Automatic Classification and Pathological Staging of Confocal Laser Endomicroscopic Images of the Vocal Cords

Kim Vo[1], Christian Jaremenko[1], Christopher Bohr[2], Helmut Neumann[3], Andreas Maier[1]

[1]Pattern Recognition Lab, Department of Computer Science, FAU Erlangen-Nürnberg, Germany
[2]Department of Otorhinolaryngology - Head and Neck Surgery, University Hospital Erlangen
[3]Department of Medicine 1, University Hospital Mainz
Kim_Vo@web.de

Abstract. Confocal laser endomicroscopy is a novel imaging technique which provides real-time in vivo examination and histological analysis of tissue during an ongoing endoscopy. We present an automatic classification system that is able to differentiate between healthy and cancerous tissue of the vocal cords. Textural as well as CNN features are encoded using Fisher vectors and Vector of Locally Aggregated Descriptors while the classification is performed using random forests and support vector machines. Two experiments are investigated following a leave-one-sequence-out cross-validation and a fixed training and test set approach. Classification rates reach up to 87.6 % and 81.5 %, respectively.

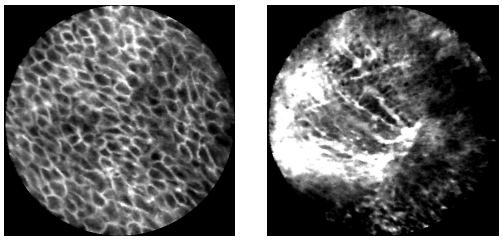
1 Introduction

Head and neck cancer is a collective term that comprises cancers of the upper aerodigestive tract including laryngeal cancer. Over 90% of laryngeal cancers are squamous cell carcinomas whereof half involve the vocal cords. To date, the standard of care for the diagnosis of laryngeal cancer is white light examination followed by biopsies and histopathology of suspicious lesions to confirm malignancy. These treatments are time consuming and resections may lead to permanent voice disorders. Moreover the accuracy of the diagnosis is highly dependent on the experience of the surgeon, the pathologist and the quality of biopsy. Recently a novel optical imaging method called confocal laser endomicroscopy (CLE) has been proposed, allowing subsurface analysis of the epithelium in real time and thus enables optical histology during ongoing endoscopy. In order to acquire high-contrast visualization of the surface epithelium, contrast agents such as fluorescein is administered intravenously to stain the cellular architecture and extracellular matrix. Thus, gained images allow the comparison between healthy epithelium and malignant lesions.
CLE has been successfully applied in gastroenterology and was recently introduced in the context of head and neck cancer. To assist the decisions of the

surgeon during an ongoing endoscopy, several approaches for the automatic detection and classification of healthy and cancerous tissue exist. For example Dittberner et al. [1] propose an automated image analysis algorithm for the classification of head and neck cancer using distance map histograms. Another approach, introduced by Jaremenko et al. [2], uses various textural features for the automatic classification of CLE images of the oral cavity.
This paper presents a bag of words (BoW) approach, based on the framework of [2], to differentiate between images of healthy and cancerous tissue using Vector of Locally Aggregated Descriptors (VLAD) and Fisher vectors (FV). Additional textural features are evaluated and since features extracted from convolutional neural networks (CNN) arise to be a strong competitor to the state-of-the-art methods in image classification [3] their performance is compared with the textural features.

2 Materials and Methods

In this study, 45 video sequences from 5 patients were obtained using a probe-based CLE (pCLE) system from Cellvizio (UHD GastroFlex, Mauna Kea Technologies, Paris, France). These sequences are separated into single images leading to a database consisting of 1767 physiological images and 2675 images containing carcinoma. The images were labeled by an expert of the University Hospital Erlangen, Germany. While images of healthy epithelium show flat and relatively uniform scale-like cells with alternating bright and dark bands, images of carcinoma show a completely disorganized cell structure with fluorescein leakage as visualized in Fig. 1.



(a) Healthy (b) Cancerous

Fig. 1. Examples of pCLE images of healthy and cancerous squamous epithelium

2.1 Features

Following the pre-processing step proposed by Jaremenko et al. [2], features are extracted from small rectangular patches with an edge length of 105 pixels and 50 % overlap. From each of the image patches, Histogram of Oriented Gradients

(HoG) [4], Gray level co-occurrence matrices (GLCM) [5], Local binary patterns (LBP) [6], Local derivative patterns (LDP) [7] and CNN features are extracted as local descriptors. In [2] the average and standard deviation of each feature over all patches is used to describe each image, whereas here the concatenation of all patch descriptors depicts each image.

2.2 Convolutional neural network model design and training

For the extraction of the CNN features, a CNN architecture based on the LeNet-5 network [8] is used. LeNet-5 consists of a convolutional layer followed by a max-pool layer, another convolutional layer followed by a max-pool layer and two consecutive fully connected layers. Additionally a fully connected layer is added and the sigmoid activations are replaced by Rectified Linear Unit (ReLU) activations. The network is trained on a training set consisting of 154440 image patches (2790 images) and 2 classes. The weights are updated by stochastic gradient descent, accompanied by momentum term of 0.9 and the learning rate is set to 0.0005 for all epochs.
Data augmentation is used to align the distribution of both classes of the original training set and to increase its variance. For this purpose, the CLE images are rotated arbitrarily and additional patches are extracted. Following this procedure, the training set is increased to 374972 image patches (7211 images).

2.3 Bag of words framework

The BoW model requires the construction of a visual codebook based on k-means clustering of features extracted from training images. The codebook consists of a set of visual words (cluster centers) which is used to compute a histogram of visual word frequencies to encode a given image.
FV and VLAD have shown to outperform the classical BoW model in the context of image classification. In this study, both methods are used to encode the image features proposed in chapter 2.1, followed by a classification step using support vector machines (SVM) and random forests (RF).

Fisher vector encoding [9] uses a Gaussian mixture model (GMM) as a generative model, where the parameters of the K components can be denoted as $\lambda = \{(\omega_k, \mu_k, \Sigma_k), k = 1, 2..., K\}$, where ω_k, μ_k and Σ_k are the mixture weight, mean and covariance matrix of the k-th component learned from a training set, respectively. Given a feature vector $X = \{x_1, ..., x_T\}$ extracted from an image, the gradients of the FV with respect to the weight parameters, mean and standard deviation can be computed with following equations:

$$\mathcal{G}^X_{\alpha_k} = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} (\gamma_t(k) - w_k) \tag{1}$$

$$\mathcal{G}^X_{\mu_k} = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) (\frac{x_t - \mu_k}{\sigma_k}) \tag{2}$$

$$\mathcal{G}^X_{\sigma_k} = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \frac{1}{\sqrt{2}} [\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1] \tag{3}$$

where $\gamma_t(k)$ is the posterior probability. By concatenating $\mathcal{G}^X_{\alpha_k}$, $\mathcal{G}^X_{\mu_k}$ and $\mathcal{G}^X_{\sigma_k}$ for all K components, the final FV of the image is obtained with size $(2 \times D + 1)K$, where D is the dimension of the local feature vectors x. Subsequently, ℓ_2-normalization and power normalization of the form $f(z) = sign(z)|z|^\alpha$ is applied to improve the performance of Fisher vectors.

VLAD encoding [10] is a simplification of the FV encoding. A codebook $\{\mu_1, \mu_2, ..., \mu_K\}$ is generated by using k-means. The VLAD descriptor for each μ_k can be computed by accumulating the differences $x - \mu_k$, where x is the image feature having μ_k as its nearest cluster center $\mu_k = NN(x)$:

$$v_k = \sum_{x_j : NN(x) = \mu_k} x_j - \mu_k \tag{4}$$

The final VLAD encoding vector is obtained by concatenating v_k over all μ_k and has the dimension $D \times K$, where D is again the dimension of the local image feature vectors x. As for FV, the VLAD descriptor is also normalized subsequently. In this study, intra-normalization is performed, where the sum over each cluster center μ_k is ℓ_2-normalized before applying the standard ℓ_2-normalization of the entire VLAD descriptor.

3 Results

To estimate the generalization performance of the BoW approach using FV and VLAD, a leave-one-sequence-out cross-validation (LOSO-CV) model is used, to evaluate the classification performance.
As the LOSO-CV model would lead to exhausting computation times in case of the CNN approach, the performance is evaluated and compared to the textural features using a fixed train and test set following a 70:30 split ratio. To avoid correlation effects, complete sequences are used as hold out test set consisting of at least two sequences of each subject, one being physiological and one being pathological. The number of visual words are empirically set to 5 for both FV and VLAD as the performance did not improve using a larger vocabulary size in preliminary experiments.
The accuracy (Acc) and average recall (Rec) for the two feature encoding methods FV and VLAD are illustrated in Tab.1. Overall, VLAD encoding outperformed FV and reaches the best result with an accuracy of 87.6% and average recall of 86.7% using LBP and the SVM classifier.
The results of [2] using the same image database, are listed on the bottom of Tab.1. As comparison the two best performing features of [2] were chosen. The approach reaches an accuracy above 89.1% and average recalls above 90.3% using the SVM classifier and similar results also apply for the RF classifier.

Table 1. Classification results using FV, VLAD*, the approach of [2]$^+$ and LOSO-CV: Accuracy (Acc) and average recall (Rec)

		SVM		RF	
Features	Property	Acc	Rec	Acc	Rec
HOG	—	66.1%	63.0%	62.0%	55.3%
GLCM	QuantLvl 8	77.1%	74.1%	76.2%	72.1%
GLCM	QuantLvl 32	78.6%	75.8%	78.1%	74.2%
LBP	R5 N16	75.8%	71.0%	72.0%	65.3%
LDP	3rd order R5	83.4%	82.5%	81.5%	79.6%
HOG*	—	75.4%	73.6%	71.9%	66.4%
GLCM*	QuantLvl 8	83.4%	82.5%	84.3%	83.7%
GLCM*	QuantLvl 32	80.0%	76.5%	78.6%	74.6%
LBP*	R5 N16	87.6%	86.7%	87.5%	86.6%
LDP*	3rd order R5	82.9%	81.9%	79.6%	76.8%
GLCM$^+$	QuantLvl 8	89.8%	90.5%	86.4%	88.6%
GLCM$^+$	QuantLvl 32	89.6%	90.3%	86.7%	88.7%
LBP$^+$	R5 N16	89.1%	91.3%	89.3%	91.6%

Table 2. Comparison of CNN features with residual features using VLAD* and the approach of [2]$^+$: Accuracy (Acc) and average recall (Rec)

		SVM		RF	
Features	Property	Acc	Rec	Acc	Rec
CNN*	–	72.6%	69.5%	76.1%	74.7%
CNN*	Data augmentation	76.0%	75.7%	81.5%	81.7%
GLCM*	QuantLvl 8	61.4%	55.1%	72.0%	70.4%
LBP*	R5 N16	72.1%	68.2%	74.2%	72.2%
CNN$^+$	–	77.6%	80.1%	76.5%	78.8%
CNN$^+$	Data augmentation	79.9%	80.4%	81.3%	81.7%
GLCM$^+$	QuantLvl 8	77.9%	81.3%	76.4%	80.1%
LBP$^+$	R5 N16	79.5%	82.1%	80.5%	81.8%

For the evaluation of the CNN features, we only consider the approach of [2] and VLAD encoding as they consistently outperformed FV. Moreover for the comparison, we focus on the features GLCM and LBP due to their superior performances. In Tab.2, the results of CNN features and the residual features are illustrated. Using VLAD, CNN features exceed the classification results of all residual features with an accuracy of 76.1% and an average recall of 74.7% using the RF classifier. By using data augmentation, the results further improve to an accuracy of 81.5% and an average recall of 81.7%.
Using the approach of [2], CNN features using data augmentation and LBP show comparable classification results and outperform the residual features with average accuracies of 79.5% and 77.9% and average recalls of 82.1% and 80.4%, respectively.

4 Discussion

Despite of the very small visual vocabulary size, FV and VLAD already reach decent classification results and may have the potential to excel the algorithm

proposed by [2]. However, with the current setup, the approach of [2] outperforms our proposed method in case of all features. This might be due to the fact that Jaremenko et al. incorporated additional information in terms of the mean and standard deviation of all features and patches of an image, that is neglected within the VLAD approach.
CNN features show comparable results and slightly outperform any other of the tested features but still leave room for improvements using different CNN models. As expected, using data augmentation the performance of CNN increases as a result of the larger size and increased variance of the training set. Most likely the results could be improved further with additional augmentation, but this was not the aim of this paper. Considering the small amount of subjects of the dataset, it would be beneficial to increase its variance by investigating additional patients rather than performing augmentation using rotation. As a next step, with an increased patient database it would be possible to perform a leave-one-patient-out cross-validation to avoid intra-patient correlation effects during the training of the classifier that yet may be existent within the LOSO-CV and fixed dataset approach. The current results are promising but nonetheless, additional effort is needed, to further develop the proposed approach to be able to reliably support and improve diagnosis of vocal cord cancer during endoscopy.

References

1. Dittberner A, Rodner E, Ortmann W, et al. Automated analysis of confocal laser endomicroscopy images to detect head and neck cancer. Head & Neck. 2016;38:1419–1426.
2. Jaremenko C, Maier A, Steidl S, et al. Classification of Confocal Laser Endomicroscopic Images of the Oral Cavity to Distinguish Pathological from Healthy Tissue. Bildverarbeitung für die Medizin. 2015; p. 479–485.
3. Razavian AS, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In: Conf Comput Vis Pattern Recognit Workshops. IEEE; 2014. p. 512–519.
4. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2005;1:886–893.
5. Haralick R, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;(6):610–621.
6. Pietikäinen M, Hadid A, Zhao G, et al. Computer Vision Using Local Binary Patterns. London: Springer; 2011.
7. Zhang B, Gao Y, Zhao S, et al. Local Derivative Pattern Versus Local Binary Pattern: Face Recognition With High-Order Local Pattern Descriptor. IEEE Trans Image Process. 2010;19(2):533–544.
8. LeCun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE. 1998;86(11):2278–2324.
9. Sánchez J, Perronnin F, Mensink T, et al. Image Classification with the Fisher Vector: Theory and Practice. Int J Comput Vis. 2013;105(3):222–245.
10. Jegou H, Perronnin F, Douze M, et al. Image Classification with the Fisher Vector: Theory and Practice. IEEE Trans Pattern Anal Mach Intell. 2013;34(9):1704–1716.