Supplemental material for the paper *"Accurate and Efficient Linear Structure Segmentation by Leveraging Ad Hoc Features with Learned Filters"*

Roberto Rigamonti and Vincent Lepetit

CVLab, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{roberto.rigamonti,vincent.lepetit}@epfl.ch,
WWW home page: http://cvlab.epfl.ch

1 Additional results

We have evaluated the quality of the segmentations we obtained by using several analytic measures, to provide a fair comparison of the different approaches. In particular, we considered the following:

- Area Under Curve (AUC): represents the area subtended by the ROC curve. It assumes values in $[0, 1]$, the higher the better;
- F-measure;
- Variation of Information (VI) [3]: assumes values in $[\,0, \infty)$, the lower the better;
- Rand Index (RI) [5]: assumes values in $[0, 1]$, the higher the better.

Both the VI and the RI values are computed on the classification results thresholded at the value found using the F-measure. The results we have obtained are summarized in Tab. 1. The reported timings do not include the reading/writing time, both of which strongly impact on the performance of the pure learning-based approaches, that would be otherwise even more at disadvantage in the comparison. Also, no considerations have been done on the memory footprint, which is more than 10 times bigger for pure learning-based approaches. Handcrafted approaches have been computed at fewer scales when used to leverage learned features with respect to the case where they are used independently, because this experimentally resulted in no decrease in the performance but reduced the computational burden.

From the table it can be seen that, while for smaller images the methods that ground solely on learned filters are still competitive in terms of the computation time, their performance quickly degrades on bigger images such as those in the BF2D dataset. The classification time for approaches based on small filter banks might be bigger than the case when handcrafted features are added to the same descriptor, despite the increased descriptor size, because of the improved separability of the data in feature space. Also, ℓ_1-regularized logistic regression emerges as a valuable alternative, capable of providing good-quality segmentations at a tiny fraction of the competitors' computation time.

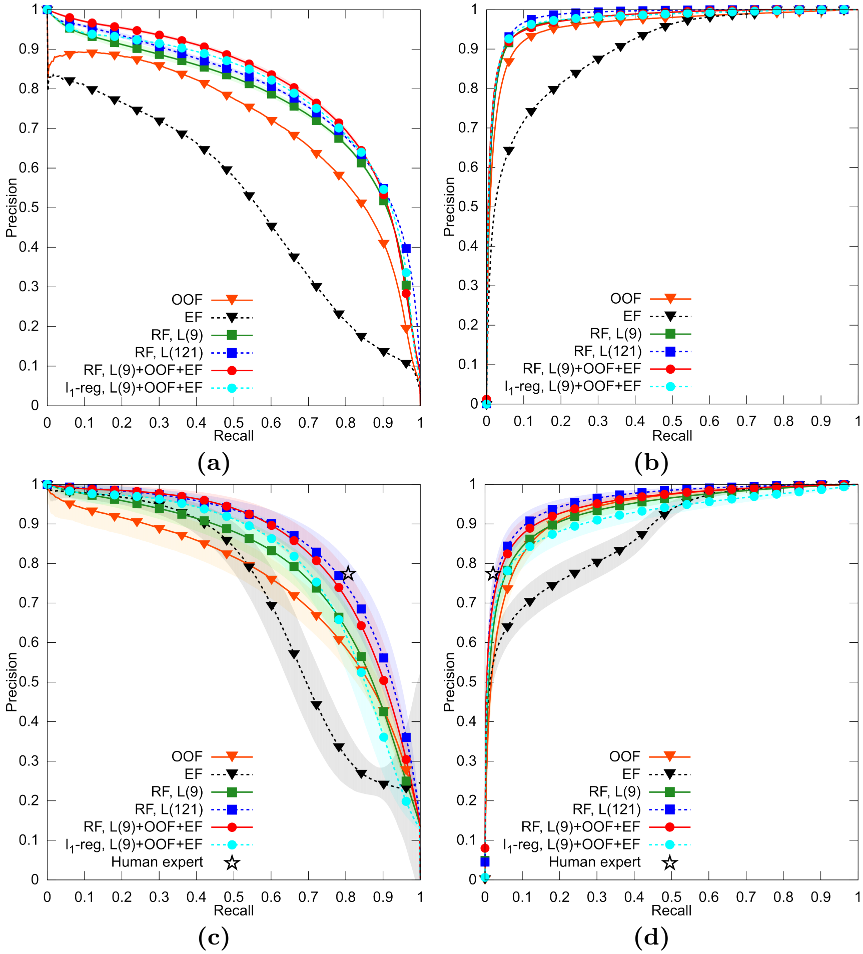
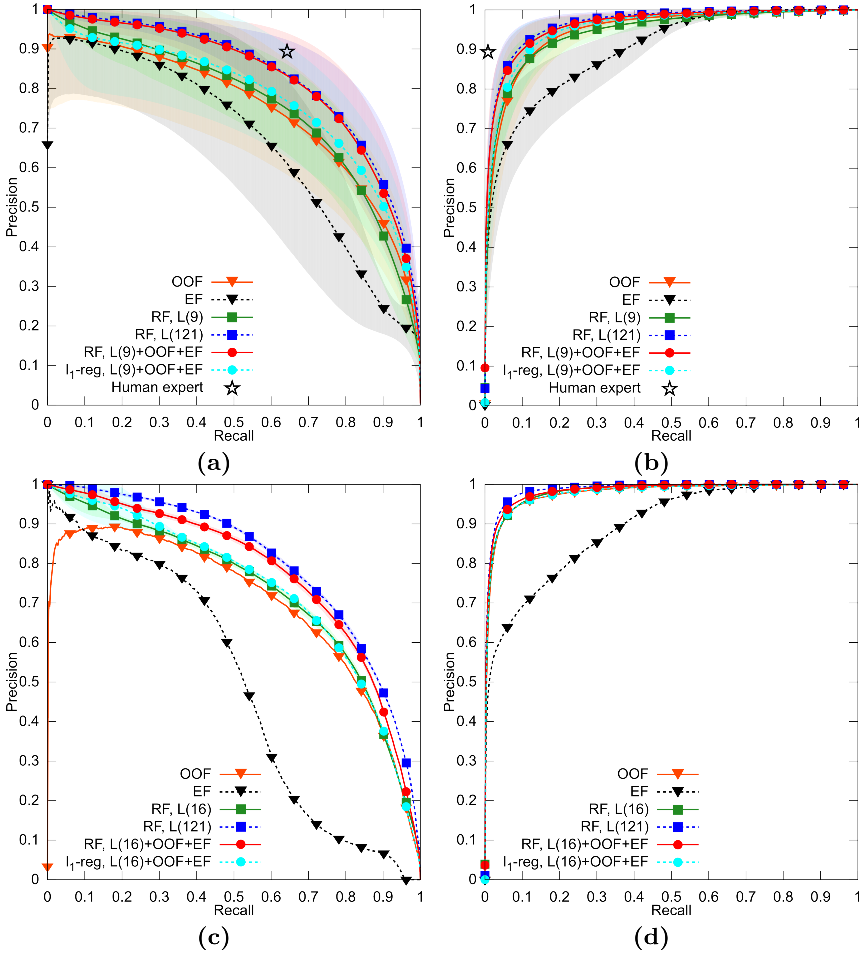


Fig. 1. PR (left) and ROC (right) curves computed on the BF2D **(a,b)** and DRIVE **(c,d)** datasets.

The Precision/Recall (PR) [1] and the corresponding Receiver Operating Characteristic (ROC) [2] curves averaged over 10 random trials for different approaches are given in Figs. 1, 2. By visual comparison of the obtained responses, Precision/Recall curves and the F-measure revealed to be the most reliable way to assess segmentation's quality, while the ROC curves revealed to be the most unreliable one, despite their popularity.

Figure 3 shows some filter banks learned on the DRIVE and VC6 datasets, while Fig. 4 shows two filter banks learned on the BF2D dataset along with an example of the extracted feature maps.

Figure 5 depicts some segmentation examples for the test image of the BF2D dataset, while the segmentations in Figs. 6, 7 are computed on the first test image



(a) **(b)** **(c)** **(d)**

Fig. 2. PR (left) and ROC (right) curves computed on the STARE **(a,b)** and VC6 **(c,d)** datasets. Please note that the filters used for the STARE images are those learned on the DRIVE dataset, and therefore they are not specifically tuned on STARE images' statistics.

of the DRIVE dataset. The colorized segmentations are obtained by thresholding the responses at a value giving a False Positive Rate (FPR) of 0.05. True positives are outlined in red, false positives in green, and false negatives in blue.

2 Framework's parametrization

The framework we adopted is composed by two main building blocks, namely the *filter learning* and the *pixel classification* parts.

For the former part no exact recipe exists in literature for devising a good parametrization. The parameters currently set in the example configuration files

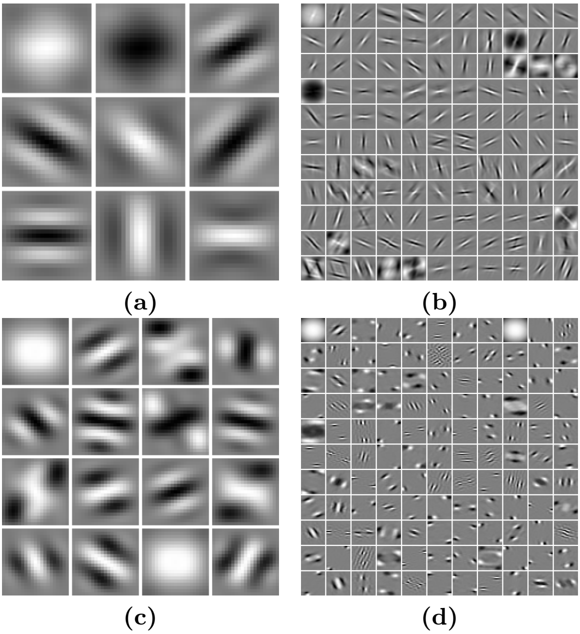


Fig. 3. Filter banks learned on the DRIVE **(a,b)** and VC6 dataset **(c,d)** and used in the feature extraction process.

available in the companion source code[1] should work flawlessly in most situations. In particular, we have used filters with size 21×21 pixels, and we observed that as few as 4 filters are enough to start leveraging handcrafted features. While the performances obviously increase with the number of filters, we found that a good speed/accuracy trade-off is represented by 9 filters (with the exception of the projections in the VC6 dataset, where 16 filters have to be considered in order to achieve reasonable results).

The performance measures for those approaches when considered alone were computed on responses derived from a dense discretization of the scale space. In particular, for the Oriented Flux Filter a good choice is to compute it at 14 scales in $[1, 7]$, and for the vesselness Enhancement Filter (EF) we have chosen 7 scales in $[1, 7]$, and we set $\gamma = 0.5, \beta = 1$. Nonetheless, when combined with learned filters, the extraction of handcrafted features showed to be very robust against the parameter choice. Experiments revealed that in this case just few scales are enough to boost the performances. We considered $\sigma = \{2, 3\}$ for the OOF method, and $\sigma = \{1, 3, 7\}$ for the Enhancement Filter.

The pixel classification code presents few tunable parameters that influence the behavior of the classifiers. We have empirically observed that 600 trees represent a reasonable number for a Random Trees classifier, and that a good value for the regularization parameter of the ℓ_1-regularized logistic regressor is $\lambda = 0.01$.

[1] Available at `http://cvlab.epfl.ch/~rigamont`

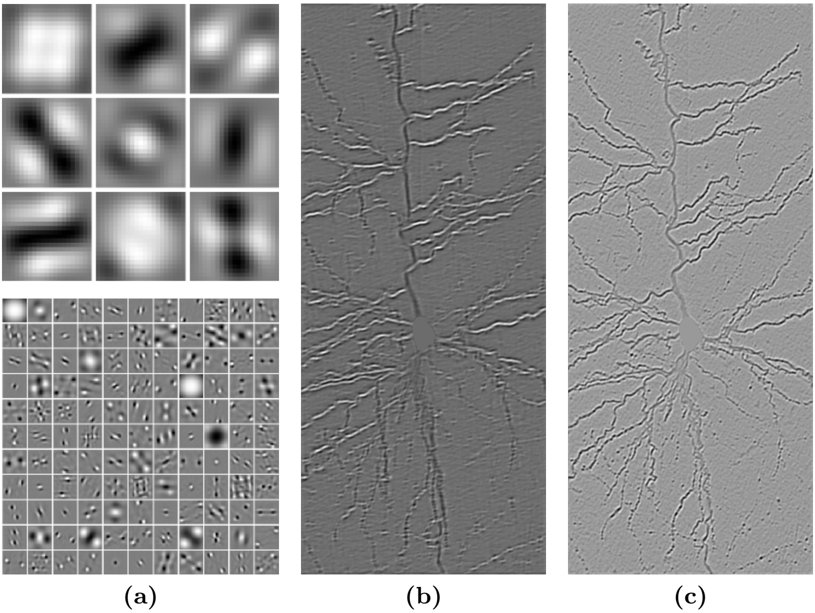


Fig. 4. Example filter banks learned on the BF2D images and corresponding extracted feature maps. **(a)** Filter banks learned on the BF2D images, with the small filter bank composed by 9 filters on the top, and the larger with 121 filters on the bottom. The size of the filters is the same (21 × 21 pixels). **(b)** Feature map extracted by the filter in row 3, column 1, in the filter bank of (a, top). **(c)** Feature map extracted by the filter in row 1, column 2, in the filter bank of (a, bottom).

The number of training samples can be relatively small, indicatively 10,000 positive and 10,000 negative samples are enough in most cases. We have observed that, for the DRIVE case, the performance of the learned filter bank with 121 filters improves with a very large number of training samples (around 100,000 positive and 100,000 negative samples), but this phenomenon has not been observed for the other datasets.

References

1. Davis, J., Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. In: ICML (2006)
2. Hanley, J.: Receiver Operating Characteristic (ROC) Methodology: The State of The Art. Crit. Rev. Diagn. (1989)
3. Meilă, M.: Comparing Clusterings - an Information Based Distance. J. Multivariate Anal. (2007)
4. Rigamonti, R., Türetken, E., González, G., Fua, P., Lepetit, V.: Filter Learning for Linear Structure Segmentation. Tech. rep., EPFL (2011)
5. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward Objective Evaluation of Image Segmentation Algorithms. PAMI (2007)

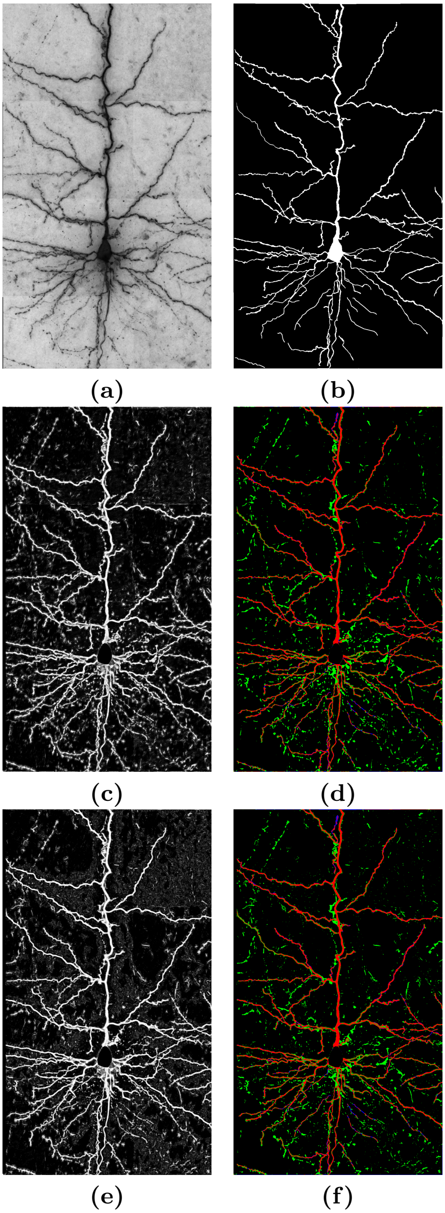


Fig. 5. Segmentation examples for the test image of the BF2D dataset. **(a)** Original image. **(b)** Ground truth. **(c)** Response of the Random Forests classifier when 121 learned filters are used. **(d)** Corresponding colorized segmentation. **(e)** Response of the Random Forests classifier when handcrafted features, leveraged by 9 learned filters, are used. **(f)** Corresponding colorized segmentation.

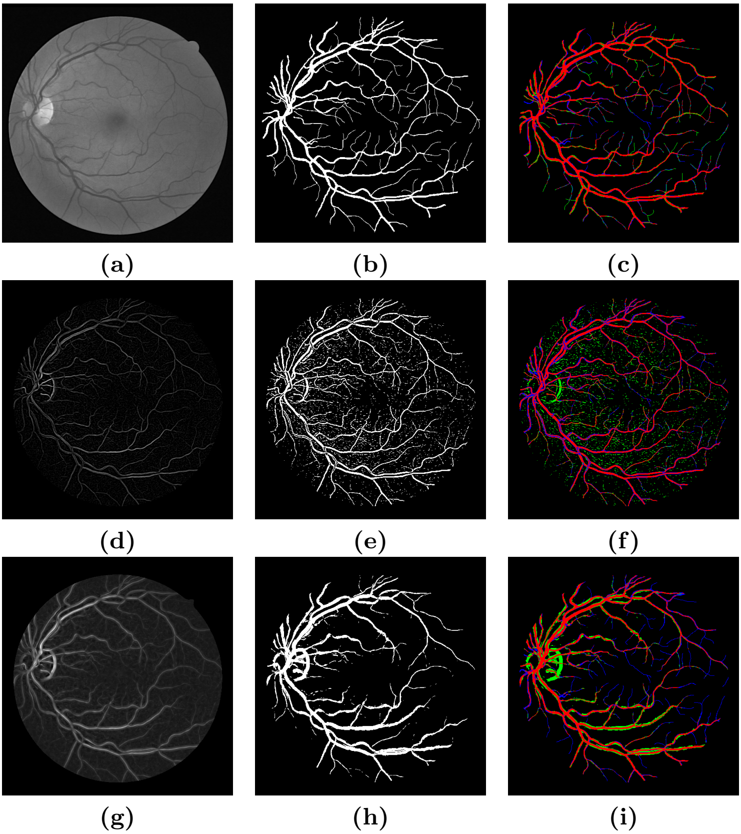


Fig. 6. Segmentation examples for the first test image of the DRIVE dataset. **(a)** Original image. **(b)** Ground truth. **(c)** Colorized segmentation based on the classification given by the second expert. **(d)** Response obtained by the Enhancement Filtering (EF) approach. **(e)** Binarized response of the EF method with FPR set at 0.05. **(f)** Colorized segmentation obtained by the EF method with FPR set at 0.05. **(g)** Response obtained by the Oriented Flux Filter (OOF) approach. **(h)** Binarized response of the OOF method with FPR set at 0.05. **(i)** Colorized segmentation obtained by the OOF method with FPR set at 0.05.

Table 1. Analytic measures of the quality of the pixel classification task for the different datasets. The values are averages over 10 random trials and over the whole dataset images. All the algorithms were artificially restricted to operate on a single core to have a fair comparison. The number after the ± sign is the standard deviation. For the filter banks subscripts indicate the cardinality, while for the handcrafted methods they indicate the number of scales at which they have been computed. For the learning-based approaches, a training set of 10,000 positive and 10,000 negative samples has been used. The time is expressed in seconds and per image, and accounts for the feature extraction and testing time only.

BF2D

Method	**AUC**	**F-measure**	**VI**	**RI**	**Time**
OOF$_{14}$	0.958	0.677	0.325	0.891	15.88
EF$_{7}$	0.899	0.537	0.334	0.894	8.27
RF,L$_{9}$	0.974 ±0.000	0.725 ±0.004	0.379 ±0.005	0.925 ±0.001	488.36
RF, L$_{121}$ [4]	0.981 ±0.000	0.736 ±0.003	0.336 ±0.004	0.937 ±0.001	855.07
RF, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.975 ±0.001	0.747 ±0.004	0.343 ±0.006	0.935 ±0.002	471.24
l1reg, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.974 ±0.000	0.740 ±0.001	0.351 ±0.004	0.933 ±0.001	16.69

DRIVE

Method	**AUC**	**F-measure**	**VI**	**RI**	**Time**
Ground truth		0.788	0.380	0.930	
OOF$_{14}$	0.933 ±0.012	0.695 ±0.028	0.569 ±0.034	0.770 ±0.018	5.70
EF$_{7}$	0.875 ±0.023	0.687 ±0.095	0.569 ±0.036	0.770 ±0.018	1.47
RF,L$_{9}$	0.934 ±0.013	0.731 ±0.026	0.690 ±0.062	0.860 ±0.017	110.79
RF, L$_{121}$ [4]	0.958 ±0.010	0.779 ±0.030	0.555 ±0.063	0.891 ±0.016	150.00
RF, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.949 ±0.013	0.766 ±0.030	0.589 ±0.068	0.883 ±0.017	112.08
l1reg, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.919 ±0.020	0.741 ±0.030	0.581 ±0.056	0.879 ±0.015	3.38

STARE

Method	**AUC**	**F-measure**	**VI**	**RI**	**Time**
Ground truth		0.740	0.424	0.909	
OOF$_{14}$	0.946 ±0.027	0.691 ±0.086	0.488 ±0.064	0.815 ±0.031	6.12
EF$_{7}$	0.899 ±0.039	0.632 ±0.103	0.486 ±0.062	0.814 ±0.032	1.87
RF,L$_{9}$	0.943 ±0.020	0.706 ±0.067	0.669 ±0.144	0.847 ±0.058	172.78
RF, L$_{121}$ [4]	0.966 ±0.015	0.761 ±0.059	0.534 ±0.134	0.889 ±0.042	189.72
RF, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.963 ±0.017	0.757 ±0.065	0.602 ±0.136	0.886 ±0.045	131.131
l1reg, L$_{9}$+OOF$_{2}$+EF$_{3}$	0.954 ±0.022	0.716 ±0.083	0.581 ±0.127	0.878 ±0.039	3.82

VC6

Method	**AUC**	**F-measure**	**VI**	**RI**	**Time**
OOF$_{14}$	0.977	0.671	0.192	0.945	5.35
EF$_{7}$	0.894	0.538	0.195	0.949	1.44
RF,L$_{9}$	0.981 ±0.000	0.686 ±0.007	0.263 ±0.013	0.947 ±0.004	82.66
RF, L$_{121}$ [4]	0.987 ±0.000	0.727 ±0.003	0.213 ±0.011	0.961 ±0.003	114.54
RF, L$_{16}$+OOF$_{2}$+EF$_{3}$	0.983 ±0.000	0.716 ±0.007	0.239 ±0.009	0.954 ±0.003	85.56
l1reg, L$_{16}$+OOF$_{2}$+EF$_{3}$	0.978 ±0.000	0.689 ±0.002	0.258 ±0.005	0.949 ±0.002	5.12

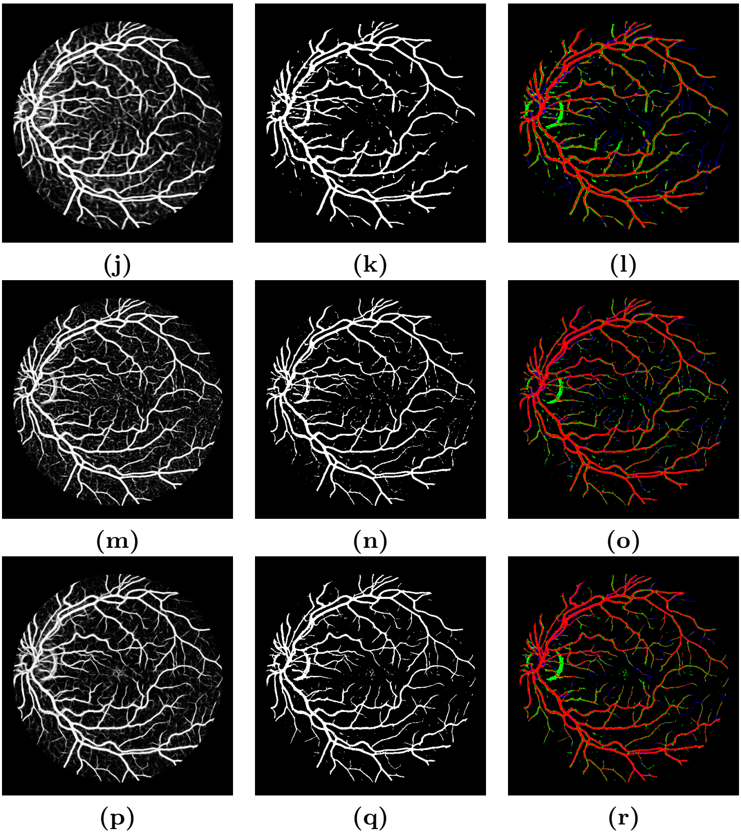


Fig. 7. Segmentation examples for the first test image of the DRIVE dataset (continued). **(j)** Response obtained by 9 learned filters. **(k)** Binarized response of 9 learned filters with FPR set at 0.05. **(l)** Colorized segmentation obtained by 9 learned filters with FPR set at 0.05. **(m)** Response obtained by handcrafted features leveraged by 9 learned filters. **(n)** Binarized response of handcrafted features leveraged by 9 learned filters, with FPR set at 0.05. **(o)** Colorized segmentation obtained by handcrafted features leveraged by 9 learned filters, with FPR set at 0.05. **(p)** Response obtained by the 121 learned filters. **(q)** Binarized response of 121 learned filters with FPR set at 0.05. **(r)** Colorized segmentation obtained by 121 learned filters with FPR set at 0.05.