Proceedings of the Joint Austrian Computer Vision and Robotics Workshop 2020 DOI: 10.3217/978-3-85125-752-6-36

Towards Identification of Incorrectly Segmented OCT Scans

Verena Renner and Jiří Hladůvka
Pattern Recognition and Image Processing Group
TU Wien
e1527272@student.tuwien.ac.at jiri@prip.tuwien.ac.at

Abstract. *Precise thickness measurements of retinal layers are crucial to decide whether the subject requires subsequent treatment. As optical coherence tomography (OCT) is becoming a standard imaging method in hospitals, the amount of retinal scans increases rapidly, automated segmentation algorithms are getting deployed, and methods to assess their performance are in demand.*

In this work we propose a semi-supervised framework to detect incorrectly segmented OCT retina scans: ground-truth segmentations are (1) embedded in 2D feature space and (2) used to train an outlier scoring function and the corresponding decision boundary.

We evaluate a selection of five outlier detection methods and find the results to be a promising starting point to address the given problem. While this work and results are centred around one concrete segmentation algorithm we sketch the possibilities of how the framework can be generalized for more recent or more precise segmentation methods.

1. Introduction

It is known that frequent eye screening helps to early-diagnose the diabetic macular edema (DME) [14] and therefore raises the effectiveness of needed treatments. Additionally, the number of age-related macular degeneration (AMD) patients is increasing, because of ageing population [9], as well as those suffering from DME due to the rising number of diabetes cases. OCT technology is nowadays minimally invasive, very fast, and therefore widely spread, so that a large number of OCT scans needs to be preprocessed automatically. Ophthalmological departments are developing or deploying systems to deal with the large amount of OCT data produced. One such instance to segment retinal layers from OCT scans is based on the work [5]. While accurate in most of cases, the method occasionally exhibits imperfections. An improvement is desirable, as the correct segmentation is essential for further automatic evaluation of OCT scans. This is because the thickness of the retinal layers is highly related to the presence of diseases, like AMD or DME [5]. They are caused by intraretinal and subretinal fluids, leading to a swelling of the retinal layers [10], exerting pressure on the light-receptors damaging them and thus eyesight.

Imperfections in segmentation can be caused by different reasons such as bad contrast of parts of the scan, noise, artefacts or an unsupported edge-case of the segmentation algorithm.

This work aims to support the identification of incorrectly segmented OCT scans with a two-fold purpose in mind. First, it is of interest to increase the trust of ophthalmologists in the algorithm by flagging segmentations that may potentially require manual inspection. Second, to improve segmentation algorithms, it is desirable to automatically identify incorrect segmentations of previously unseen scans and focus on improvements for such cases.

2. Dataset

A set of 100 OCT scans, each accompanied with both manual ground truth (GT) and algorithmic (A) segmentation [5] have been provided for this study. Each OCT scan is a stack of 200 1024 × 200 gray scale images. Both the ground truth and the algorithmic segmentation are available as slice-wise boundaries of 13 retina layers. Figure 1 shows boundary examples of the first retina layer (L1). There is no expert assessment available on whether the algorithmic segmentations are accepted as correct or not.

For legal issues, this dataset is currently unavailable for public use.

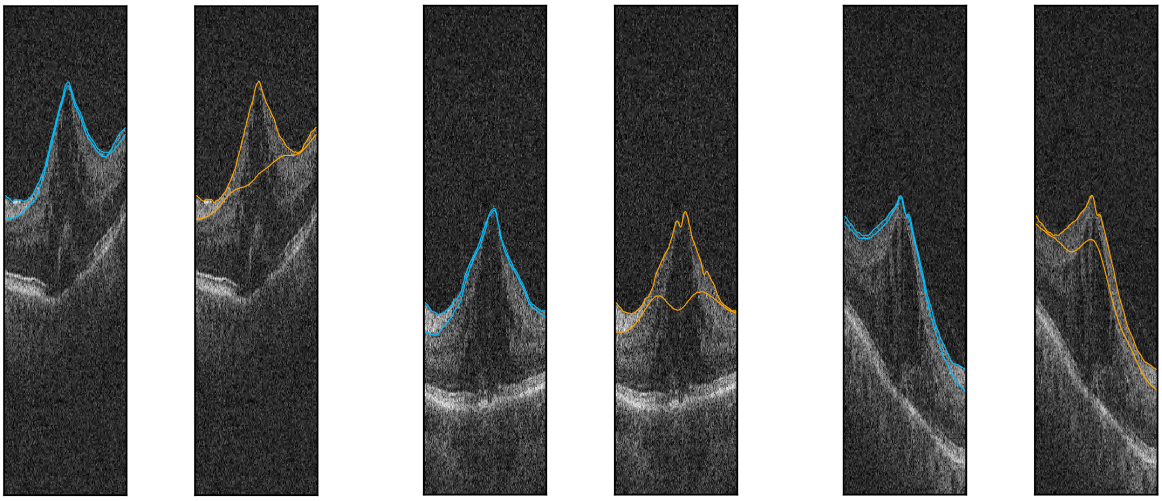

Figure 1: Three examples (scans 1, 2, and 5) of ground truth (left) and incorrect algorithmic (right) segmentations of the first retina layer (L1) in mid-stack slices.

3. Method

In order to identify incorrectly segmented OCT scans we suggest in section 3.1 to embed the segmentation results in as few dimensions as possible. While this is certainly motivated by curse of dimensionality it is additionally motivated by an increase of interpretability – ophthalmologists may desire to visually relate a particular case to cases inspected previously.

Methods of outlier detection can be divided in three branches [6]. Supervised classification, when both inliers and outliers are labeled and in balance; unsupervised when training data of both inliers and outliers are unlabeled; and semi-supervised when training data consists only of observations describing normal behavior. In section 3.2 we follow semi-supervised methods for the following reasons. First, there is no assessment of algorithmic segmentations available and we only can roughly estimate the class based on some metric (e.g., the Dice coefficient). Second, the outlier class (wrong segmentations) is expected to be under-represented. Third, it is likely there are several sources of segmentation error which could map to low-density clusters. We aim to detect outliers in low-density regions, too. We model the distribution of the inliers (correct segmentation) and compare the test points to this distribution.

3.1. Area curves and their representation

While for each retinal layer a list of region properties can be thought of, for sake of interpretability the slice-wise area values are of special interest. Furthermore, the focus of this work was restricted to layer 1. This decision is based on the observation that a segmentation error in L1 layer propagates to subsequent layers while correct L1 segmentations tend to correlate with correctly segmented scans.

For each segmented OCT, we introduce the vector $\mathbf{a} = [a_0, \dots, a_{199}]^\top$ of layer-1 area values and refer to is as the area curve. Examples of how area curves look like for both ground truth and algorithmic segmentations are given in figure 2.

Looking at the (orange) area curves calculated from the algorithmic segmentation, which are of the main interest, two types of shape appear: Those exhibiting a maximum (cf., scan 1, 2 or 5 of figure 2), or a minimum (cf. scan 0) around the middle of the slices.

In healthy eyes, the layers get thinner around the cavity of the fovea [11], causing the area curves to exhibit a global minimum and tend to be convex. The first hypothesis about the curves with dominant concave bumps therefore was that they may correspond to pathologies where fluid intruded into the retinal layers and caused them to thicken.

Closer inspection of the corresponding scans and a comparison to the (blue) GT area curves, however, quickly disproved this hypothesis and revealed that the issue of a too thick segmented layer 1 appeared in all scans that exhibit a global maximum in the area curve or tend do be concave.

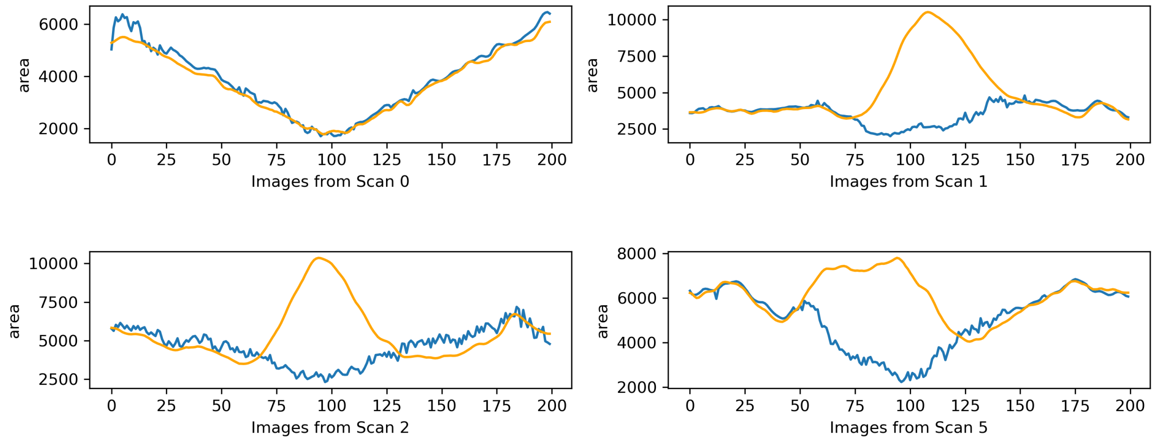

Figure 2: Examples of area curves resulting from ground truth (blue) and algorithmic (orange) segmentations. Scan 0 is an example of correct segmentation, the remaining three cases (scans 1, 2, 5) correspond to incorrect segmentations shown in figure 1.

3.1.1 Curve Embedding

To grasp the convex-vs-concave nature of the area curves and to embed them in a lower dimensional space we chose to approximate them by second order polynomials $a(x;\mathbf{w}) \approx w_0 + w_1 x + w_2 x^2$ and to represent them by the three regression coefficients w_i.

Following [3] the regression coefficients $\mathbf{w} = [w_0, w_1, w_2]^\top$ for each area curve are calculated by means of regularized least-squares, i.e, by solving $\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{a}$, where λ is the regularization term, $\mathbf{I}$ the 3×3 identity matrix, $\mathbf{\Phi}$ the 200×3 design matrix with rows $[1, x, x^2]$, and x indexes the slices $x \in \{0..199\}$.

The optimal regularization coefficient was determined close to zero $\lambda \approx 0$, which can be explained by the fact that fitting a low-grade polynomial to 200 values does not suffer from overfitting. This reduces the curve fitting to ordinary least squares, i.e., multiplication of the area curve vector by the psuedoinverse of the design matrix: $\mathbf{w} = (\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{a} = \mathbf{\Phi}^\dagger\mathbf{a}$.

3.1.2 Regression Coefficients

Regression coefficients corresponding to all 100 ground truth (blue) as well as algorithmic (orange) segmentations are scatter-plotted in the first row of figure 3. Its second row shows the three corresponding kernel density estimation (KDE) plots.

The two w_0 KDE plots indicate very similar distributions and therefore the w_0 coefficients do not seem to be discriminative.

The too-thick segmented layers are mapped to concave area curves. Therefore, the distribution of w_2 coefficients is of special interest, as they are responsible for the positive/negative curvature of the polynomials. Looking at the KDE plot of w_2 coefficients, there is a high peek from the ground truth coefficients between 0 and 0.5, showing that there are almost no negative w_2 coefficients. Therefore the assumption that ground truth curves tend to exhibit convexity (positive curvature) holds. In contrast, the orange KDE resulting from algorithm segmentations is more flat in the GT area and also exhibits a minor peek around -0.5. This indicates the presence of a cluster of negative w_2 values, which corresponds to concave area curves. This distribution can be confirmed looking at scatter plots including w_2. For example in the w_1–w_2 plot there is a (blue) cluster formed by ground truth coefficients while several negative w_2 algorithm coefficients are scattered outside of it.

Interestingly the w_1 coefficients exhibit a very similar behaviour to the w_2 coefficients: almost no positive w_1 GT coefficients and a tendency to bimodal distribution of the algorithm ones forming a small peek around value of 100.

The highly correlated coefficients w_1 and w_2 encourage for further dimensionality reduction. Indeed,

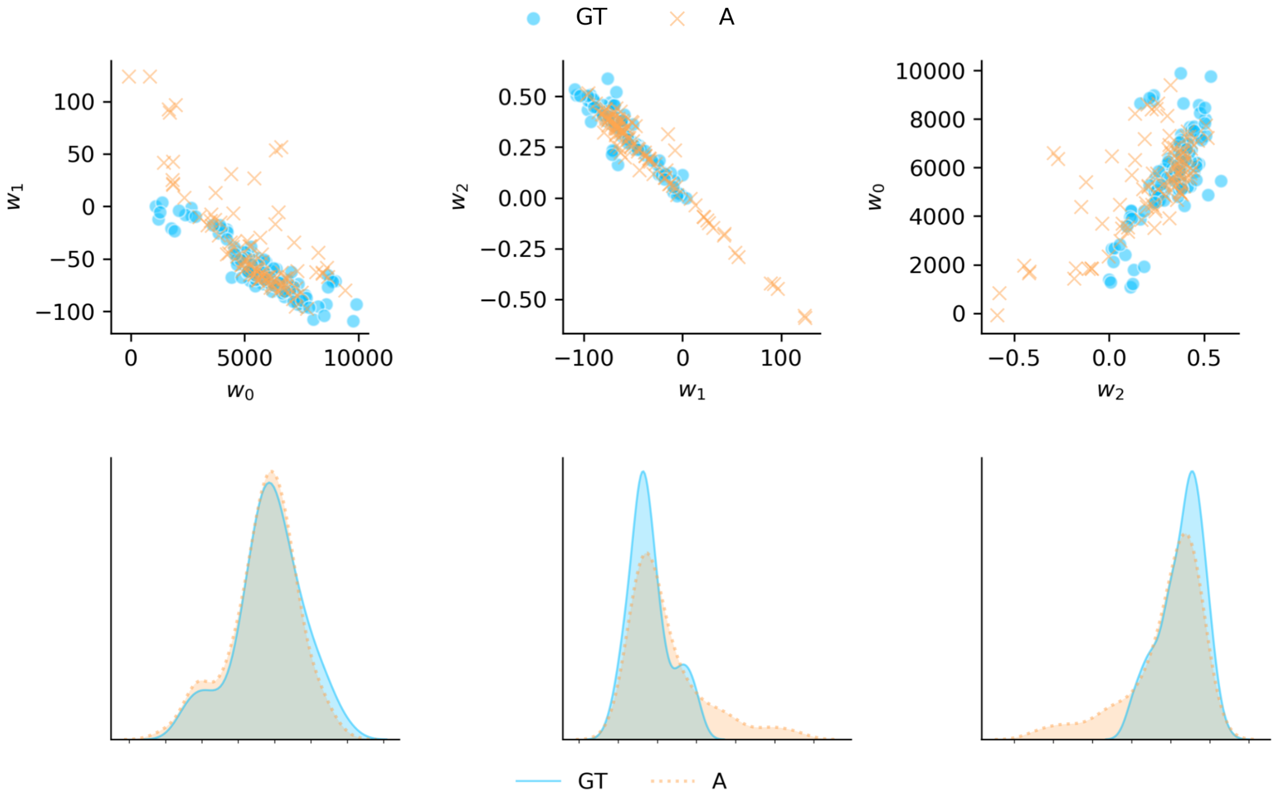

Figure 3: Scatter and kernel density estimation plots of the L1-area curve regression coefficients. The blue and orange dots/curves correspond to the respective coefficients of ground truth and algorithmic segmentation.

an interactive 3D scatter plot revealed the points close to a 2D linear manifold embedded in three dimensions. Projection of both ground-truth and algorithm-segmentation coefficients onto the first two PCA eigenvectors yields 2D scatter plot shown in the left part of figure 4.

In the following the problem of identifying incorrect segmentations is thus cast to outlier detection in a 2D feature space.

3.2. Outlier Detection Using Projected Regression Coefficients

Our approach to outlier detection is a semi-supervised one: we reuse the ground-truth coefficients to fit a model that represents the expected segmentation behavior. Subsequently the likelihood of an algorithmic segmentation to be generated by the learned model is tested.

While there is a broad spectrum of methods for outlier (novelty) detection, we show a digest of 5 algorithms resulting from our experiments and discuss their performance.

Feature Bagging (FB) [7] fits several base detectors on sub-samples of the dataset and use averaging to combat over-fitting. We used the LOF (see below) as the base detector.

Nearest Neighbors (KNN) [2] the distance of the sample to its most distant k-th neighbor is used as the outlier score. We set $k = 5$.

Local Outlier Factor (LOF) [4] Samples with much lower local density than their neighbors are declared as the outliers. The local density was estimated by 20 nearest neighbors.

Minimum Covariance Determinant (MCD) [12] fits the minimum covariance determinant model to the data. The outlier-ness of a sample is proportional to its Mahalanobis distance.

One-class SVM (OCSVM) introduced in [13] aims to find a smooth boundary modelling a user-specified probability that randomly drawn point will land outside.

4. Results and Discussion

To evaluate the outlier detectors quantitatively, notion of positives (incorrect segmentation) and negatives is necessary for the test data, i.e. for algorithmic segmentations. As this information was not present we chose to disambiguate the two classes by setting a Dice coefficient threshold. To figure out a sufficiently high Dice threshold we refer to the score-vs-dice scatter plot in the middle column of figure 4. Here the blue margin corresponds to the ground-truth region, proposed by the detector. The orange cluster within this margin then corresponds to true negatives (correct segmentations), and suggests the dice threshold of 0.87.

Figure 4 shows three plots for each of the five methods. In the following its columns are described in detail.

Left column: in addition to feature scatter plot the decision boundary and the scoring function of the respective detector are shown. In the following texts the orange test points falling outside the blue region will be referred to as the positives, points inside the blue region as the negatives.

Middle column shows scatter plots of Dice vs outlier scores. The horizontal line is the Dice threshold. The vertical lines are the thresholds of the scoring function proposed by the respective algorithms. The four quadrants correspond to TNs, FPs, FNs, and TPs, respectively. These four numbers are typeset in the top center of the plot and the recalls and precisions computed thereof are displayed in the titles.

Right column shows the ROC and Precision-Recall curves corresponding to the possible thresholds in the scoring function. The areas under these curves are abbreviated by auROC and auPR, respectively, and are displayed in the title.

The performance numbers are summarized in Table 1. In terms of precision, the areas under ROC and PR, the kNN seems to be the method of choice. However, the OCSVM wins in term of recall, because of its steep narrow margin which determines the outlier score. While LOF and FB are of lower recall, they are less over-fitted than earlier two, and we can observe an improvement when an ensemble of LOFs is aggregated into the FB. The MCD is easily interpretable but unfortunately not performing well.

Looking at the result of the well-fitting OCSVM, there are four FNs with a low Dice coefficient. Investigation on these revealed that such cases indeed might appear, because the area curves of the ground truth do not show a minimum around the middle of the slices, but have a nearly a rising shape. While the segmentation algorithm did not perform well on these scans, it still exhibits a convex fit to the area curve.

method	Rec.	Prec.	auROC	auPR
FB	0.73	0.95	0.96	0.93
KNN	0.77	1.00	0.97	0.94
LOF	0.65	0.89	0.96	0.91
MCD	0.54	0.88	0.95	0.87
OCSVM	0.85	0.92	0.88	0.89

Table 1: Summary of results

Analyzing the two false positives, one of them appeared close to the OCSVM boundary. The less over-fitted detectors (e.g the MCD), however, have classified this point correctly. The second false positive was a FP in all methods, except for the KNN. This could be because the ground truth data again shows an unusual shape: in contrast to the other ground truth shapes it starts with a high maximum, then falls down, but does not rise up again. There are few additional ground truth curves having this kind of shape which we consider unusual. When the segmentation algorithm yields such a shape, it is more likely to be a wrong segmentation.

Whether an ROC curve should be used to assess an outlier detector depends on the imbalance of the test set. In the current setting, the segmentation algorithm [5] does not seem to be mature enough as it produces around 25 percent of incorrect segmentation. As more reliable segmentation methods will be developed, the test set becomes increasingly more imbalanced and the validation by ROC and its area will have to be replaced by the precision-recall curves.

5. Conclusion and Future Work

We proposed a semi-supervised method to detect incorrectly segmented OCT retina scans: ground-truth segmentations are used, after feature extraction and projection to 2D, to train the decision boundary and the outlier scoring function. This function is subsequently used to flag the incorrectly segmented scans.

We evaluated a selection of five outlier detection methods and find the results to be a promising starting point to address the given problem.

While in this work the data-pipeline components are tailored to a specific segmentation algorithm and its pitfalls, we would like to sketch how the presented approach can be generalized. Firstly, higher-degree polynomials (i.e., more regression coefficients) could be used if it turns out that the segmentations can not be discriminated by the concave/convex shapes. Secondly, we concentrated only on description of layer 1, as imperfections in its segmentation propagated to subsequent layers. As the segmentation algorithms mature, descriptors of remaining layers could be incorporated. With an increased number of features, the ensemble-based detectors (FB in this work) may improve in their performance. Finally, after the segmentation algorithms become very advanced, it may turn out that the area-related descriptors loose their discriminative power and a need for completely new set descriptors may arise. In the proposed semi-supervised framework, the manually crafted features can be replaced by ones proposed by auto-encoders [1] or generative adversarial neural networks [8].

References

[1] C. C. Aggarwal. Outlier analysis. In *Data mining*, pages 75–79. Springer, 2015.

[2] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 15–27. Springer, 2002.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 3.1.4: Regularized least squares, pages 144–145. Springer, 2006.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[5] 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. *IEEE Transactions on Medical Imaging*, 28(9):1436–1447, Sep. 2009.

[6] V. Hodge. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 10 2004.

[7] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 157–166. ACM, 2005.

[8] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[9] P. J. Mekjavic, V. J. Balciüniene, L. Ceklic, J. Ernest, Z. Jamrichova, Z. Z. Nagy, I. Petkova, S. Teper, I. G. Topcic, and M. Veith. The burden of macular diseases in central and eastern Europe —- implications for healthcare systems. *Value in Health Regional Issues*, 19:1–6, 2019.

[10] T. Otani, S. Kishi, and Y. Maruyama. Patterns of diabetic macular edema with optical coherence tomography. *American Journal of Ophthalmology*, 127(6):688–693, 1999.

[11] Pro Visu Foundation. Fovea Centralis. https://www.provisu.ch/cgi/en/anatomical-structure.pl?en+alp+F+A09.371.729.522.436, 2018. [Online; accessed 13-October-2019].

[12] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[13] 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.

[14] W. D. Strain, X. Cos, and C. Prünte. Considerations for management of patients with diabetic macular edema: Optimizing treatment outcomes and minimizing safety concerns through interdisciplinary collaboration. *Diabetes Research and Clinical Practice*, 126:1–9, 2017.

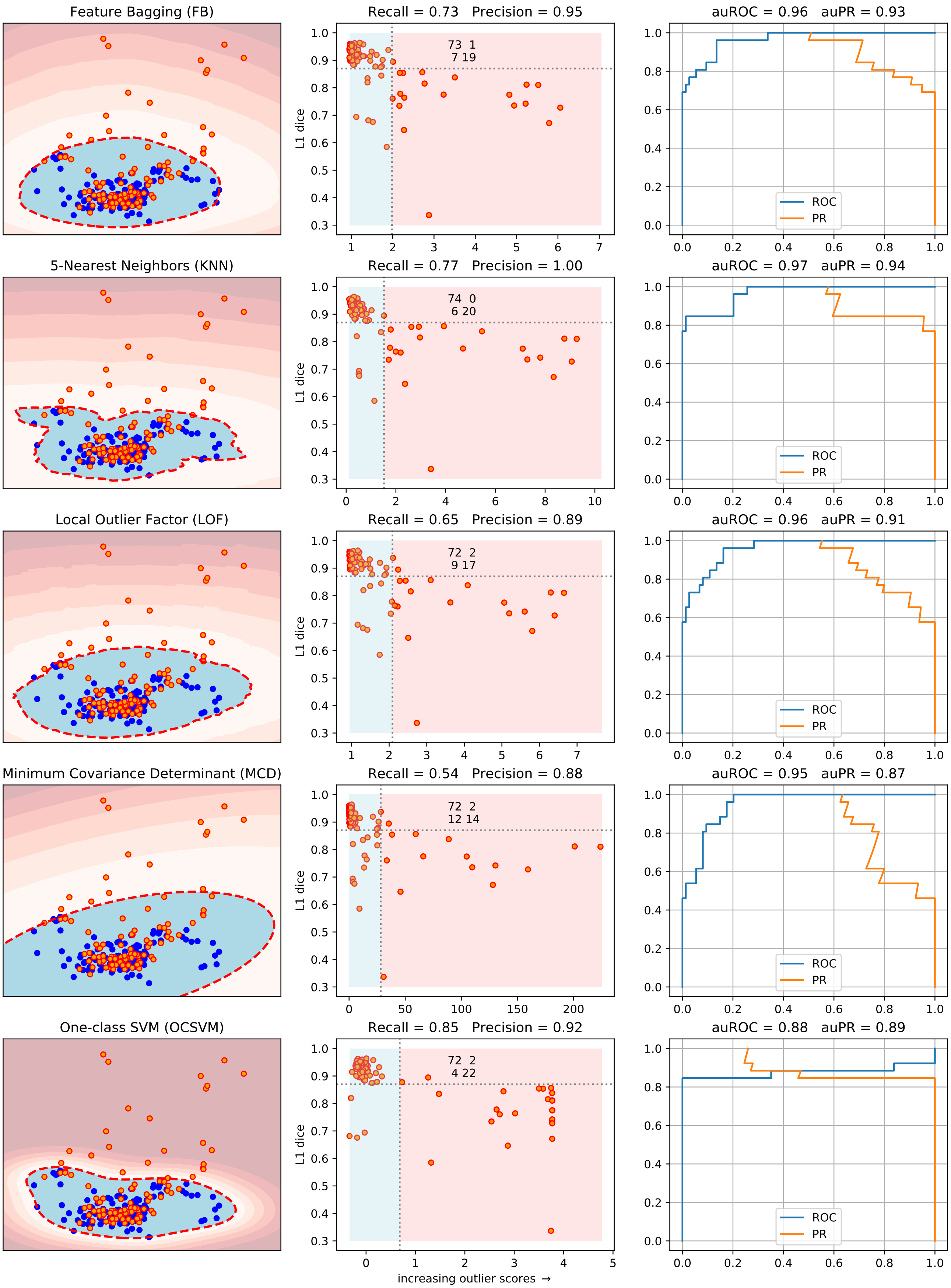

Figure 4: Selected outlier detectors and their performance on test set, i.e., the segmentation results of the algorithm.