21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Training Data Selection for Cancer Detection in Multispectral Endoscopy Images

Cuong V. Dinh[1,2], Marco Loog[1], Raimund Leitner[2], Olga Rajadell[3], Robert P.W. Duin[1]
[1]*Pattern Recognition Laboratory, Delft University of Technology, the Netherlands*
[2]*Carinthian Tech Research AG, Austria*
[3]*Institute of New Imaging Technologies, University Jaume I, Spain*

Abstract

Multispectral endoscopy images provide potential for early stage cancer detection. This paper considers this relatively novel imaging technique and presents a supervised method for cancer detection using such multispectral data. The data under consideration include different types of cancer. This poses a challenge for the detection as different cancer types may exhibit different spectral signatures. Consequently, it is not always feasible to transfer the knowledge learnt from one data set to another data set. In our approach, we select suitable training data for a given test set based on a similarity measurement between data sets. Experimental results demonstrate that the classification results can be significantly improved if a few data sets that are presumably similar to a given test set are selected for training instead of using all available data sets.

1 Introduction

Early cancer detection plays an important role in increasing the chance for successful cancer treatment. A common technique for early cancer diagnosis is taking biopsies, which requires physical removal of specimens followed by a histopathological analysis [6]. It is difficult to determine the dysplastic and malignant regions for biopsies and therefore the procedure may have to be repeated many times, which delays the necessary treatment.

Optical techniques, such as the autofluorescence spectroscopy, have been investigated for early cancer diagnosis. Autofluorescence is the light emission of specific substances of biological tissues, e.g. porphyries and proteins if the tissues are excited by a light source. Those substances then emit light of specific wavelengths. The spectra of the tissues then correspond to different wavelengths measured by the spectroscopy. Previous studies, e.g. [2] have shown that there is a significant difference in the fluorescent properties, such as their spectral shape and intensity, between malignant and normal tissues. Therefore, they have been used to identify early instances of diseases in the colon, larynx, lung, and other organs.

The advantage of optical techniques lies in their potential to perform *in vivo* detection without the need for tissue removal. Therefore, they facilitate the determination of the dysplastic and malignant regions for the biopsy. These spectroscopic diagnosis techniques are often referred to as point-measurement methods as they attempt to obtain the spectra of a single tissue.

Multispectral/hyperspectral endoscopy techniques developed recently provide three-dimensional images of the area of interest in both spatial and spectral domains [3, 4, 6]. Multispectral images provide richer information than point-measurement techniques as they are able to acquire the spectra of thousands to millions of malignant and normal pixels at the same time. In [4], a thresholding algorithm is used to assign pixels to normal/malignant spectra based on the observation that the intensity of a malignant area is brighter than that of a normal area. In this paper, we present a supervised method, in particular, we focus on the issue of transferring knowledge among data sets.

The data under consideration consist of eight multispectral endoscopy images belonging to different types of cancers. As different cancer types may exhibit different spectral signatures [1], the discriminant information between normal and malignant tissues learnt from a data set may not be applicable to another data set. We address this problem by selecting suitable training data sets for a given test set. Data sets are only selected for training if they are similar to the test set, i.e. they stay close to it in the feature space. Experimental results show that the classifications can be significantly improved if a few data sets which are similar to a test set

978-4-9906441-1-6

are selected for training instead of using all data sets.

2 Materials

Data were collected from patients with different kinds of cancer at the hospitals for otolaryngoscopic and thorax surgeries in Stuttgart, Germany. Multispectral images of the investigated tissue areas were recorded quickly after surgery so the *in vivo* conditions of the tissues are commonly believed to be conserved. The hyperspectral images were produced using an electron multiplying charge coupled device (EMCCD) camera, with a resolution of 1002 $\times$ 1004 pixels, an acousto-optic tunable filter (AOTF) with wavelengths ranging from 400nm to 650nm (FWHM 5nm), and a 10 mm laparoscope with a 300W Xenon light source.

The eight data sets under consideration (called M1, M2 $\cdots$ M8) belong to different types of cancer: Laryngeal cancer (data sets M3, M4, M5, and M8), Pharyngeal cancer (M1), Esophageal cancer (M2), Diaphragm cancer (M6), and Parotid cancer (M7). For the M4 data set, the exact boundary of the cancer area is unclear since the cancer tissue is under the surface. Therefore, it is not easily detectable by a non-penetrating optical method. All data sets are acquired in a white light condition and the number of spectral bands is 51.

3 Methods

3.1 Data preprocessing

First, each reflectance spectrum is normalized using the area under the curve normalization in the spectral domain. Second, spectra corresponding to the specular reflection are removed by a simple thresholding algorithm. Third, the principal component analysis (PCA) is used to reduce the number of features from the original space. The reconstruction of all data sets are then based on their first eight eigenvectors which preserve 99% of the total variance. Finally, a unit variance normalization is applied to each data set so that each spectral band has a zero mean and a unit variance. The main aim of this normalization is to align all the data sets, i.e. to force them to stay close to each other in the feature space.

3.2 Data selection for training

As the data sets are different from one to another with respect to their class distributions, the discriminant information between normal and malignant tissues learnt from a data set might not be suitable for another data set. Therefore, it is essential to select suitable training sets for a given test set. We first use the Gaussian data domain description [5] to model the distribution of each data set. Denote q the percentage of outliers in each data set, a pixel is considered as an outlier if its probability density $p(x_i)$ is smaller than a threshold θ determined by:

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta - p(x_i)) = q$$

where N is the total number of pixels in the data, $h(.)$ the unit step function, and $p(x_i)$ the probability density of pixel x_i. We then measure the similarity between two data sets by the fraction of pixels they share in their data domain. For two data sets M_i and M_j, we calculate M_{ij} the set of all pixels in M_i that belong to the domain of M_j and M_{ji} the set of all pixels in M_j that belong to the domain of M_i. The similarity between M_i and M_j denoted by S_{ij} is defined as: $S_{ij} = |M_{ij}|/|M_i| + |M_{ji}|/|M_j|$. The similarity among data sets are then used as the criterion to select the training set for a given test set. Note that we model a data set using all the pixels contained. It is therefore possible to measure the similarity between any two data sets, e.g. between a training set and a test set, even when we do not have label information of the test set.

4 Experimental results

As knowledge about the prior probabilities of normal and malignant classes is not available, we set them to be equal in all experiments. We use the quadratic discriminant classifier (QDC) for the classification between normal/malignant tissues.

4.1 All available data sets are used for training

We first evaluate the classification results for two training scenarios: i) training and test data are from the same data set, i.e. a part of a data set is used for training and the remainder is for testing; ii) training and test data are from different data sets. For the latter, we follow the leave-one-dataset-out cross validation configuration, i.e. seven data sets are used for training and the remaining data set is used for testing. Moreover, for the second scenario we also investigate the influence of the unit variance normalization in the data preprocessing step. Since the QDC is invariant to affine transformations, the classification results for the first scenario remain unchanged whether this normalization is applied or not. Table 1 shows the classification results with respect to different training and normalization options. The table clearly shows the difference between

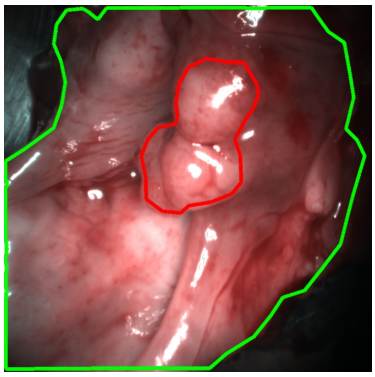

Figure 1. Reconstructed color image of the data set M8. Normal and malignant areas are marked by green and red contours.

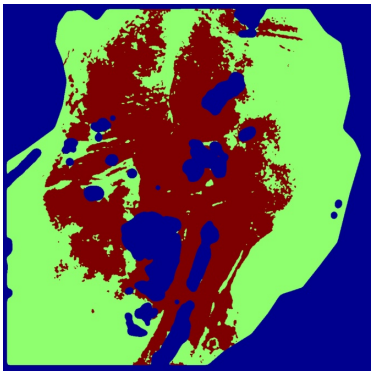

Figure 2. Classification result for the M8 data set using the unit variance normalization and QDC.

Table 1. Error rate (%) for different training and normalization options

Training scenario	Normalization	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Same set	No	09.9	11.0	16.0	10.1	05.8	15.8	10.0	07.6	10.8
Different sets	No	39.8	48.9	34.0	28.8	51.6	46.0	30.2	22.8	37.7
Different sets	Yes	**30.1**	**26.5**	**36.0**	**29.6**	**17.6**	**38.1**	**30.5**	**23.3**	**29.0**

Table 2. Error rate (%) when training data selection is used

	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Case 1	42.0	27.0	45.4	51.2	27.2	42.7	31.9	24.0	36.4
Case 2	**26.8**	25.4	**30.1**	**24.1**	26.5	50.9	**26.5**	**16.0**	**28.3**
Case 3	30.8	27.6	32.6	28.4	20.1	45.6	28.9	18.4	29.1
Case 4	31.8	26.2	39.0	25.2	18.7	44.0	29.8	24.3	29.9
Case 5	26.8	**25.3**	36.9	27.7	19.0	39.6	34.3	25.1	29.3
Case 6	29.2	24.7	36.0	28.7	**17.2**	**36.1**	30.9	24.7	28.4
Case 7	30.1	26.5	36.0	29.6	17.6	38.1	30.5	23.3	29.0

Table 3. Best error rate (%) and the corresponding number of data sets used for training

	M1	M2	M3	M4	M5	M6	M7	M8	Mean
Errors	21.6	20.0	28.2	22.1	13.3	29.7	23.8	13.8	21.5
#training sets	3	2	2	2	2	4	3	2	2.5

the two training scenarios. The error rate increases substantially when training and test data are not from the same data set as they are far different from each other. In addition, the unit variance normalization is demonstrated to significantly improve the classification results when the training and test data are from different data sets. Therefore, we apply the normalization step in all of the following experiments. Note that we also used other classification methods, such as Parzen classifier and the linear SVM; however, they often perform worse than the QDC. Figure 1 displays the reconstructed color image for the data set M8 and its normal and malignant areas. The border of the malignant area has been annotated by a medical expert. The classification on this data set is shown in Figure 2. Detected malignant and normal areas are displayed in red and green, respectively. Blue depicts background. The blue areas within the tissues correspond to the specular reflections. They are removed during the preprocessing step mentioned in Section 3.1 and therefore not considered in the classifica-

tion. The figure shows that the detected malignant area tends to expand into the normal area. One reason might be that malignant tissues in the training data are different from one another as they belong to different cancer types. Thus, they exhibit mixture data distribution. However, the QDC assumes only a single Gaussian for each malignant/normal class. Consequently, the estimated distribution of the malignant class becomes more flat and therefore more tissues become false positive.

4.2 Training data selection

We evaluate the classification results when the training data contain similar data sets for a given test data set. We model the data sets by using the Gaussian domain description in which the percentage of outlier q is set to 0.1. For each data set, we first selected the training data as the most one, two $\cdots$ seven similar data sets (denoted by Case 1, $2 \cdots 7$) according to the similarity measurement defined in the Section 3.2. The QDC is then trained on the selected training data and subsequently used for the classification of normal/malignant tissues for the data set under consideration. Table 2 shows the error rates for all seven cases. Numbers in bold emphasize the best results achieved for each data set in all cases. Note that Case 7 corresponds to the results shown in the third row of Table 1 as all seven data sets are included in the training data. On average, the best classification results are obtained if the two most similar data sets are used for training (Case 2). Increasing the number of training data sets then, in most of the time, worsens the classification as irrelevant data are included in the training process. Case 2 yields the best results for five over eight data sets. Case 2 does not perform well on the data set M6 as the data set itself is challenging: the cancer type (diaphragm) is totally different from the other cancer types.

We also carry out experiments in which for each data set, the training data is manually selected according to the classification results. Table 3 shows the best error rates and the corresponding number of data sets used for training. Similar to the above results (cf. Table 2), the classifier performs best when two or three data sets are selected for training. We also noted that for any of the three data sets M3, M5, and M8, the best classification result is achieved if the other two data sets are included in the training data except for the M5 where the best training set contains M7 and M8 yielding an error rate of 13.3%. Nevertheless, the training set containing M3 and M8 produces a comparable result of 15.1% error rate. This confirms the fact that the three data sets are similar as they exhibit the same type of cancer (Laryngeal cancer).

5 Conclusion

This paper presents a study of normal/malignant tissue classification for eight multispectral endoscopy data sets in a supervised manner. The data are heterogeneous as they are collected from different patients and with different types of cancer. We showed that the classification result is improved if a subset of the data that are similar to the test set is used for training (cf. Table 2 & 3). In other words, it is not always good to combine all available data for training as the difference between the data sets may result in poor classification.

We introduce an approach to select training data based on the similarity between data sets using the Gaussian data domain description. Experimental results show that the method substantially improves the classification results for our heterogeneous data. Note that we measure the similarity between data sets based on all the pixels, i.e. from both normal and malignant classes. For data from a patient who does not have cancer, all the pixels should fall into the normal region of the selected training data; therefore, our method correctly classifies the data set as normal.

In the present paper we use PCA to reduce the dimensionality of the feature space. To find subspaces that provide discriminant information between normal/malignant tissues in the data may also improve the performance of the classifiers. Finally, more data sets are essential to fully evaluate the applicability of our method.

References

[1] S. Demos, R. Gandour-Edwards, R. Ramsamooj, and R. deVere White. Near-infrared autofluorescence imaging for detection of cancer. *Journal of biomedical optics*, 9:587, 2004.

[2] M. Harries, S. Lam, C. MacAulay, J. Qu, and B. Palcic. Diagnostic imaging of the larynx: autofluorescence of laryngeal tumours using the helium-cadmium laser. *The Journal of Laryngology & Otology*, 109(02):108–110, 1995.

[3] R. Leitner, T. Arnold, and M. De Biasio. High-sensitivity hyperspectral imager for biomedical video diagnostic applications. In *Proceedings of SPIE*, volume 7674, 2010.

[4] M. Martin et al. Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection. *Annals of biomedical engineering*, 34(6):1061–1068, 2006.

[5] D. Tax. *One-class classification; Concept-learning in the absence of counter-examples (Chapter 3)*. PhD thesis, Delft University of Technology, 2001.

[6] T. Vo-Dinh. A hyperspectral imaging system for in vivo optical diagnostics. *Engineering in Medicine and Biology Magazine, IEEE*, 23(5):40–49, 2004.