

Lappeenranta University of Technology
Faculty of Industrial Engineering and Management
Degree Program in Computer Science

Bachelor's Thesis

Teemu Huovinen

**SENSITIVITY OF RETINAL IMAGE SEGMENTATION ON
GROUND TRUTH ACCURACY**

Examiners: Lasse Lensu D.Sc. (Tech.)

Supervisor: Lasse Lensu D.Sc. (Tech.)

ABSTRACT

Lappeenranta University of Technology
Faculty of Industrial Engineering and Management
Degree Program in Computer Science

Teemu Huovinen

Sensitivity of retinal image segmentation on ground truth accuracy

Bachelor's Thesis

2014

28 pages, 13 figures, 1 table, and 2 appendices.

Examiners: Lasse Lensu D.Sc. (Tech.)

Keywords: eye fundus, image segmentation, sensitivity analysis, ground truth

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto
Tuotantotalouden tiedekunta
Tietotekniikan koulutusohjelma

Teemu Huovinen

Silmänpohjakuvien segmentoinnin herkkyyssä asiantuntijatiedon tarkkuudelle

Kandidaatin työ

2014

28 sivua, 13 kuva, 1 taulukko ja 2 liitettä.

Tarkastajat: TkT Lasse Lensu

Hakusanat: silmänpohja, segmentointi, herkkyyssanalyysi, pohjamerkintä

Keywords: eye fundus, image segmentation, sensitivity analysis, ground truth

PREFACE

I would like to thank the Laboratory of Machine Vision and Pattern Recognition of Lappeenranta University of Technology for giving me this opportunity to carry out my research. I would like to give special thanks to D.Sc. Lasse Lensu for the endless guidance and support during my summers as an intern.

I would like to thank University of Bristol for their retinal image database and its accurate ground truth data, which was extensively used in this thesis.

Lappeenranta, October 19th, 2014

Teemu Huovinen

CONTENTS

1 INTRODUCTION	7
1.1 Background	7
1.2 Objectives and Restrictions	7
1.3 Structure of the Thesis	8
2 RETINAL IMAGES AND THEIR SEGMENTATION	9
2.1 Structure of the eye fundus	9
2.2 Optic disc detection	10
2.3 Blood vessel detection	10
2.4 Color transform	12
2.5 Unsupervised methods	14
2.5.1 Kirsch operator	14
2.5.2 Morphological operations	14
2.6 Supervised methods	15
2.6.1 Gaussian Mixture Models	15
2.6.2 Naive Bayes	15
2.6.3 Features	16
2.7 Evaluation metrics	16
3 EXPERIMENTS AND RESULTS	18
3.1 The ground truth	18
3.2 Supervised methods	18
3.2.1 Tests	18
3.2.2 Results	19
3.3 Unsupervised methods	20
3.3.1 Finding the best parameters	20
REFERENCES	21

ABBREVIATIONS AND SYMBOLS

RGB	Red Green Blue
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FPR	False Positive Rate
FNR	False Negative Rate
CLAHE	Contrast Limited Adaptive Histogram Equalization
GMM	Gaussian Mixture Model
pdf	Probability Density Function

1 INTRODUCTION

1.1 Background

The growing amount of diabetes patients and (arguably) more importantly the estimated amount of undiagnosed patients motivate the research for an effective mass screening method for monitoring and early detection of diabetes. The most common complication of diabetes, diabetic retinopathy, causes abnormalities in the eye, and detecting these abnormalities in the eye fundus is a promising mass screening method. [1] When developing a method for detecting these abnormalities, handmade annotations of the objects in the image are used as a ground truth to train classifiers and to evaluate the results. In eye fundus image segmentation research, ground truths are usually done by medical experts of the field (ophthalmologists) marking these abnormalities, such as exudates.

Optimal ground truth would be a pixel-accurate binary representation of the abnormalities, but as ground truths are done by a human hand, such accuracy is not possible. Because the marking of an accurate ground truth takes a good amount of time and patience, it is often necessary to have to settle for rough markings of the present abnormalities. Clusters of exudates are circled, rather than each small finding specified separately.

1.2 Objectives and Restrictions

This thesis will address two main questions in its research:

- How will inaccurate ground truth affect features and segmentation methods?
- How will other features than colour perform?

This thesis focuses on exudate detection in terms of segmentation performance, and Bristol database is used as it is readily available and has accurate ground truths of exudates. To enable comparison between features, colour, edge and texture features are used. Blood vessel detection is explored only to create a mask for them. A rough method for optic disk detection is also implemented as a preprocessing step for masking reasons. Structure of the eye fundus is briefly explored for context.

Both supervised and unsupervised segmentation methods are used. In supervised methods, ground truths are used to label observations as either exudate or background. In unsupervised methods, ground truth is used to evaluate segmentation results. Best parameters for each method are chosen based on their performance.

1.3 Structure of the Thesis

Section 2 takes a look at the different features of eye fundus images, and how they are relevant in this thesis. It also explains the theory behind the applied pre-processing and segmentation methods. Section 3 describes the experiments in detail, and presents the results for each experiment.

2 RETINAL IMAGES AND THEIR SEGMENTATION

2.1 Structure of the eye fundus

The function of the human eye on an abstract level can be compared to a camera. Light reflected from an object passes through the cornea, pupil and lens and is focused on the retina (the inner part of the eye). It is then processed by photoreceptors and then transmitted to the brain via the optic nerve. [2] Rather than covering the eye as a whole, this chapter will focus on the structure of the eye fundus.

The eye fundus is the area located in the retina, at the very back of the eye. The location of the eye fundus is illustrated in Figure 1a. Its most noticeable parts are the optic disk, the macula, and the arteries. Macula is a highly light sensitive area in the central region of the retina. Fovea is a round area located in the macula, which is densely populated by cones (cells sensitive to color). [3]

Optic disc is where the optic nerve and main arteries connect with the eye. There are no light sensitive cells inside the optic disc, which creates a blind spot in the retina. Arteries inside the eye together with the choroid (vascular layer surrounding the retina) provide nutritional support to the eye. [1] These parts of the eye fundus are illustrated in Figure 1b.

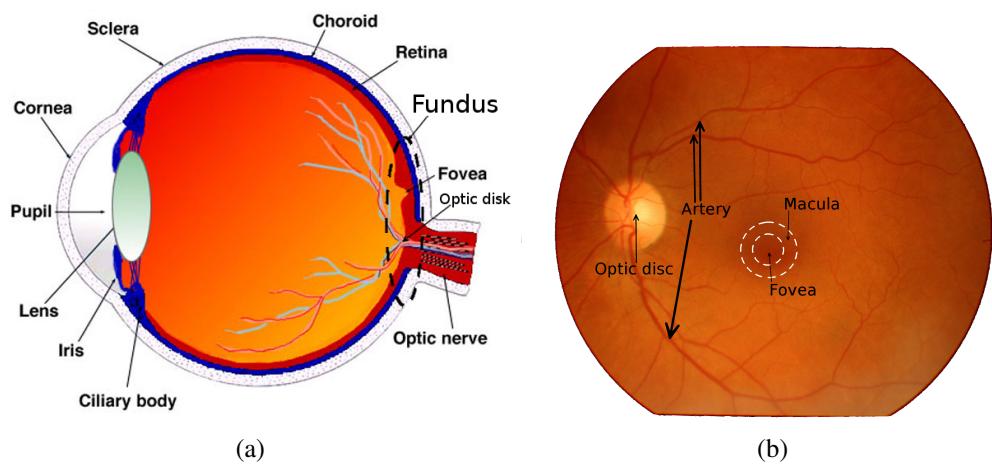


Figure 1. Structure of the human eye (a) Cross section of the eye (modified from [3]) (b) Structure of the fundus

2.2 Optic disc detection

Optic disc is very similar to exudates in terms of color and intensity, so detection and masking of the optic disk is an important preprocessing step in exudate detection. There are papers dedicated to the localisation of the optic disk [4], and it is also covered in papers concerning the detection of other parts of the eye fundus, such as exudates [5].

This method is based on the brightness of the optic disk, and the vertical blood vessels inside it. The horizontal image gradient is calculated using Sobel operator, the result is shown in Figure 2a. Image is then divided into slightly overlapping square areas with a side of 140 pixels (size is adjusted when operating close to image borders). The area with the highest sum of gradients is considered as region of interest, i.e. to hold the optic disk. This is because the dark blood vessels inside the bright optic disk result in a strong horizontal gradient. Images with a “camera glare”, i.e. a high intensity strip in the corner of the eye fundus are problematic, as that area also has a high horizontal gradient. Region of interest is shown in Figure 2b.

Inside this area of highest sum of horizontal gradients, the pixel with the highest intensity is considered to be inside the optic disk. This pixel is then used as a center of a circle that will mask out the optic disk. Final masking result is shown in Figure 2c.

2.3 Blood vessel detection

The issue of blood vessel detection in fundus images has been a popular topic of research [6] [7] [8]. In the context of this thesis, however, the purpose of blood vessel detection is to create a mask, and to use that mask to remove false positives from exudate segmentation results. For example, edge detection techniques often highlight the borders of vessels as well as exudates. For this purpose, accurate vessel detection itself is not important and including other dark areas of the image is even beneficial.

The mask is formed by first using CLAHE (contrast limited adaptive histogram equalization) [9] to enhance contrast in the green channel of the image, the result for this is shown in Figure 3a. This contrast enhanced image is then thresholded with Otsu’s method [10], which separates the image into foreground and background by minimizing the intra-class variance. This results in all the vessels and other darker areas showing as black (or background), and all brighter areas as white (foreground). This is shown in Figure 3b. To create a binary mask of the darker areas, we use the complement of this thresholded im-



Figure 2. Locating and masking the optic disk: (a) Horizontal gradient (b) Region of interest (c) Optic disk masked out

age. Final version of the mask is shown in Figure 3c.

This method is inadequate for blood vessel detection as it also includes other darker areas of the image, such as the fovea. As a mask however, it clearly reduces the amount of false positives in exudate segmentation results. It also doesn't remove true positives, as only the darker areas of the image are included in the mask.



Figure 3. Creating the blood vessel mask: (a) CLAHE (b) Thresholding using Otsu's method (c) Final mask, complement of thresholded image

2.4 Color transform

The human eye is capable of correcting the effect of varying light sources (illuminants) in perception of color [11]. In contrast, an image of an object taken under different light sources is perceived differently in terms of color by a computer. This is why color features need to be normalized in image sets where there's high variation in color properties.

The illumination and general color properties of the eye fundus images in Bristol database vary quite heavily. To be able to effectively use color features in teaching a classifier, the variance between image needs to be minimized. This is achieved by estimating the

illuminant by applying the gray-world assumption [12]: *the average reflectance in a scene under a neutral light source is achromatic*. Color feature normalization was then done by multiplying each channel with a coefficient defined as $K_r = I_{avg}/r_{avg}$, where I_{avg} is the mean of all RGB-values in the image, and r_{avg} is the mean of the specific channel. Results are shown in Figure 4.

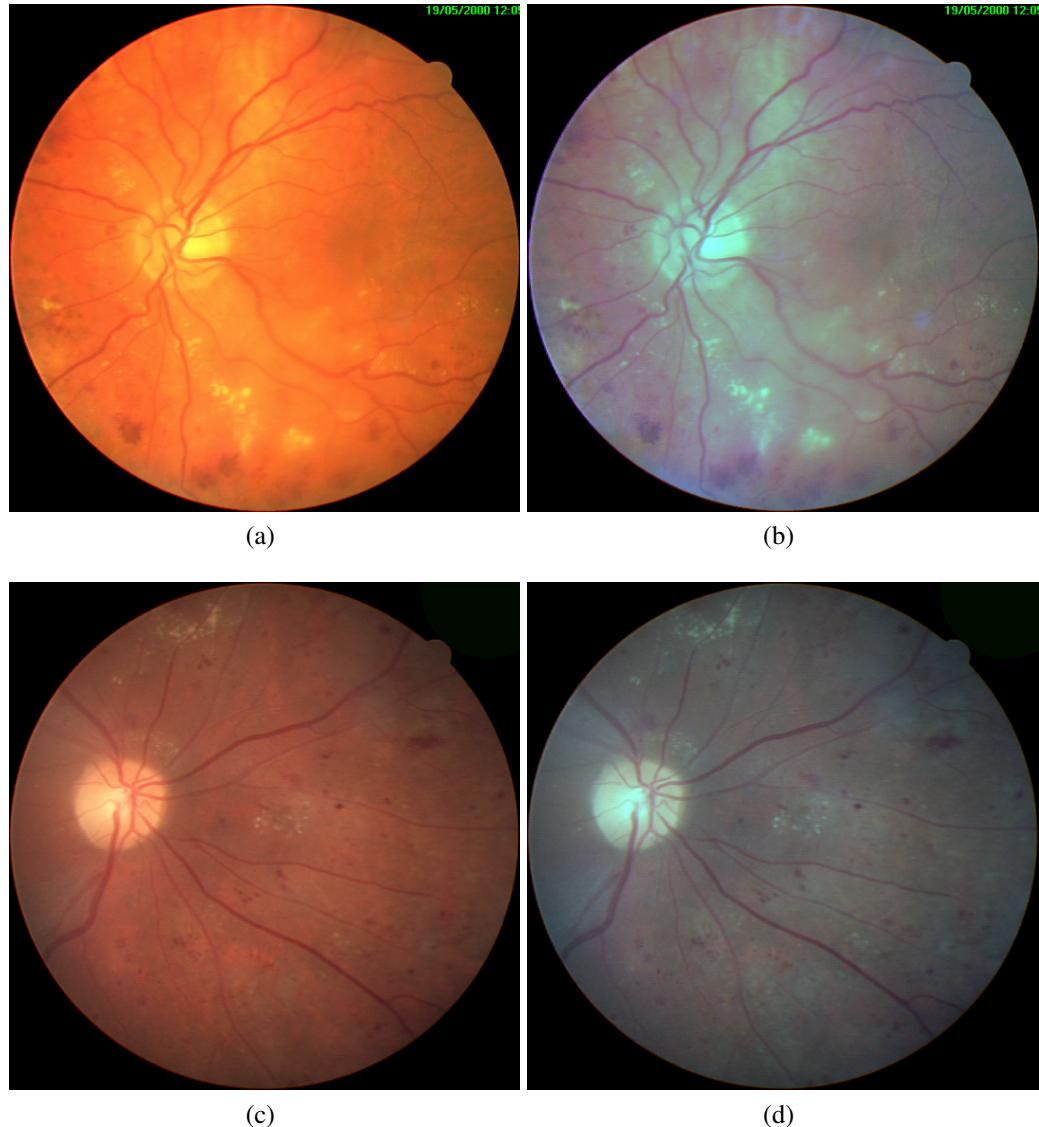


Figure 4. Color space transform: Original images on the left, adjusted images on the right.

2.5 Unsupervised methods

2.5.1 Kirsch operator

Edge detection operators that only detect gradients in specified directions are called compass operators. Kirsch operator applies a compass operator in eight directions by rotating a mask in 45° shifts. [13] It is defined as:

2.5.2 Morphological operations

Mathematical morphology operations use a structuring element to perform an operation on an input image. Most morphology operations are based the two basic morphological operations, dilation and erosion. Top-hat transform is an operation that highlights bright areas in an image, whereas bottom-hat (also known as black top-hat) operation highlights the dim areas of an image. [14] These transforms are defined as follows:

$$\text{TopHat}(f) = f - (f \circ B) \quad (1)$$

$$\text{BottomHat}(f) = (f \bullet B) - f \quad (2)$$

where f is the input image, B is the structuring element, \circ denotes the opening operation and \bullet denotes the closing operation. Opening and closing are operations based on dilation and erosion, and they are defined in [14].

The exudate detection method proposed in [15] was also implemented, except for the optic disc detection, for which the method described in 2.2 was used. The method is designed to use top-hat operation to highlight the exudates, and bottom-hat operation to highlight and ultimately remove dim areas, e.g vessels from the results. The method is defined as follows:

$$F(f) = \text{TopHat}(f) - \text{BottomHat}(f) \quad (3)$$

where $Tophat(f)$ and $BottomHat(f)$ denote the Equations 1 and 2, respectively. During the computations performed for this thesis, thresholding is applied after these operations even though it is not mentioned in [15].

2.6 Supervised methods

2.6.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a single probability density function (pdf) that describes a weighted sum of multiple Gaussian pdfs [16]. When using a GMM to model image classes (more specifically, class conditional pdfs), parameters of the underlying Gaussian pdfs are estimated by fitting the model to training data. After the formation of the model, it is possible to use Bayesian classification to classify pixels to the given classes.

Application, parameter estimation and performance of a classifier using GMMs and Bayesian classification is discussed in [17]. GMMBayes Toolbox [18] is a MATLAB implementation of said methods. For parameter estimation, Expectation Maximization (EM), Figuero-Jain (FJ) and greedy EM algorithms are implemented in the Toolbox and described in both [17] and [18]. The main difference between these algorithms is that EM requires a fixed amount of Gaussian components, while FJ estimates the amount (maximum amount is fixed in the implementation). In this thesis, only Figuero-Jain is used.

2.6.2 Naive Bayes

Naive Bayes is a probabilistic classifier that gets its name from the assumption that all features are independent from each other. Each feature contributes to the classification regardless of the absence or presence of other features. As the name also implies, the Naive Bayes classifier uses the Bayes' theorem as a basis of classification. [19] - The Naive Bayes classifier calculates the posterior probability for the sample to belong in each of the possible classes based on its features, and then classifies the sample to the class with the highest posterior probability. The classes are modeled with a single probability density function, and in the case of binary pixel classification, classes are foreground and background. In this thesis, exudates represent the foreground and the rest is considered background. The posterior probability is calculated with the Bayes' theorem as follows:

$$p(C_1) = \frac{P(C_1)p(\bar{x}|C_1)}{P(C_1)p(\bar{x}|C_1) + P(C_2)p(\bar{x}|C_2)} \quad (4)$$

where $p(C_1)$ is the posterior probability and $P(C_1)$ is the prior probability for the sample to belong in class one, and $p(\bar{x}|C_1)$ denotes the conditional probability of the feature vector having the values of the sample pixel, if it were of class one.

2.6.3 Features

Color, edge and texture features were used to teach the classifiers of the supervised methods. Color was extracted from the image as-is. Edge was represented with the differential of the maximum and minimum value in a 3-by-3 neighborhood (implemented with *rangefilt*-function in MATLAB [20]). Location invariant local binary patterns (LBP) [21] were used for texture.

2.7 Evaluation metrics

In evaluation, correctly classified samples can be split into true positives and true negatives. Errors are then similarly split into false positives and false negatives. In this thesis, true positives are findings correctly classified as exudates in the segmentation results, and true negatives are consequently findings correctly classified as background. False negatives are classified as background in results, but are actually positive (classified as exudate) in ground truth, and false positives are classified as positive in results, when they are in fact background [22]. These terms are illustrated with

Sensitivity represents the amount of samples correctly classified as positive. Its values range from 0 to 1, where 1 means all foreground samples were correctly classified. Specificity in turn represents the amount of background pixels correctly described as negative. Its values also range from 0 to 1, where 1 means all background pixels were correctly classified. [22] They are defined as:

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

The Dice coefficient [23], Jaccard index and F-score all measure set agreement, representing the success of segmentation in a single number. All of these coefficients can have values from the range [0,1], where 0 indicates no similarities, and 1 indicates perfect agreement. The coefficients are defined as follows:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \quad (7)$$

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

$$F-score = \frac{2PR}{P + R} \quad (9)$$

where A is the segmented set, and B is the ground truth, P stands for precision and R stands for recall (sensitivity).

3 EXPERIMENTS AND RESULTS

3.1 The ground truth

The ground truth of exudates in Bristol database is very accurate, and to enable comparison of results and sensitivity analysis, more inaccurate ground truths were made by hand. Instead of the original color images, markings for the inaccurate ground truth were made on the black-and-white images of Bristol ground truth. This was to ensure every exudate present in the Bristol ground truth was also present in the inaccurate ground truth. Markings were done in a way that estimated the way doctors marked their findings when given the freedom to make inaccurate markings. Essentially this means that clusters of exudates are grouped together, and single exudates were more loosely circled. This is illustrated in Figure 5.

To create a basis for more comprehensive testing, different stages of inaccurate ground truths were created by dilating the accurate ground truth. Three different stages were created; one, three and five iterations of dilation by a disk-shaped structuring element, with a radius of 1 pixel. The dilation was restricted to stay inside the inaccurate ground truth.

The accuracy of ground truth has a direct impact on training and performance of supervised methods. In general, inaccurate ground truth will mean a significant amount of background samples close to exudates will be categorized as positive. This will result in higher variation of feature distributions and an increase of false positives.

3.2 Supervised methods

3.2.1 Tests

Preprocessing was kept to a minimum as only the timestamp was removed from the eye fundus images. As the optic disk's colour is very close to the exudates, a mask was generated with the method described in Sec. 2.2. The eye outline was also included in the mask, as it was included in the results by edge detection methods. Vessels were not included in the mask.

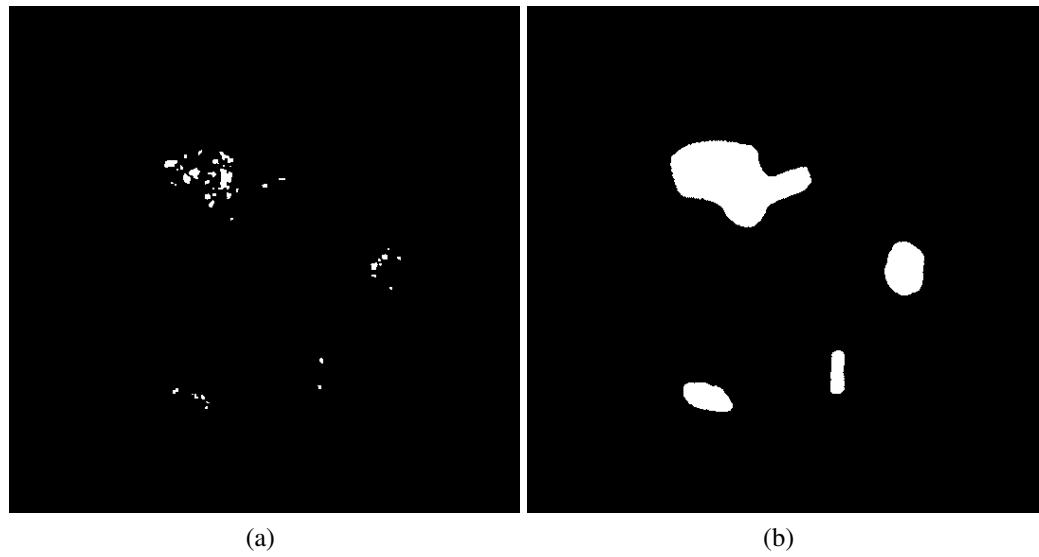


Figure 5. Ground truths: (a) Original accurate ground truth (b) Inaccurate ground truth

The amount of background pixels in the images is considerably higher than the amount of exudate pixels. This was balanced by applying sampling to the training data, using equal amounts of background and exudate pixels. All exudate pixels were used, and background pixels were selected randomly.

Classifiers were trained with colour, edge and texture features, or a combination of the three. Feature extraction methods are described in Sec. 2.6.3. No contrast enhancement was used, so raw colour info from the image was used for the colour feature.

The Bristol image set was split in half, using the other half for training and the other half for evaluation, and then vice versa. This was done to study the effect the selection of training data has on the results. Classifiers were trained with each level of inaccurate ground truth described in Sec. 3.1. Two different image sets were used; the original Bristol images, and the color-normalized images. Colors were adjusted with the method described in Sec. 2.4, with the objective of reducing the color variation in the images. With this setup, each calculation used a specific ground truth accuracy, image set, and teaching set.

3.2.2 Results

KUVAKOHTAISET TULOKSET SUPERVISED METHODS, LISÄÄ APPENDIX VIITE

GRAY WORLD TULOKSISTA VIELÄ KUVAAJIA, tarkista gray world kappale

Using inaccurate ground truth is expected to result in more false positives in segmentation results. This can be seen in the results as an increase in sensitivity and a decrease in specificity with all features and classifiers when the level of inaccuracy in ground truth increases. This is also visible in the image-specific results show in **APPENDIX VIITE**.

Comparing the sensitivity and specificity of Naive-Bayes and GMM-Bayes classifiers shown in Figures 6 and 7, a slightly higher overall performance of GMM-Bayes classifiers can be seen. The Naive-Bayes edge-classifier is an exception, clearly outperforming any other single-feature trained classifier or even the classifier with all features included. Based on the specificity of the texture-based classifiers, their performance was below the others. The image-specific results explain this, as the texture-based classifier behaves similarly to a random classifier. This could be because of the feature's variation in the training set, but this was not confirmed.

Color normalization didn't have a big impact on the results. The expected improvement on the performance of color-based classifiers is minimal, and the performance of edge-based classifiers was slightly worse.

3.3 Unsupervised methods

3.3.1 Finding the best parameters

When researching feature sensitivity to ground truth accuracy, standard segmentation with unsupervised methods isn't really useful as it doesn't use the ground truth. In this thesis, the ground truth is used in finding the best parameters for unsupervised methods. First, the images used for teaching are segmented with each method and a large set of parameters. The results are then evaluated with the ground truth.

For this purpose, it was desired to have a single value describing the "goodness" of segmentation results to enable easier comparison and ranking of results. Different values, such as sensitivity, specificity and precision (defined in Sec. 2.7) were considered, but these were discarded as they only described one aspect of the result, either the amount of samples correctly positive or negative. The used coefficient would have to describe the "goodness" in a more wholesome way, and for that reason, the Dice coefficient, Jaccard index and F-score were considered (defined in Sec. 2.7).

Dice coefficient and F-score values are interchangeable in practice as their values are exactly equal. Jaccard index behaves similarly to the two, though being considerably more critical. The use of Jaccard was discarded because of its criticality, and the choice between Dice and F-score was arbitrary. Dice coefficient was chosen as it is simple to implement.

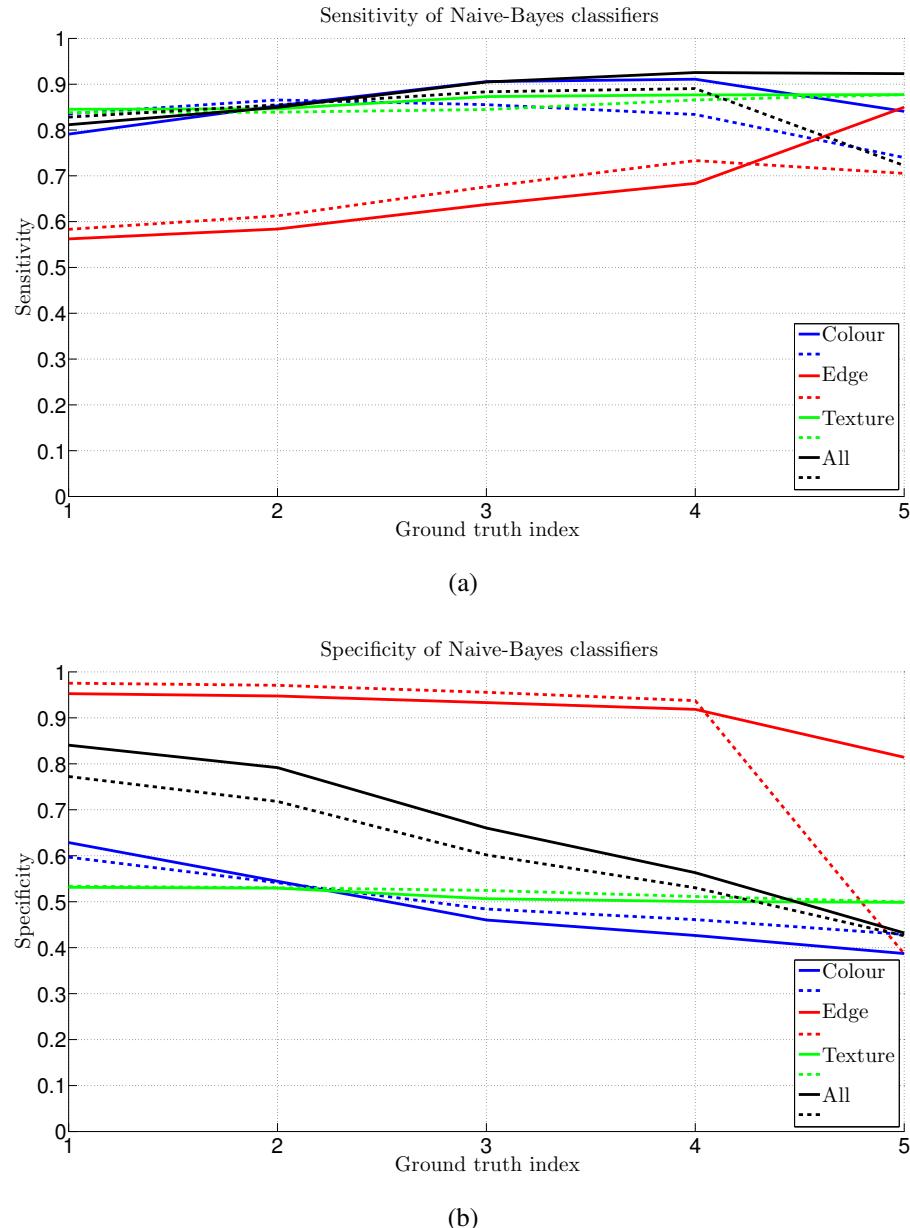


Figure 6. Performance of Naive-Bayes classifiers: (a) Sensitivity (b) Specificity. Note that these graphs represent the methods' performance over images that have exudates present. Original Bristol ground truth is on the left on the x-axis, and the most inaccurate ground truth is on the right.

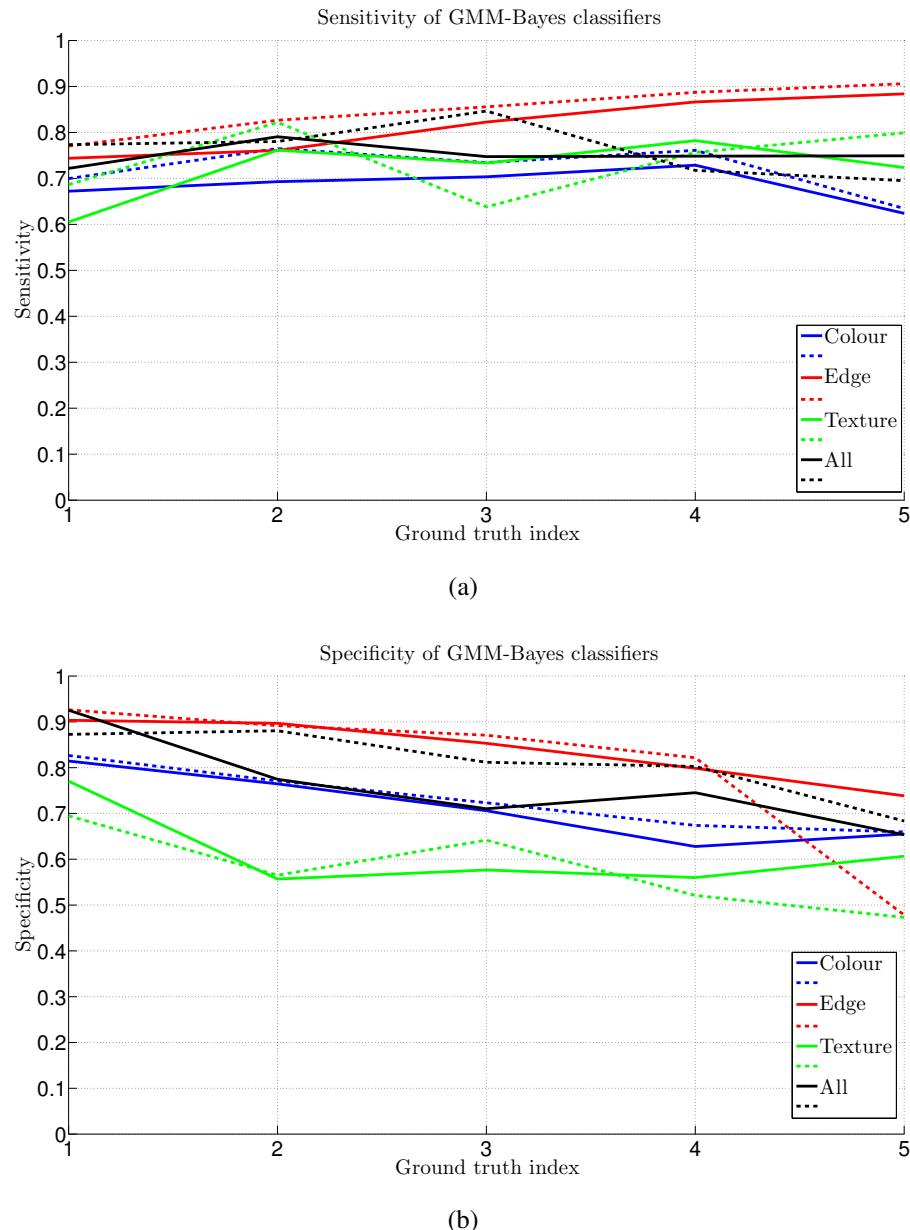


Figure 7. Performance of Naive-Bayes classifiers: (a) Sensitivity (b) Specificity. Note that these graphs represent the methods' performance over images that have exudates present. Original Bristol ground truth is on the left on the x-axis, and the most inaccurate ground truth is on the right.

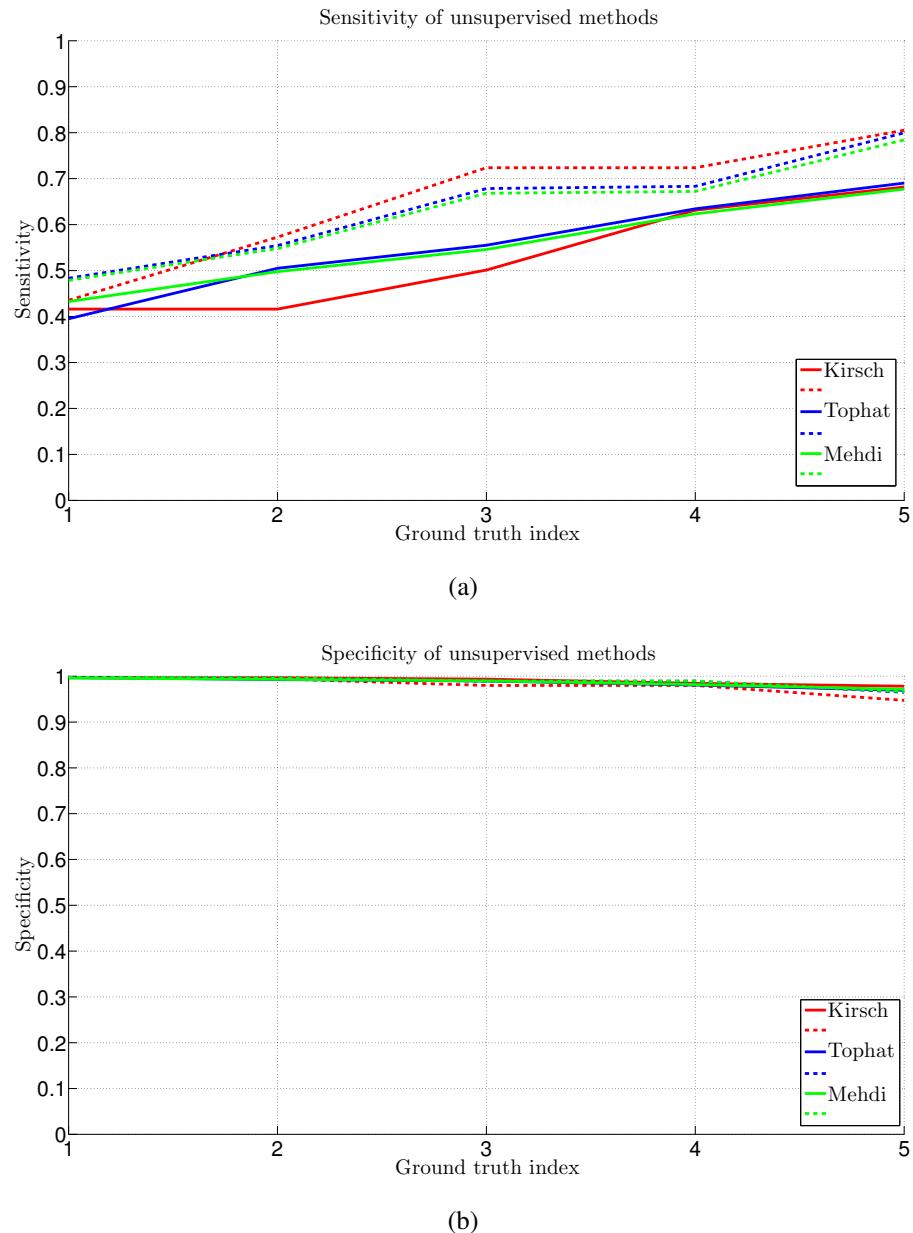


Figure 8. Performance of unsupervised methods: (a) Sensitivity (b) Specificity

REFERENCES

- [1] Tomi Kauppi et al. *Eye fundus image analysis for automatic detection of diabetic retinopathy*. Lappeenranta University of Technology, 2010.
- [2] Brian A Wandell. *Foundations of vision*. Sinauer Associates, 1995.
- [3] Helga Kolb. Simple Anatomy of the Retina. <http://webvision.med.utah.edu/book/part-i-foundations/simple-anatomy-of-the-retina/>. Accessed: 2014-09-30.
- [4] Sribalamurugan Sekhar, Waleed Al-Nuaimy, and Asoke K Nandi. Automated localisation of retinal optic disk using hough transform. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 1577–1580. IEEE, 2008.
- [5] Thomas Walter, J-C Klein, Pascale Massin, and Ali Erginay. A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina. *Medical Imaging, IEEE Transactions on*, 21(10):1236–1243, 2002.
- [6] Frédéric Zana and Jean-Claude Klein. A multimodal registration algorithm of eye fundus images using vessels detection and hough transform. *Medical Imaging, IEEE Transactions on*, 18(5):419–428, 1999.
- [7] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram van Ginneken. Ridge-based vessel segmentation in color images of the retina. *Medical Imaging, IEEE Transactions on*, 23(4):501–509, 2004.
- [8] Meindert Niemeijer, Joes Staal, Bram van Ginneken, Marco Loog, and Michael D Abramoff. Comparative study of retinal vessel segmentation methods on a new publicly available database. In *Medical Imaging 2004*, pages 648–656. International Society for Optics and Photonics, 2004.
- [9] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44, 2004.
- [10] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- [11] David H Foster. Color constancy. *Vision research*, 51(7):674–700, 2011.

- [12] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.
- [13] Anil K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.
- [14] Xiangzhi Bai, Fugen Zhou, and Bindang Xue. Image enhancement using multi scale image features extracted by top-hat transform. *Optics & Laser Technology*, 44(2):328–336, 2012.
- [15] Mehdi Ghafourian Fakhar Eadgahi and Hamidreza Pourreza. Localization of hard exudates in retinal fundus image by mathematical morphology operations. In *Computer and Knowledge Engineering (ICCKE), 2012 2nd International eConference on*, pages 185–189. IEEE, 2012.
- [16] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663, 2009.
- [17] Pekka Paalanen, Joni-Kristian Kamarainen, Jarmo Ilonen, and Heikki Kälviäinen. Feature representation and discrimination based on gaussian mixture model probability densitiesâpractices and algorithms. *Pattern Recognition*, 39(7):1346–1358, 2006.
- [18] Pekka Paalanen, Joni-Kristian Kamarainen, and Heikki Kälviäinen. Bayesian classification using gaussian mixture model and em estimation: implementations and comparisons. 2004.
- [19] Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.
- [20] MATLAB rangefilt-function. <http://se.mathworks.com/help/images/ref/rangefilt.html>. Accessed: 2014-11-18.
- [21] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *Computer Vision-ECCV 2000*, pages 404–420. Springer, 2000.
- [22] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [23] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Appendix 1. Image-wise results

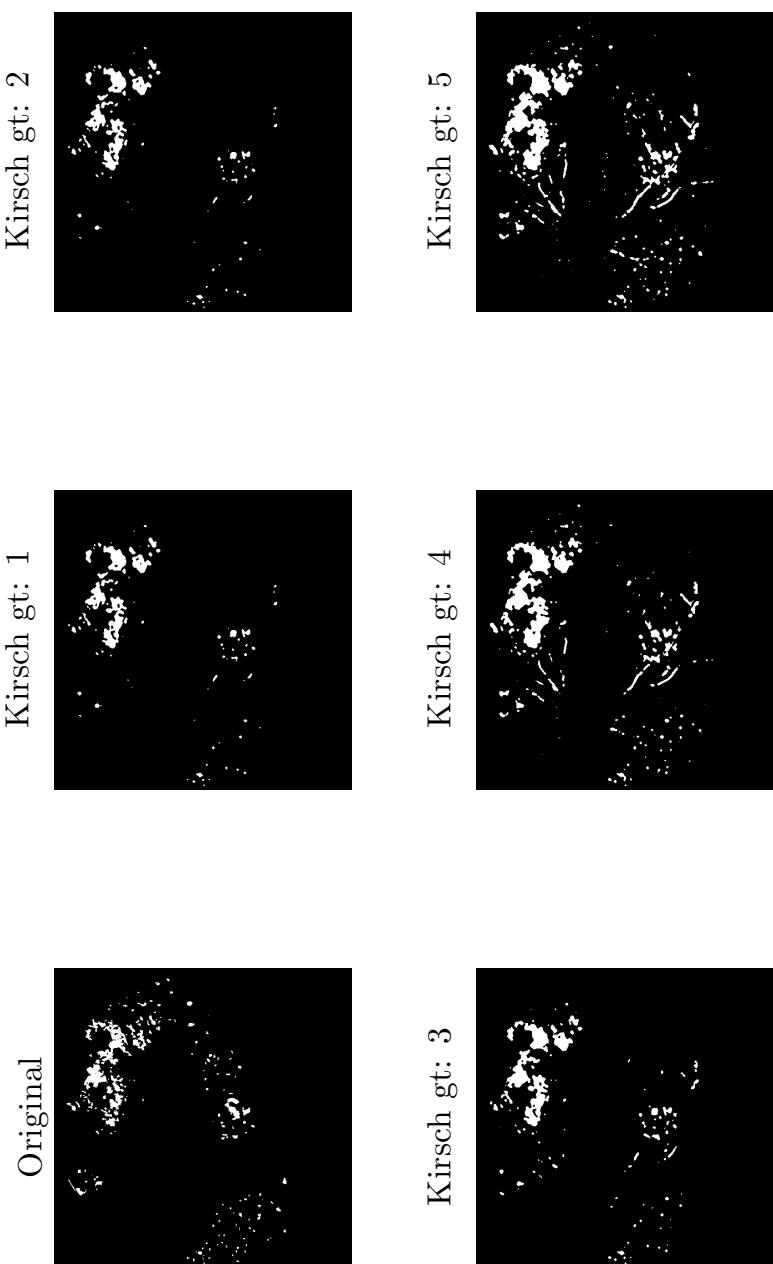


Figure A1.1. Imagewise results of Kirsch

(continues)

Appendix 1. (continued)

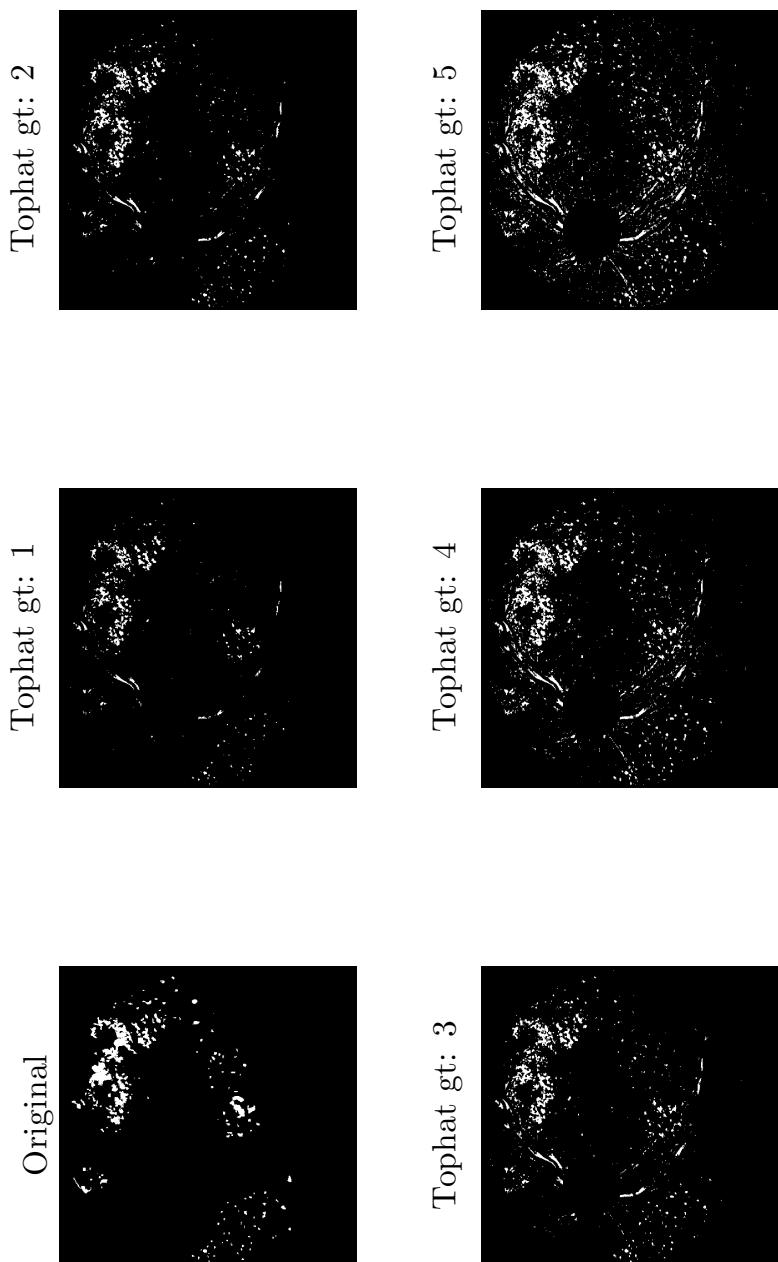


Figure A1.2. Imagewise results of Tophat

(continues)