International Journal of Computer Assisted Radiology and Surgery manuscript No.
(will be inserted by the editor)

Deep Learning-based Detection of Motion Artifacts in Probe-based Confocal Laser Endomicroscopy Images

Marc Aubreville · Maike Stoeve · Nicolai Oetter · Miguel Goncalves · Christian Knipfer · Helmut Neumann · Christopher Bohr · Florian Stelzle · Andreas Maier

This is a **pre-print** of an article published in International Journal of Computer Assisted Radiology and Surgery. The final authenticated version is available online at: https://doi.org/10.1007/s11548-018-1836-1

Abstract Purpose: Probe-based Confocal Laser Endomicroscopy (pCLE) is a subcellular in-vivo imaging technique capable of producing images that enable diagnosis of malign structural modifications in epithelial tissue. Images acquired with pCLE are, however, often tainted by significant artifacts that impair diagnosis. This is especially detrimental for automated image analysis, which is why said images are often excluded from recognition pipelines.

Methods: We present an approach for the automatic detection of motion artifacts in pCLE images and apply this methodology to a data set of 15 thousand images of epithelial tissue acquired in the oral cavity and the vocal folds. The approach is based on transfer learning from intermediate endpoints within a pre-trained Inception v3 network. For detection within the non-rectangular pCLE images, we perform geometrically motivated pooling within the activation maps of the network and evaluate this at different network

Marc Aubreville and Maike Stoeve contributed equally to the research in this paper.

Marc Aubreville · Maike Stoeve · Andreas Maier
Pattern Recognition Lab, Computer Science
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
E-mail: marc.aubreville@fau.de

Nicolai Oetter · Florian Stelzle
Department of Oral and Maxillofacial Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Miguel Goncalves
Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Christian Knipfer
Department of Oral and Maxillofacial Surgery, University Medical Center Hamburg-Eppendorf, Universität Hamburg, Germany

Christopher Bohr
Department of Otolaryngology, University Hospital Regensburg, Universität Regensburg, Germany

Helmut Neumann
First Department of Internal Medicine, University Hospital Mainz, Johannes Gutenberg-Universität Mainz, Germany

depths.

Results: We achieved area under the ROC curve values of 0.92 with the proposed method, compared to 0.80 for the best feature-based machine learning approach. Our overall accuracy with the presented approach is 94.8 %.

Conclusion: Over traditional machine learning approaches with state of the art features, we achieved significantly improved overall performance.

Keywords deep convolutional neural networks · confocal laser endomicroscopy · motion artifact detection

1 Introduction

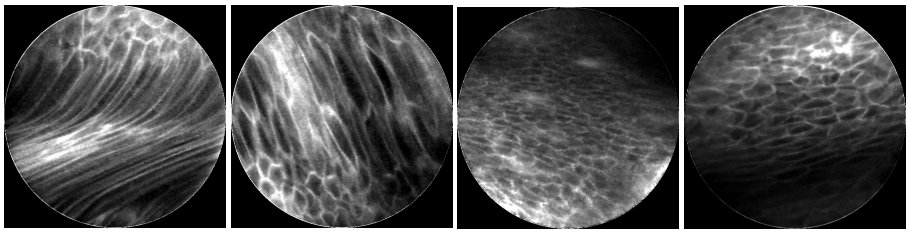
Squamous cell carcinoma (SCC) accounts for over 90 percent of all cancer types in the oral cavity and pharynx, as well as for almost all malignancies in the larynx [27]. Tobacco and alcohol consumption are regarded as the most important risk factors for head and neck cancer [19,31]. While this malignancy has had a higher prevalence for men in their 6th and 7th decade, in the last two decades rising incidence rates for patients below the age of 40 years have been observed [20]. An increasing incidence of cancer of the oral cavity and pharynx in younger patients has been attributed at least in part to the tumor-inducing properties of human papillomavirus [31,2]. Only around a third of the patients with head and neck cancer is diagnosed in an early tumor stadium (T1, i.e. with less then 2 cm in greatest diameter), which reduces treatment options. This, in turn, also increases the radicality of treatment as well as the mortality [27]. One promising option for earlier diagnosis are optical, non-invasive imaging methods.

Confocal Laser Endomicroscopy (CLE) is an imaging technique, that enables imaging of cellular microstructure of superficial mucosal layers with a high magnification (up to 1000x) and resolution. Probe-based CLE has a fixed focal length, thus being able to investigate the area of interest in a narrow plane [6]. The pCLE used in this study (CellVizio GastroFlex UHD probe, Mauna Kea Technologies, Paris, France) has an absorbing wavelength of around 660 nm and a penetration depth of 60 μm. Around 30 seconds before the examination, fluorescein as the contrast agent is given intravenously to the patient. Fluorescein distributes within intercellular spaces without diffusing through cell membranes, thus enabling the outline visualization and structural analysis of cellular tissue. Video sequences are obtained and visualized on a screen in real-time.

The classification of mucosal lesions using pCLE requires training to yield good accuracy ratings. The interpretation of such images is subject, however, to a rather large interobserver variability, which is why automated analysis could be advantageous for screening [1, 22,23,12]. A reliable and accurate evaluation of CLE images could potentially be used in the future to improve the location of tumor (margins) reducing the need for unnecessary tissue removal in areas so sensitive as the vocal cords or help reduce radicality during oncological surgery.

It is well known, that pCLE images can suffer from severe, deteriorating artifacts (see Fig. 1) that impede diagnosis [13,21,25]. Since artifacts are furthermore at times correlated with physical tissue properties, they should be excluded from classification tasks in order reduce biases [4].

Motion artifacts are a major impairment occurring during acquisition of pCLE images. Different CLE scanners have different frame refresh rates, e.g. CellVizio scanners typically



(a) streaky patterns (b) stretched cells (c) compressed cells (d) physiologically stretched cells

Fig. 1 Motion artifacts occurring in pCLE images. As depicted here, motion artifacts can cover a whole image or only parts of it. Stretched-appearing cells can however also occur physiologically (d).

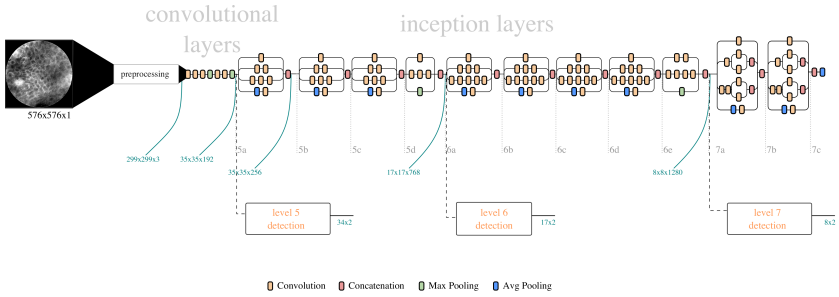


Fig. 2 Inception v3 network [29] with different attachment points on levels 5 to 7. The network has been pre-trained on ImageNet [9]. The motion detection extension is shown in detail in Fig. 3.

achieve a rate of 8 Hz while Optiscan (Optiscan Pty Ltd, Australia) scanners feature a frame rate of up to 1.2 Hz [13]. Motion artifacts are generated, when the probe and the tissue are moved relative to each other during sampling because of the construction principle of the device: Combining a horizontal oscillating mirror and a vertical galvanometric mirror, a meander-shaped sampling pattern is achieved [17]. The change in spatial position between both components can be expressed as a motion vector. If the vertical projection of the motion vector is negative, i.e. tissue and scanner direction coincide, the same line of tissue is potentially being scanned multiple times. This leads to two patterns, as depicted in Fig. 1: For small amplitudes of the motion vector, cells appear stretched or skewed (cf. Fig. 1(b)). For large amplitudes, streaky patterns occur (cf. Fig. 1(a)). If the tissue and the probe move in opposite directions, the vertical sampling is sparse, i.e. cell components appear compressed. While for stretched or compressed cells diagnostic value might still be present, this can clearly be neglected for strong motion artifacts leading to streaky patterns. Motion artifacts may appear in the complete image or only in parts of the image, with the restriction that they have no horizontal limitation, i.e. a horizontal line of the image is either artifact-tainted or not.

This work focuses on the detection of motion-induced artifacts in CLE images, and investigates machine learning techniques applicable for detection of image areas tainted by said impairments. Integrated into a diagnosis toolchain, such detection mechanisms help to interpret images and have the potential to improve the sensitivity, specificity and overall robustness of automatic malignancy detection systems for CLE images.

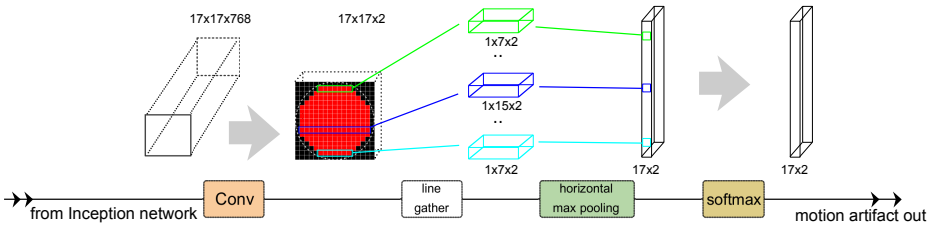


Fig. 3 Geometrical pooling for motion detection within the fully convolutional network. Here: Dimensions for Inception level 6 with horizontal/vertical resolution of 17. After an initial 2D convolutional layer, relevant horizontal slices are extracted and horizontal max pooling is applied. Finally, a softmax layer is applied.

2 Related Work

CLE has recently been proven to be of great value in several fields of diagnosis. In gastroenterology, it is clinically used successfully [21], and for some diseases even discussed as new gold standard for diagnosis [18]. In the field of urology disease assessment, Wiesner *et al.* have found that typical tumor growth patterns in urothelial bladder tissue were visible in CLE images. They argue that the tool could improve sensitivity and specificity over white light cytoscopy [32]. Pavlov *et al.* have shown the diagnostic potential for surgical guidance in glioma detection in the human brain [26].

Computerized image recognition plays an important role in medical image diagnosis. Convolutional neural networks (CNNs) have been successfully used for various tasks in biomedical image processing, such as mitosis detection in histology images [7] or retinopathy classification [34], to only name a few. In bright light microscopy, deep learning-based methods are emerging to be the leading pattern recognition tool [33].

Most recently, deep learning pipelines have successfully been applied also on pCLE images [13,4], where they outperformed state of the art approaches (e.g. textural feature-based approaches). While the automatic detection of malignant structures in CLE images has been investigated by several authors [15,30,4,11], previous work on preprocessing of CLE images is limited. Bier *et al.* have shown, that noisy CLE images can be improved using frequency-domain manipulations [5].

Izadyyazdanabadi *et al.* introduced a binary classification of CLE images into diagnostically useful and non-diagnostic images, i.e. images that are either tainted by artifacts or simply not containing visible features useful for diagnosis of the underlying tissue [13,14]. They showed, that deep learning techniques, in their case based on the well-known AlexNet and GoogleNet architectures, were well able to perform this task.

Besides this rather coarse differentiation, our previous work showed that it is possible to differentiate motion artifacts from regular image parts within a single image [28]. For this, we employed transfer learning based on the trunk of a pre-trained network [28]. This work, however, was restricted to a square-shaped image extracted from the round field of view of pCLE images. It was also evaluated only on a limited data set of 12 patients. Our contribution in this work is:

- We introduce the idea of pooling with a spatial constraint within the deep convolutional network. This enables us to make use of the complete CLE image, which also allows joint network topologies for malignancy detection or other detection tasks, that require analysis of the overall pCLE image.
- We extend the evaluation to a database of 22 patients.

– We present an extensive evaluation of our architectural optimization.

3 Materials

For this study, image material of $N = 22$ patients acquired by two independent hospitals was used. At the Department of Oral and Maxillofacial Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, 12 patients with suspected SCC in the oral cavity were examined. The other part of our study group ($N = 10$ patients) were patients undergoing diagnosis for squamous cell carcinoma of the vocal cords. This group was examined at the Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg. For artifact detection, images of verified squamous cell carcinoma as well as of presumably and verified non-cancerous tissue have been included.

At both departments, a probe-based CLE scanner (CellVizio, Mauna Kea Technologies, France) was used for image acquisition. Images acquired with both CLE devices were between 512×512 pixels and 576×576 pixels. Acquisition frame rate was 8 Hz for both devices.

In total, 201 sequences containing 15,018 images were used for this work. Images with high-grade noise level were excluded, since even manual motion artifact labeling is unreliable in this case. Each image was manually assessed for motion artifacts, and tainted areas were annotated and stored in a relational database.

4 Methods

For evaluation, the study group of 22 patients has been split up for a five-fold cross-validation scenario, resulting in a group of patients used for training ($N = 14$ or $N = 15$, depending on the fold), validation ($N = 3$) and testing ($N = 5$ or $N = 4$). The patients to be included in each run were chosen randomly but equal for all methods compared in this work. The split on a patient level ensures, that individual patient variances are covered by the statistical evaluation. Further, we can exclude potentially strong image correlations induced by sequences showing identical anatomical locations.

4.1 Deep Learning Pipeline

Our deep learning recognition pipeline consists of a preprocessing stage, followed by a convolutional neural network (CNN) based on Szegedy's Inception v3 [29] architecture, and an attached motion detection extension (see Fig. 2 and Fig. 3). All code can be found on the author's web page[1].

4.1.1 Preprocessing

The used CLE scanner produces monochrome images with a nominal depth of 16 bit. The round images that are produced by the scanner, however, are not well suited for automatic image recognition with CNNs, due to convolutions being very sensitive to the steep edges at

[1] `https://www5.cs.fau.de/~aubreville/` - Available after paper has been accepted

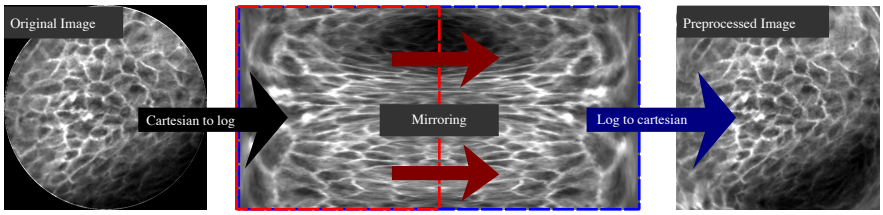


Fig. 4 Preprocessing for pCLE images. The contents of the round field of view are effectively point-mirrored around the outer border of the delimiting circle.

the field of view circle that yet do not comprise information. This problem can be circumvented by using only a rectangular square in the middle [28], or by extracting patches [15, 4, 30], which is also commonly done in other domains [34]. For our application we want the outer black border to replicate the field of view area statistically, which is why we propose to use a circular extrapolation technique.

In log polar coordinates, the image coordinates x, y are transformed according to [3]:

$$\rho = \log\left(x^2 + y^2\right) \tag{1}$$

$$\Phi = \arctan\left(\frac{y}{x}\right). \tag{2}$$

By concatenating a log-polar transformed image with a flipped copy of itself (along the ρ-axis), and back-transformation into cartesian space, mirroring at the border of the field of view is achieved (see Fig. 4).

This method is, while not preserving motion artifact information correctly outside the circular field of view, achieving similar statistical properties for the complete image.

Finally, the resulting image is rescaled and converted to grayscale in order to match the input dimensions of the pre-trained network of $299 \times 299 \times 3$, and normalized to unit variance and zero mean.

4.1.2 Geometrically Motivated Pooling from Fully Convolutional Networks

Convolutional neural networks are often seen as black boxes with hidden ingredients and unpredictable outcomes. One key to understanding this class of networks was proposed by Oquab *et al.* [24]: The basic idea of this approach is, that localization is kept within a network by solely employing convolutional filters and pooling/striding operations (fully convolutional network) to a certain stage, where a global max pooling operation is located (some authors also use averaging for this). Following this, there is a final fully connected and subsequent softmax layer, that is in a later object analysis/network segmentation phase being attached without modification to the network before the global max pooling operation. This way, the network is trained up to this central average pooling element to have equally important weights in all regions, i.e. independent from the localization of the object. The output of this analysis step then can be interpreted as a heat map, indicating which parts of the image contributed to the class discrimination.

Special to many pCLE images, however, is the fact that a round field of view - caused by the round optic fibre - leads to areas in the image that carry no information, but cause very steep gradients in the image. As convolutional filters are sensitive to gradients, this leads to a significant decrease in performance of the networks. For cases like these, we suggest to

use masking operations within the network. The masking operation cuts position-dependent fragments out of the original image, and subsequent operations have to be chosen in order to no longer rely on the spatial relationship between the fragments like average pooling.

We denote the respective previous layer of a network to be $\mathbf{U} \in \mathbb{R}^{W \times H \times C}$ with its elements $u_{i,j,c}$, where i, j and c are indices denoting the horizontal, vertical and channel component, respectively.

In order to restrict the attention of the algorithm to areas that have a valid image, we apply a masking operation $F_{\mathrm{M}} : U \rightarrow V'$ with $V' \in \mathbb{R}^{1 \times 1 \times C}$ the elements of the resulting tensor v'_c as:

$$v^*_c(i,j) = \delta(i,j) \cdot u_c(i,j) \tag{3}$$

with:

$$\delta(i,j) = \sigma\left(r^2 - \left(j - \frac{H}{2}\right)^2 - \left(i - \frac{W}{2}\right)^2\right) \tag{4}$$

where r is a constraining radius of the field of view and $\sigma(x)$ is the step function. The radius r needs to be set taking into account the dimensions of the previous network layer w.r.t. the original image. For the Inception v3 architecture, we chose $r_8 = 3.2$, $r_{17} = 17.65$, $r_{35} = 15.75$ for attachment in layers with width and height of 8, 17 and 35, respectively.

4.1.3 Network Architecture

Our proposed network extension for motion detection begins with a convolutional layer to reduce the number of channels to a binary classification with one-hot encoding (2 filters with 1x1 kernel size).

In order to be able to have a joint detection of motion artifacts on the whole image, we formulate a masking and extraction operation $F_{\mathrm{ME}} : U \rightarrow (M_1, M_2, \ldots, M_N)$, $U \in \mathbb{R}^{W \times H \times C}$, $M_n \in \mathbb{R}^{A_n \times 1 \times C}$ of vertical slices from the previous network layer. The line vectors M_n consist of $A_n \leq W$ elements $m_c(n)$, where:

$$A_n = \sum_i^W \delta(i,n) \tag{5}$$

$$m_c(i,n) = \delta(i,n) \cdot u_c(i,n) \tag{6}$$

To extract the motion information, a max pooling operation along the slice vector $F_{\mathrm{MLP}} : M \rightarrow P$, $P \in \mathbb{R}^{HxC}$ is performed:

$$p_c(n) = \max_i^{A_n} m_c(i,n) \tag{7}$$

Finally, a softmax operation is applied to retain probabilities for the two classes: motion artifact absence and motion artifact presence:

$$c_{\mathrm{c}}(n) = \frac{exp(p_c(n))}{\sum_i^C exp(p_i(n))} \tag{8}$$

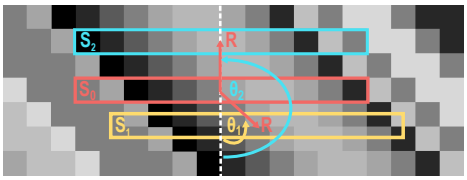


Fig. 5 Visualization of angle of maximum correlation feature.

4.1.4 Transfer Learning and Training

As previously stated, our deep learning pipeline is based upon an Inception v3 [29] network with a custom extension for the task. The network's weights are initialized from a network pre-trained on ImageNet [9], and the custom extension's weights are randomly initialized. In order to train this custom tail faster than the original network and keep the generalization properties of the Inception network stem high, we use a higher initial step sizes for the ADAM optimizer [16] for the tail ($5 \cdot 10^{-4}$) than the stem ($5 \cdot 10^{-6}$).

Since the classes *motion artifact* and *artifact free* are heavily skewed, we employ undersampling of the majority class: For each training mini-batch, an equal number of images containing motion artifacts and artifact free images are chosen randomly from the data set. Due to this random picking process, a training epoch is longer deterministically defined, but statistically (while some duplicates may be included and some may be left out for a single epoch).

Further, we utilize an early stopping scheme, where the network weights are stored after each statistical epoch. We train for a minimum of 2 epochs, and afterwards stop the training and restore the weights, if the accuracy on the complete validation set is below the respective value for the preceding run.

4.2 Feature-based Machine Learning Pipeline

Our feature-based machine learning pipeline consists of overlapping slice extraction, feature extraction and classification by a random forest classifier. Since our distribution is heavily skewed, we employed undersampling of the majority class (i.e. *artifact free*) to reach even distributions for the classifier training.

In order to detect motion artifacts, we employ three hand-crafted features, out of which the following two are known from literature:

1. Histogram of oriented gradients (HOG) as described by Dalal and Triggs [8] is a feature successfully used in object detection [10]. Streaky patterns (cf. Fig. 1(a)) esentially represent strong geometric gradients along a certain spatial direction, which motivates the use for this work. HOG vectors are calculated for square-shaped blocks within the current patch, resulting in a concatenated HOG descriptor of varying length (due to the varying width in dependency of the vertical position). From this HOG descriptor matrix, the first four central moments are calculated and used as a feature.
2. Local binary patterns (LBP) have successfully been employed in the field of CLE malignancy detection [15] and in other image recognition fields. They describe the relationship of a pixel with its circular neighborhood, in that its pixel value is compared to the neighboring pixels, resulting in a binary mask. For this work, we use uniform, rotation invariant LBPs, i.e. patterns rotated bitwise to their minimum binary representation. In

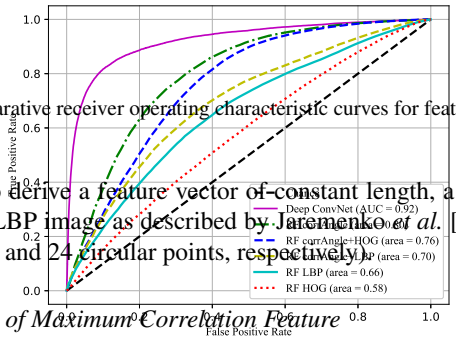


Fig. 6 Comparative receiver operating characteristic curves for feature-based and deep learning methods.

order to derive a feature vector of constant length, a histogram operation is performed on the LBP image as described by Jaremenko *et al.* [15]. We use LBP-radii of 1, 3 and 5 (8, 16 and 24 circular points, respectively).

4.2.1 Angle of Maximum Correlation Feature

Motion artifacts, especially when the motion vector coincides with the scanning direction, induce strong correlations between one line S_0 of the image and a line S_1 of the image that is sampled with an angular offset (according to the motion vector) at a fixed radius. For the angle of maximum correlation (corrAngle) feature, lines S_k are being interpolated from the image with a fixed radius R and varying angle θ_k (see Fig. 5). The coordinates $(x_{s_\theta}, y_{s_\theta})$ of an image line S_k with vertical offset o used for pairwise comparison with the reference line S_0 are defined as:

$$\begin{pmatrix} x_{s_\theta}(n) \\ y_{s_\theta}(n) \end{pmatrix} = \begin{pmatrix} r \cdot \sin(\theta_k) + n \\ r \cdot \cos(\theta_k) + o \end{pmatrix} \tag{9}$$

Finally, utilizing Pearson's ρ, the angle of maximum correlation is defined as:

$$\rho_{\max} = \underset{k}{\operatorname{argmax}} \frac{\sum \left(S_0 - \overline{S_0}\right)\left(S_k - \overline{S_k}\right)}{\sqrt{\sum \left(S_0 - \overline{S_0}\right)^2} \sqrt{\sum \left(S_k - \overline{S_k}\right)^2}} \tag{10}$$

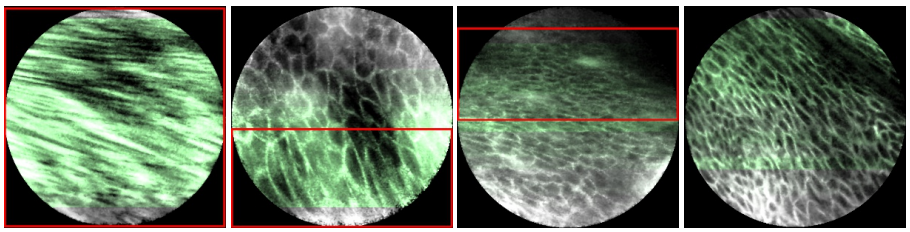
5 Results

As expected from our previous results [28], the convolutional network clearly outperformed all feature-based approaches with AUC values of 0.92 (see Fig. 6). The next best performing classifier was using solely the newly designed feature *angle of maximum correlation* (AUC=0.82), while commonly used features like LBP and HOG were inferior. The combination of features did not increase the performance, which could indicate an overfit to the features. Dimensionality reduction methods have not been employed in this work, as we assumed they would likely not increase the performance to that of the deep neural network.

To compare recognition performance across different motion artifact manifestations, we calculated the mean accuracy over all frames where said artifact was annotated (see Table 1). The comparison shows the general superiority of the presented CNN-approach over all

Artifact type →	stretched	streaky	compressed
Deep Net	0.746	0.780	0.764
corrAngle (RF)	0.676	0.701	0.749
HOG (RF)	0.551	0.609	0.610

Table 1 Mean accuracies calculated on different artifact types. Accuracy is only calculated on the respective frames with annotated artifacts, explaining the low values compared to the overall accuracy. Note that the low a priori probability of motion artifacts was not considered by the classifier, since undersampling of the majority class was applied.



(a) streaky patterns (b) stretched cells (c) compressed cells (d) physiologically stretched cells

Fig. 7 Exemplary detection of motion artifacts. Color-coded in green is the probability of motion artifacts, as detected by the network. The manually labeled ground truth is annotated with red rectangles.

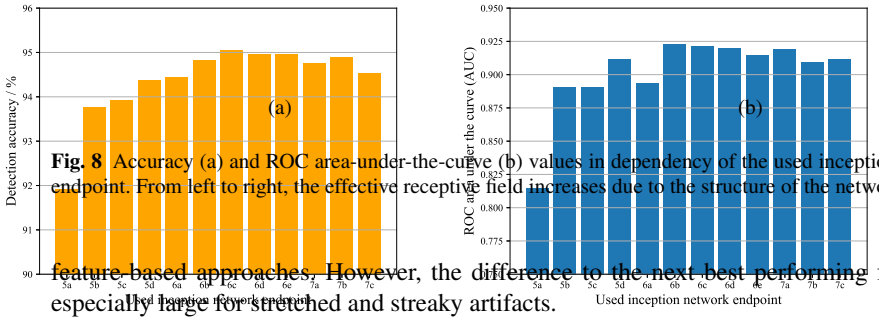


Fig. 8 Accuracy (a) and ROC area-under-the-curve (b) values in dependency of the used inception network endpoint. From left to right, the effective receptive field increases due to the structure of the network.

feature-based approaches. However, the difference to the next best performing feature is especially large for stretched and streaky artifacts.

5.1 Depth-Performance Relation

For product-grade applications, complexity of a method is of almost equal importance as performance. For convolutional networks, the complexity is linked to the depth of the network. We thus evaluated the performance in dependency of the layer where the motion detection extension was attached (cf. Fig. 2). The results (see Fig. 8) indicate, that the performance increases up to the layer commonly denoted as 6b.

6 Discussion

The used network topology (Inception v3) can be described as a fully convolutional network, de-facto achieving a downscaling in its first convolutional layers (up to 5a), i.e. only a limited receptive field is considered. The inception layers following afterwards add horizontal and vertical stripe convolutions, effectively taking a more widespread context into account. While there is still a direct relation between a coordinate within the output of a inception block and its respective receptive field at the input, a direct 1:1 connection can not be made. This underlines why our pre-processing is so important, as the pre-trained network uses a very broad receptive field and steep edges in the input image around the round field of view hinder convergence. The broadened receptive field, as induced by the inception blocks, however adds accuracy up to a certain layer of the network (as shown in Fig. 8). This can also be motivated by the fact that motion artifacts usually cover not only single parts of the image, and thus a broader receptive field can help the network to increase its confidence.

Some false positives are induced by physiological variations (see Fig. 7(d)). Considering single still images, the decision on motion artifact contents within these images is, even for the human observer, a difficult distinction to make. For this, embedding of sequential information into the pipeline would be required and is expected to improve results.

The high accuracy ratings of up to 0.95 should not neglect the fact that for relatively underrepresented events as motion artifacts, this is only part of the truth. As our mean accuracies (Table 1) indicate, the detection on those images tainted by artifacts is only true in less than 80 % of the cases. However, it should be noted, as acceleration is not infinite, the precise starting or ending position of an artifact is hard to define for some images (see e.g. Fig. 7(b)). Visual inspection of the results indicates, that there is generally a very low subjective false positive rate, which is also expressed by the generally high accuracies of the network (Fig. 8(a)). It should also be pointed out, that quantization of the vertical axis, as performed by both, the CNN and the patch extraction method, has an impact on performance, since the annotation resolution is on a pixel level. Further, we did not distinguish between weak (as in: only noticeable by comparing sequential images) and strong artifacts. Images tainted by weak motion artifacts should not be problematic for expert grading, nor should they be a significant problem for automated approaches.

In our work, we applied motion detection solely on CLE images of epithelial cells, however CLE is being applied in a variety of other scenarios. At this time, no data set that could be used to demonstrate applicability to those other domains was available for our study. Yet, we are confident that the approach is indeed transferable to other pCLE images, since the origin of the artifacts is identical.

7 Summary

Confocal Laser Endomicroscopy is an imaging method that has been proven to be highly suitable for the detection of squamous cell carcinoma in the head and neck region.

In this work, we introduced a new method for motion artifact detection in pCLE images, a common deterioration induced by doctor's hand or patient movement. This new approach, which is based on masking and pooling from within a pre-trained network architecture, enables detection of said artifacts on the complete image using a deep learning pipeline.

Over previous approaches with state-of-the-art features, the method shows clearly superior properties, yielding better detection accuracies while taking into account almost the complete image. This was demonstrated on two data sets of epithelial tissue, both of which

include images of presumably physiological and malign tissue (in total: 15,018 single images). Our five-fold cross-validation on patient level found accuracies of 94.8 % with area under the ROC curve values of 0.92.

This, from our point of view, represents a key step towards a fully automatic classification and analysis of pCLE images. We expect further insights about robustness and clinical applicability with the acquisition of more image sequences, which is part of our ongoing research.

Compliance with Ethical Standards

All authors (M. Aubreville, M. Stoeve, N. Oetter, M. Goncalves, C. Knipfer, H. Neumann, C. Bohr, F. Stelzle and A. Maier) declare, that they do not have conflicts of interest regarding the work covered by this manuscript.

The local ethics committee approved the studies (ethics committee of the University of Erlangen-Nürnberg; reference numbers 243 12 B and 60 14 B) and all patients gave their written informed consent. All procedures involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments.

References

1. Abbaci, M., Breuskin, I., Casiraghi, O., De Leeuw, F.: Confocal laser endomicroscopy for non-invasive head and neck cancer imaging: a comprehensive review. Oral Oncology **50**(8), 711–6 (2014). DOI 10.1016/j.oraloncology.2014.05.002
2. Agaimy, A., Weichert, W.: Grading of Head and Neck Neoplasms. Der Pathologe **37**(4), 285–292 (2016). DOI 10.1007/s00292-016-0173-9
3. Araujo, H., Dias, J.M.: An introduction to the log-polar mapping [image sampling]. In: Proceedings, 2nd Workshop on Cybernetic Vision, pp. 139–144. IEEE (1996). DOI 10.1109/CYBVIS.1996.629454
4. Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, Christian, Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., Maier, A.: Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. Scientific Reports **7**(1), 11979 (2017). DOI 10.1038/s41598-017-12320-8
5. Bier, B., Mualla, F., Steidl, S., Bohr, C., Neumann, H., Maier, A., Hornegger, J.: Band-Pass Filter Design by Segmentation in Frequency Domain for Detection of Epithelial Cells in Endomicroscope Images. In: Bildverarbeitung für die Medizin 2015, pp. 413–418. Springer Berlin Heidelberg, Berlin, Heidelberg (2015). DOI 10.1007/978-3-662-46224-9_71
6. Chauhan, S.S., Dayyeh, B.K.A., Bhat, Y.M., Gottlieb, K.T., Hwang, J.H., Komanduri, S., Konda, V., Lo, S.K., Manfredi, M.A., Maple, J.T., Murad, F.M., Siddiqui, U.D., Banerjee, S., Wallace, M.B.: Confocal laser endomicroscopy. Gastrointestinal Endoscopy **80**(6), 928–938 (2014). DOI 10.1016/j.gie.2014.06.021
7. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013 **16**(Pt 2), 411–418 (2013). DOI 10.1007/978-3-642-40763-5_51
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886–893. IEEE (2005). DOI 10.1109/CVPR.2005.177
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE (2009). DOI 10.1109/CVPR.2009.5206848
10. Deniz, O., García, G.B., Salido, J., De la Torre, F.: Face recognition using Histograms of Oriented Gradients. Pattern Recognition Letters **32**(12), 1598–1603 (2011). DOI 10.1016/j.patrec.2011.01.004
11. Dittberner, Andreas, Rodner, Erik, Ortmann, Wolfgang, Stadler, Joachim, Schmidt, Carsten, Petersen, Iver, Stallmach, Andreas, Denzler, Joachim, Guntinas-Lichius, Orlando: Automated analysis of confocal laser endomicroscopy images to detect head and neck cancer. Head & Neck **38**(S1), E1419–E1426 (2016). DOI 10.1002/hed.24253

12. Goncalves, M., Iro, H., Dittberner, A., Agaimy, A., Bohr, C.: Value of confocal laser endomicroscopy in the diagnosis of vocal cord lesions. European Review for Medical and Pharmacological Sciences **21**, 3990–3997 (2017)
13. Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., Preul, M.C.: Improving utility of brain tumor confocal laser endomicroscopy - objective value assessment and diagnostic frame detection with convolutional neural networks. In: Medical Imaging 2017, vol. 10134. SPIE (2017). DOI 10.1117/12.2254902
14. Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., Preul, M.C., Yang, Y.: Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization for neurosurgical cle images. Journal of Visual Communication and Image Representation **54**, 10 – 20 (2018). DOI 10.1016/j.jvcir.2018.04.004
15. Jaremenko, C., Maier, A., Steidl, S., Hornegger, J., Oetter, N., Knipfer, C., Stelzle, F., Neumann, H.: Classification of Confocal Laser Endomicroscopic Images of the Oral Cavity to Distinguish Pathological from Healthy Tissue. In: Bildverarbeitung für die Medizin 2015, pp. 479–485. Springer Berlin Heidelberg, Berlin, Heidelberg (2015). DOI 10.1007/978-3-662-46224-9_82
16. Kingma, D., Ba, J.: ADAM: A method for stochastic optimization. arXiv preprint **arXiv:1412.6980** (2014)
17. Laemmel, E., Genet, M., Le Goualher, G., Perchant, A., Le Gargasson, J.F., Vicaut, E.: Fibered confocal fluorescence microscopy (Cell-viZio) facilitates extended imaging in the field of microcirculation. A comparison with intravital microscopy. Journal of vascular research **41**(5), 400–411 (2004). DOI 10.1159/000081209
18. Macé, V., Ahluwalia, A., Coron, E., Le Rhun, M., Boureille, A., Bossard, C., Mosnier, J.F., Matysiak-Budnik, T., Tarnawski, A.S.: Confocal laser endomicroscopy: A new gold standard for the assessment of mucosal healing in ulcerative colitis. Journal of Gastroenterology and Hepatology **30**, 85–92 (2015). DOI 10.1111/jgh.12748
19. Maier, H., Dietz, A., Gewelke, U., Heller, W., Weidauer, H.: Tobacco and alcohol and the risk of head and neck cancer. The Clinical Investigator **70**(3-4), 320–327 (1992). DOI 10.1007/BF00184668
20. Nachalon, Y., Alkan, U., Shvero, J., Yaniv, D., Shkedy, Y., Limon, D., Popovtzer, A.: Assessment of laryngeal cancer in patients younger than 40 years. The Laryngoscope (2017). DOI 10.1002/lary.26951
21. Neumann, H., Langner, C., Neurath, M.F., Vieth, M.: Confocal Laser Endomicroscopy for Diagnosis of Barrett's Esophagus. Frontiers in Oncology **2**, 42 (2012). DOI 10.3389/fonc.2012.00042
22. Neumann, H., Vieth, M., Atreya, R., Neurath, M.F., Mudter, J.: Prospective evaluation of the learning curve of confocal laser endomicroscopy in patients with IBD. Histology and histopathology **26**(7), 867–872 (2011)
23. Oetter, N., Knipfer, C., Rohde, M., Wilmowsky, C., Maier, A., Brunner, K., Adler, W., Neukam, F.W., Neumann, H., Stelzle, F.: Development and validation of a classification and scoring system for the diagnosis of oral squamous cell carcinomas through confocal laser endomicroscopy. Journal of Translational Medicine **14**(1), 1–11 (2016). DOI 10.1186/s12967-016-0919-4
24. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 685–694. IEEE (2015). DOI 10.1109/CVPR.2015.7298668
25. Parikh, N.D., Gibson, J., Nagar, A., Ahmed, A.A., Aslanian, H.R.: Confocal laser endomicroscopy features of sessile serrated adenomas/polyps. United European Gastroenterology Journal **4**(4), 599–603 (2016). DOI 10.1177/2050640615621819
26. Pavlov, Vladislav, Meyronet, David, Meyer-Bisch, Vincent, Armoiry, Xavier, Pikul, Brian, Dumot, Chloé, Beuriat, Pierre-Aurelien, Signorelli, Francesco, Guyotat, Jacques: Intraoperative Probe-Based Confocal Laser Endomicroscopy in Surgery and Stereotactic Biopsy of Low-Grade and High-Grade Gliomas. Neurosurgery **79**(4), 604–612 (2016). DOI 10.1227/NEU.0000000000001365
27. Robert Koch Institut. Zentrum für Krebsregisterdaten: Krebs in Deutschland für 2013/2014, 11 edn. Robert Koch Institut, Berlin (2017)
28. Stoeve, M., Aubreville, M., Oetter, N., Knipfer, C., Neumann, H., Stelzle, F., Maier, A.: Motion Artifact Detection in Confocal Laser Endomicroscopy Images. In: Bildverarbeitung für die Medizin, pp. 328–333. Springer Vieweg, Berlin, Heidelberg (2018). DOI 10.1007/978-3-662-56537-7_85
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). DOI 10.1109/CVPR.2015.7298594
30. Vo, K., Jaremenko, Christian, Bohr, C., Neumann, H., Maier, A.: Automatic Classification and Pathological Staging of Confocal Laser Endomicroscopic Images of the Vocal Cords. In: Bildverarbeitung für die Medizin 2017, pp. 312–317. Springer Vieweg, Berlin, Heidelberg, Berlin, Heidelberg (2017). DOI 10.1007/978-3-662-54345-0_70

31. Westra, W.H.: The pathology of HPV-related head and neck cancer: Implications for the diagnostic pathologist. Seminars in Diagnostic Pathology **32**(1), 42–53 (2015). DOI 10.1053/j.semdp.2015.02.023
32. Wiesner, C., Jäger, W., Salzer, A., Biesterfeld, S., Kiesslich, R., Hampel, C., Thüroff, J.W., Goetz, M.: Confocal laser endomicroscopy for the diagnosis of urothelial bladder neoplasia: a technology of the future? BJU International **107**(3), 399–403 (2010). DOI 10.1111/j.1464-410X.2010.09540.x
33. Xing, F., Xie, Y., Su, H., Liu, F., Yang, L.: Deep learning in microscopy image analysis: A survey. IEEE Transactions on Neural Networks and Learning Systems **PP**(99), 1–19 (2017). DOI 10.1109/TNNLS.2017.2766168
34. Yang, Y., Li, T., Li, W., Wu, H., Fan, W., Zhang, W.: Lesion Detection and Grading of Diabetic Retinopathy via Two-Stages Deep Convolutional Neural Networks. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017, pp. 533–540. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-66179-7_61