, vol.75, Jul. 1974.
[36] T. Ojala, M. Pietikäinen and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol.24, no. 7, pp. 971-987, Jul. 2002.
[37] T. Ahonen, A. Hadid and M. Pietikainen, "Face recognition with local binary patterns," in *European conference on computer vision*, Berlin, Heidelberg, 2004, pp. 469-481.
[38] D.K. Iakovidis, E.G. Keramidas and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," in *International conference image analysis and recognition*, Berlin, Heidelberg, 2008, pp. 750-759.
[39] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern recognition letters*, vol. 24, no. 9-10, pp. 1513-1521, Jun. 2003.
[40] K. Huang and S. Aviyente, "Wavelet Feature Selection for Image Classification," *IEEE Transactions on Image Processing*, vol. 17, no. 9, pp. 1709-1720, Sep. 2008.
[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167 [cs.LG]*, Mar. 2015.

Chih-Hung Chan received the M.S. degree in the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2017. Since 2017, he has been an R&D Engineer in NVIDIA, Taiwan.

Pau-Choo (Julia) Chung (S'89-M'91-SM'02-F'08) received the Ph.D. degree in Department of Electrical Engineering from Texas Tech University, USA, in 1991. She then joined the Department of Electrical Engineering, National Cheng Kung University (NCKU), Taiwan, in 1991 and has become a full professor in 1996. She served as the Head of Department of Electrical Engineering (2011-2014), the Director of Institute of Computer and Communication Engineering (2008-2011), the Vice Dean of College of Electrical Engineering and Computer Science (2011), the Director of the Center for Research of E-life Digital Technology (2005-2008), and the Director of Electrical Laboratory (2005-2008), NCKU. She was elected Distinguished Professor of NCKU in 2005 and received the Distinguished Professor Award of Chinese Institute of Electrical Engineering in 2012. She also served as Program Director of Intelligent Computing Division, Ministry of Science and Technology (2012-2014), Taiwan, and the Director General of the Department of Information and Technology Education, Ministry of Education, Taiwan.

Chih-Yang Chen received the M.S. and Ph.D. degrees in Department of Electrical Engineering from National Cheng Kung University, Tainan, Taiwan, in 2003 and 2008, respectively. Since 2010, he has been an R&D Engineer in Delta Electronics, Taiwan. His research interests include machine learning, adaptive fuzzy control, medical and robotic applications.

Chien-Cheng Lee received the Ph.D. degree in Department of Electrical Engineering from National Cheng Kung University, Tainan, Taiwan in 2003. Dr. Lee is currently an Assistant Professor in the Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan. He has been a research visitor at the Telcordia Inc. (formerly Bellcore), NJ, USA, from Oct. 2007 to Jan. 2008. He is one of the guest editors for a special issue on Signal Processing for Applications in Healthcare Systems for EURASIP Journal on Advances in Signal Processing, 2008. His research interests include image processing, pattern recognition, and machine learning.

Man-Yee Chan BDS (NDMC), MSc (Manchester, UK), EMBA (THU), PhD (CSMU). He is presently the Vice Chief of the Department of Stomatology at Taichung Veterans General Hospital. He was the former chief of Oral & Maxillofacial Surgery. His main specialty fields include the treatment of oral cancer, Orthognathic Surgery, Facial trauma surgery and Implant surgeries.

Tze-Ta Huang received the Ph.D. degrees in Molecular biological research from National Chung Chen University, Tainan, Taiwan, in 2013. Since 2014, he has been an Assistant Professor and Attending physician in Department of Stomatology and Institute of Oral Medicine, College of Medicine, National Cheng Kung University, Taiwan.