
Identifying and Categorizing Anomalies in Retinal Imaging Data

Philipp Seeböck, Sebastian Waldstein, Sophie Klimscha, Bianca S. Gerendas,
René Donner, Thomas Schlegl, Ursula Schmidt-Erfurth and Georg Langs

Christian Doppler Laboratory for Ophthalmic Image Analysis, Department for Ophthalmology
CIR - Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna
philipp.seeboeck@meduniwien.ac.at

1 Introduction

The detection of diagnostically relevant markers in imaging data is critical in medicine and treatment guidance. Typically, detectors are trained based on a priori defined categories, and annotated imaging data. This makes large-scale annotation necessary which is often not feasible, limits the use to known marker categories, and overall slows the process of discovering novel markers based on available evidence. Additionally, in contrast to natural images such as photographs, the relevant categories of retinal images cover only a small fraction of the overall imaging data, though they are the focus of diagnostic attention, and are often dominated by natural variability of even healthy anatomy.

Optical Coherence Tomography (OCT) [11] is an important diagnostic modality in ophthalmology and offers a high-resolution 3D image of the layers and entities in the retina. Each position of the retina sampled by an optical beam results in a vector, the A-scan. Adjacent A-scans form a B-scan, which in turn form the entire volume. Anomaly detection [12] in retinal images is a difficult and unsolved task, though many patients are affected by retinal diseases that cause vision loss (e.g. age-related macular degeneration has a prevalence of 9% [19]). We propose to identify abnormal regions which can serve as potential biomarkers in retinal spectral-domain OCT (SD-OCT) images, using a Deep Convolutional Autoencoder (*DCAE*) trained unsupervised on healthy examples as feature extractor and a *One-Class SVM* to estimate the distribution of normal appearance. By using only healthy examples for training, we omit the need of collecting a dataset that represents all the variabilities which may occur in the data and contains sufficient amount of anomalies.

Results show that the trained model is able to achieve a dice of 0.55 regarding annotated pathologies in OCT images. Since there are different pathologies and structures occurring in the regions detected as anomalous, a meaningful sub-classification of these areas is of particular medical interest. Therefore we further cluster regions identified as anomalous in order to retrieve a meaningful segmentation of these areas. In addition, we evaluate to which extent the learned features can be used in a classification task, where intraretinal cystoid fluid (IRC), subretinal fluid (SRF) and the remaining part of the retina are classified. A classification accuracy of 86.6% indicates the discriminative power of the learned feature representation.

Related Work Deep Convolutional Neural Networks (CNNs) in combination with large quantities of labeled data have recently improved the state-of-the-art in various tasks such as image classification [16] or object detection [13]. CNNs can automatically learn translation invariant visual input representations, enabling the adaptation of visual feature extractors to data, instead of manual feature engineering. While purely supervised training of networks is suitable if labeled data is abundant, unsupervised methods enable the exploitation of unlabeled data [2, 4, 5, 20], discovering the underlying structure of data.

Doersch *et al.* [4] use spatial context as supervisory signal to train a CNN without labels, learning visual similarity across natural images, which does not necessarily extend to medical domain. Dosovitskiy *et al.* [5] train a CNN unsupervised by learning to discriminate between surrogate image classes which are created by data augmentation. A limitation of this approach is that it does not scale

to arbitrarily large amounts of unlabeled data. Zhao *et al.* [20] propose joint instead of layer-wise training of convolutional autoencoders, but they perform experiments in a semi-supervised setting.

A variety of anomaly detection techniques are reported in the literature. Carrera *et al.* [1] use the technique of convolutional sparse models to detect anomalous structures in texture images. In the work of Erfani *et al.* [6] Deep Belief Networks are combined with One-Class SVMs in order to address the problem of anomaly detection in real-life datasets. In contrast to our paper, these works address problems regarding natural images and real-life datasets, which have considerable different characteristics compared to medical images, as explained above. An overview for anomaly detection methods can be found in [12].

Regarding unsupervised learning in OCT images, Venhuizen *et al.* [18] train random forests with features from a bag-of-words approach in order to classify whole OCT volumes, as opposed to our pixel-wise segmentation approach. Schlegl *et al.* [14] use weakly supervised learning of CNNs to link image information to semantic descriptions of image content. In contrast to our work, the aim is to identify a priori defined categories in OCTs using clinical reports.

2 Method

The following preprocessing is applied to all volumes. First we identify the top and bottom layer of the retina using a graph-based surface segmentation algorithm [7], where the bottom layer is used to project the retina on a horizontal plane. Then each B-scan is brightness and contrast normalized. Finally we perform over-segmentation of all B-scans to monoSLIC superpixels sp of an average size of 4×4 pixels [9].

To capture visual information at different levels of detail, we use a multi-scale approach to perform superpixel-wise segmentation of the visual input. We conduct unsupervised training of two DCAEs on pairs of extracted patches sampled at the same positions for two scales (32×32 patches for $DCAE_1$ and down-sampled 128×32 to 32×32 patches for $DCAE_2$) in parallel. The used network architecture for both scales is 512c9-3p-2048f-512f for the encoder, implying a matching decoder structure that uses deconvolution and unpooling operations. All layers except pooling and unpooling are followed by Exponential Linear Units (ELUs) [3]. The loss function for training is defined as the Mean Squared Error function $MSE(x, \hat{x})$, where x denotes the input patch and \hat{x} the reconstructed input. In addition, we use dropout in each layer, which corresponds to unsupervised joint training with local constraints in each layer.

The feature representations of both scales are concatenated and used as input for training a Denoising Autoencoder (DAE), its single-layer architecture denoted as 256f. All three models together form our final model $DCAE_{ent}$ that gives us a 256 dimensional feature representation z for a specific superpixel in the B-Scan.

We then train the One-Class SVM [15] using a linear kernel to find a boundary that describes the distribution of healthy examples in the feature space z , which serves as decision boundary for unseen data. New samples can be classified either as coming from the same data distribution if lying within the boundary (normal) or not (anomaly). For each OCT, features z and the corresponding class are computed for each superpixel lying within the top and bottom layer of the retina. This provides a segmentation of the retina into two classes.

Subsequently, we use spherical K-means clustering [10] with cosine distance to sub-segment regions which have been identified as anomalous in the former step into multiple clusters c . The number of cluster centroids is determined by an internal evaluation criterion called Davies-Bouldin (DB) index [8], where a small value indicates compact and well separated clusters. To segment an unseen OCT, each superpixel with the property "anomalous" gets a cluster assignment, according to the nearest cluster centroid in the feature space z .

3 Evaluation

The primary purpose of the evaluation is to test if we can identify novel marker candidates in imaging data algorithmically, instead of relying on a priori defined object categories. We evaluated (1) if we can segment anomalies, (2) if we can find categories of these regions that correspond to a fine-grained ontology of known findings, and (3) if the learned features have discriminative power for image classification.

Table 1: Dice, Precision and Recall for anomalous regions with ground-truth annotations. The same One-Class SVM settings are used for all methods.

Algorithm	Dice	Precision	Recall
PCA_{256}	0,33	0,31	0,39
$PCA_{0.95}$	0,32	0,32	0,35
$DCAE_{ent}$	0,55	0,53	0,58

Table 2: Mean classification accuracies of L2-SVMs on the generated OCT-dataset with balanced classes, trained with features from different models.

Algorithm	Accuracy (in percent)			
	IRC	SRF	Other	Overall
PCA_{256}	71.9	74.0	73.9	73.4 (\pm 3.9)
$PCA_{0.95}$	73.3	76.9	70.0	73.4 (\pm 3.9)
$DCAE_{ent}$	87.3	88.3	84.1	86.6 (\pm 1.6)

We used scans from 704 patients, where 283 healthy OCT volumes were used to create the healthy training set (277,340 pairs of image patches) for the $DCAE$ and the SVM, 411 unlabeled OCTs were used to create the anomaly training set (295,920 patches) for clustering, and the validation and test set consisting of 5 volumes each, with voxel-wise ground truth annotations of anomalous regions and additional annotations for specific pathologies (IRC, SRF), were used for model selection and evaluation, respectively. The scans were acquired using Spectralis OCT instruments (Heidelberg Engineering, GER) and have a resolution of $512 \times 496 \times 49$ depicting a $6mm \times 2mm \times 6mm$ volume of the retina, where the distance between voxel centers is about $11\mu m$ in the first, $4\mu m$ in the second, and $120\mu m$ in the third dimension.

The learned anomaly detection model is evaluated on the test set, where *Dice*, *Precision* and *Recall* are calculated for anomalous regions regarding the ground-truth annotation. Interpretation of these quantitative values is done carefully and only in combination with qualitative evaluation by clinical retina experts.

In addition, we compare our $DCAE_{ent}$ model with conventional PCA embedding. To ensure fair comparison, we train two models. In the first model PCA_{256} , the dimensionality is chosen to match the feature dimension z of our proposed model. For both scales, the first 128 principal components are kept and concatenated to obtain z . In the second model $PCA_{0.95}$, for each scale the first components that describe 95% of the variance in the dataset are kept.

For the categorization of anomalous regions the number of clusters is varied between 2 and 30, where the model with the lowest DB-index on the anomaly training set is selected and applied to the test set. We qualitatively evaluate if the categories found in the regions identified as abnormal are clinically meaningful, to provide a link to actual disease.

Beside anomaly detection, we test the learned models in a classification task in order to evaluate the discriminative power of the learned feature representation. Here, we train a linear Support Vector Machine (L2-SVM) on a balanced classification dataset with 15,000 labeled examples and three classes (33.3% IRC, 33.3% SRF, 33.3% remaining part of the retina) in feature space z . 5-fold-cross-validation is performed so that samples of a patient are only present in one fold, where this labeled dataset is only used to train the L2-SVM.

4 Results

We report quantitative and qualitative results illustrating anomaly segmentation, visualize anomaly categorization outcome, provide descriptions of clusters according to the experts and describe results regarding the classification task.

As can be seen in Table 1 our method achieves a dice of 0.55, which is a clear improvement in comparison with both PCA_{256} and $PCA_{0.95}$. This is also reflected by the visualization provided in Figure 1 (a)-(d), which shows substantial overlap between ground truth annotations and $DCAE_{ent}$ segmentation. $DCAE_{ent}$ provides a less diffuse segmentation compared to $PCA_{0.95}$, capturing the retinal pathology in a meaningful way.

Regarding anomaly categorization, the lowest DB-index was found for 7 clusters, as illustrated in Figure 1 (f). Figure 1 (e) shows 2D t-SNE embedding [17] of the learned $DCAE_{ent}$ features. Categories were identified and described as following by two clinical retina experts and are summarized in Figure 1 (g)-(h).

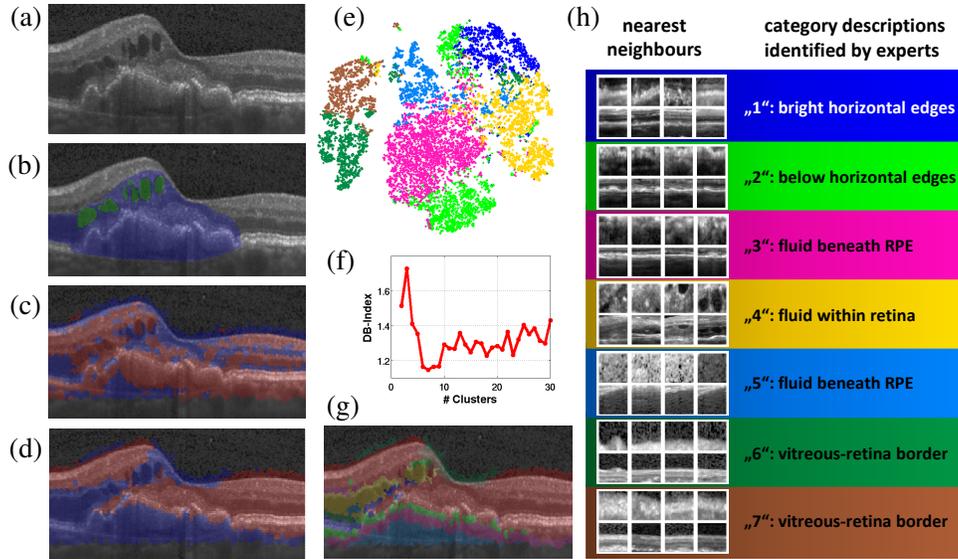


Figure 1: The same B-scan is illustrated for (a) the original scan, (b) the ground truth annotations (pathologic region in blue, IRC in green), anomaly detection results for (c) the comparison method $PCA_{0.95}$ and (d) our proposed method (normal=red, anomaly=blue). The clustering result is shown in (g), where identified anomalous regions are segmented into 7 categories. Each cluster is indicated by a separate color. The corresponding cluster descriptions which are identified by experts, are illustrated in (h) together with nearest neighbours of cluster centroids. 2D t-SNE embedding of the feature space z is shown in (e). The calculated values of the DB-Index are plotted in (f).

Bright horizontal edges are segmented in cluster "1" which is highlighted in blue. In the majority of cases this cluster corresponds to Retinal Pigment Epithelium (RPE). Cluster "2" is marked in light green and corresponds to areas below these horizontal structures. Both clusters segment areas situated next to fluid, which changes the local appearance of patches to abnormal. Cluster "4" is highlighted in yellow and corresponds to fluid within the retina. This finding is supported by the fact that 62% of manual IRC annotations located in anomalous regions are assigned to cluster "4". Marked in grey-blue and pink, Cluster "3" and "5" both segment regions that correspond to fluid beneath the RPE. Cluster "6" (dark green) and "7" (brown) highlight the border between vitreous and retina due to irregular curvature or fluid situated below, which alters the appearance of extracted patches.

The classification results are illustrated in Table 2, where mean overall accuracy as well as class specific accuracies are reported. The proposed unsupervised feature learning approach $DCAE_{ent}$ clearly outperforms both conventional methods PCA_{256} and $PCA_{0.95}$. Our proposed method achieves a mean overall accuracy of 86.6%, while the PCA-models only achieve 73.4% and 73.3%.

5 Discussion

In this paper we propose a method to detect anomalous regions in OCT images which needs only healthy training data. A deep convolutional autoencoder is used as feature extractor, while One-Class SVM is used to estimate the distribution of healthy retinal patches. In a second step, the identified anomalous regions are segmented into subclasses by clustering. Results show that our proposed method is not only capable of finding anomalies in unseen images, but also categorizes these findings into clinically meaningful classes. The identification of new anomalies, instead of the automation of expert annotation of known anomalies is a critical shift in medical image analysis. It will impact the categorization of diseases, the monitoring of treatment and the discovery of relationships between genetic factors and phenotypes. Additionally, the power of our learned feature extractor is also indicated by performance in a classification task.

Acknowledgments

This work funded by the Austrian Federal Ministry of Science, Research and Economy, and the FWF (I2714-B31). A Tesla K40 used for this research was donated by the NVIDIA Corporation.

References

- [1] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg. Detecting anomalous structures by convolutional sparse models. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.
- [2] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2015.
- [3] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
- [4] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.
- [5] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014.
- [6] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 2016.
- [7] M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka. Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *Medical Imaging, IEEE Transactions on*, 28(9):1436–1447, 2009.
- [8] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.
- [9] M. Holzer and R. Donner. Over-segmentation of 3d medical image volumes based on monogenic cues. In *Proceedings of the 19th CVWW*, pages 35–42, 2014.
- [10] K. Hornik, I. Feinerer, M. Kober, and C. Buchta. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012.
- [11] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, et al. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991.
- [12] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.
- [13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.
- [14] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.
- [15] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [17] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.
- [18] F. G. Venhuizen, B. van Ginneken, B. Bloemen, M. J. van Grinsven, R. Philipsen, C. Hoyng, T. Theelen, and C. I. Sánchez. Automated age-related macular degeneration classification in oct using unsupervised feature learning. In *SPIE Medical Imaging*, pages 941411–941411. International Society for Optics and Photonics, 2015.
- [19] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, 2(2):e106–e116, 2014.
- [20] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015.