

IMAGE SEGMENTATION USING SPARSE LOGISTIC REGRESSION WITH SPATIAL PRIOR

Pekka Ruusuvuori, Tapio Manninen, Heikki Huttunen

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

ABSTRACT

We propose a supervised learning based method for image segmentation. The method is based on feeding artificial features to a framework of logistic regression classifier with ℓ_1 norm regularization and Markov Random Field prior, which has been studied, e.g., in hyperspectral image segmentation. The novelty of our approach stems from the use of a generic artificial feature set and the embedded feature selection property of the sparse logistic regression framework, which avoids application specific feature engineering. The proposed method generates a large set of artificial features and passes them to a ℓ_1 regularized logistic regression classifier. Finally, a spatial prior imposes additional homogeneity to the classification result. We study the performance of the proposed method for two application cases, and show that the segmentation results are accurate even with simple models with high degree of sparsity.

Index Terms— Logistic regression, regularization, segmentation, classification, object detection, graph cut

1. INTRODUCTION

The approach for solving segmentation and detection problems in image analysis is typically very application oriented, and the designer is an expert with image analysis techniques. However, often the end user of the image analysis tool is not experienced in the field, and manual tuning becomes difficult. In such cases, supervised image segmentation may offer an alternative, where the non-expert end user may point a number of foreground and background areas, and the software automatically learns to separate them. However, in this framework, the use of a large collection of image analysis tools and selection among them may be difficult.

In this paper, we propose to use a linear classifier with sparse coefficient vector together with a Markov Random Field (MRF) spatial prior for the segmentation task. The inputs to the classifier are constructed by applying a collection of filters with various parameters (e.g., filtering window size). This way the training stage has a large amount of possible input filters, from which the sparse classifier can choose the most relevant ones for the particular classification task.

This manuscript has been supported by Academy of Finland project #140052 (PR).

However, the actual segmentation does not become overly complex despite a possibly large collection of filters, because those features discarded in the training stage are not necessary to be calculated in the actual segmentation. Importantly, the feature set is designed to be generic, enabling the use of the framework without building case-specific feature sets.

Our approach is based on ℓ_1 -penalized logistic regression classifier [1], which combines feature selection and classification into a single penalized maximum likelihood optimization problem, thus, embedding the feature selection problem into the classifier design. In the case of segmentation, the input features are generated by applying various spatial filtering operations to the images, and the output is an estimate of the probability of a pixel to belong to foreground. However, the estimates are pixel-wise and assume statistical independence between neighboring pixels, but the assumption does not hold in real-world images. Thus, we will take advantage of the accurate estimates of class probabilities by integrating the contextual information from neighboring pixels through a spatial prior: the Markov Random Field model.

There exists a lot of literature on design and use of image filters as feature generators for segmentation [2] including the design of sparse features [3]. On the other hand, the design of sparse classifiers is a widely studied area [4], which has also been applied in pixel classification with hyperspectral images [5] and for backscatter image segmentation [6]. Of the two latter studies, the first inputs multichannel images acquired at different wavelengths directly with no filtering, while the second uses a fixed collection of filters, but selects the most relevant ones with simple forward selection. Moreover, the latter paper does not estimate the class probabilities, which have a significant role in our framework. Thus, to the best of our knowledge, the two domains of feature generation with a large collection of filters and sparse classifier design have not been integrated together.

The non-sparse approaches to supervised classification [7] typically have a rigid classification structure with a small number of fixed features generated by a filter bank, and are thus difficult to modify by a non-expert to incorporate other features than those specified in the implementation. The famous Viola-Jones framework [8] is related to our approach, but the features are limited to have a fixed structure: only sum-of-blocks-like features are the only ones allowed.

2. SUPERVISED SEGMENTATION FRAMEWORK

The proposed segmentation framework consists of three stages: first a large pool of features is generated by applying a large set of manually defined filters with a range of parameters (see e.g., Table 1). At the training stage, we design a sparse logistic regression classifier to determine the class of each pixel. The design is based on training data collected by the user by pointing at a number of foreground and background areas. However, in the applications considered in the experiments, the foreground objects are point-like, and thus the background samples can be collected automatically from areas far enough from the foreground areas.

In the actual classification stage, the selected features are calculated, and the classifier predicts the pixelwise class membership probabilities. These are integrated with the contextual information through the MRF model.

2.1. Feature generation

The original data is in the form of 2D images. However, the image intensities alone do not necessarily form informative enough feature set for supervised object detection. Thus, we apply a number of filtering and transform operations for the image intensities in order to build a higher dimensional feature set. A large set of features is particularly desired in our case since the logistic regression framework, described in Section 2.2, is able to efficiently handle cases with high dimensional feature space. Moreover, by generating a large feature set instead of strictly optimized feature(s) for a specific purpose and leaving the task of feature selection to the classification stage, we aim at a segmentation framework which has the potential to generalize better to different applications.

Our feature set is mainly built by spatial filtering operations with multiple kernel sizes. Although the set of filters can be selected arbitrarily, e.g., taking into account problem specific knowledge, we will use the following set of filters as feature generators in the experiments of this paper. The operations include low-pass filtering, unsharp masking, morphological and edge enhancing filtering, all of which incorporate spatial information using local windows with varying kernel sizes. In addition, we use the wavelet decomposition [9] where the signal is represented as a multiscale decomposition of à trous wavelet coefficients, and the local binary patterns [10] for producing features. The features and parameter ranges are listed in Table 1, and the parameter ranges are given using the notation [min:step:max].

2.2. Regularized Logistic Regression Classifier

The feature generation produces a large set of redundant and highly correlated features. The question is now, how to find a good subset of complementary features enabling an accurate classification result. This problem of *feature selection*

Table 1. Features and parameter ranges used in the experiments. In total, 106 features were used in this study but the framework allows to extend the feature list without changes.

Feature	Parameter	Values
Gaussian lpf	kernel width σ	3:5:83
Unsharp masking	kernel width σ	3:5:83
Morphological top-hat	kernel size	3:5:33
Morphological bottom-hat	kernel size	3:5:83
Masking with h-maxima transform	h	3:5:83
Local binary patterns		
Edge enhancement	kernel size	3:5:83
Wavelet decomposition	depth layer	1,2,3

has been widely studied, and one of the most interesting directions is the use of regularization for embedded feature selection. Probably the most famous method is the LASSO [1], which uses the ℓ_1 penalty to regularize the least squares solution of a regression problem. The LASSO has later been extended to solve classification problems through the logistic link function, which results in the logistic regression classifier [4], defined as follows.

Given a p -dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^p$ corresponding to the i^{th} pixel in the analyzed image, logistic regression models the probability $p(c_i|\mathbf{x}_i)$ of the pixel i belonging to the foreground as¹

$$p(c_i|\mathbf{x}_i) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)}. \quad (1)$$

The model parameters β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ are estimated by maximizing the ℓ_1 -penalized log-likelihood

$$\sum_{\mathbf{x}_i \in F} \log p(c_i|\mathbf{x}_i) + \sum_{\mathbf{x}_i \in B} \log(1 - p(c_i|\mathbf{x}_i)) - \lambda \|\boldsymbol{\beta}\|_1, \quad (2)$$

where F and B are the training sets of foreground and background pixels, respectively [11]. The parameter $\lambda > 0$ controls the strength of the regularization and thus the sparsity of the result, and is selected by cross-validation. The model parameters β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ can be efficiently estimated by a coordinate descent algorithm [4].

2.3. MAP Segmentation using Graph Cuts

As described in Equation (1), the model estimates the probability $p(c_i|\mathbf{x}_i)$ of the i^{th} pixel belonging to the foreground given the feature vector \mathbf{x}_i . It is tempting to use the probability information for post-filtering and increasing the coherence between neighboring pixels. This can be done in an optimal manner through the MRF spatial prior.

Markov Random Fields were first proposed for vision applications by Geman and Geman [12]. The MRF assumption

¹We denote both the discrete probability $\Pr(C = c_i | \mathbf{X} = \mathbf{x}_i)$ and the continuous pdf $p(\mathbf{x} = \mathbf{x}_i | C = c_i)$ with the short notations $p(c_i|\mathbf{x}_i)$ and $p(\mathbf{x}_i|c_i)$, unless the context is ambiguous.

states that each pixel class c_i (i.e., foreground / background) depends only on its neighbors as defined by the neighborhood \mathcal{N}_i . The idea of using graphs for solving the maximum a posteriori configuration with MRF prior was originally discovered by Greig *et al.* [13], and a fast implementation was proposed by Boykov and Kolmogorov [14]. This so called graph cut method solves the equivalent problem of splitting a graph into two disconnected parts, such that the foreground and background nodes are in different partitions, and it is applicable for a large set of MRF's. An MRF that can be solved by finding a graph cut has the following form for the probability of a labeling \mathbf{c} of the set of all pixels \mathcal{P} :

$$p(\mathbf{c}) \propto \exp \left(- \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}_i} V_{i,j}(c_i, c_j) \right),$$

where the *clique potential* $V_{i,j}(c_i, c_j)$ essentially determines the cost of giving different labels to neighboring pixels. Our definition of clique potential is equivalent to that used by [5] for hyperspectral image segmentation, which gives a constant penalty $\gamma > 0$ for each pair of neighbors with different labels.

The smoothing effect of the MRF prior can be seen from the derived formulation for the prior $p(\mathbf{c})$, [5]:

$$p(\mathbf{c}) \propto \exp \left(\gamma \sum_{\{i,j\} \in \mathcal{C}} \delta(c_i - c_j) \right), \quad (3)$$

where $\delta(\cdot)$ is the unit impulse function. In the above formula, equal labels c_i and c_j for neighboring pixels i, j clearly increase the value of the prior, thus favoring segmentations with a large number of cliques \mathcal{C} having the same class label.

The prior of Eq. (3) is integrated with the pixelwise logistic regression classifier as proposed in [5]. The problem is that we would like to find the pixel labeling $\hat{\mathbf{c}}$ maximizing the posterior probability with the MRF prior, i.e., $\hat{\mathbf{c}} = \arg \max_{\mathbf{c}} p(\mathbf{x}|\mathbf{c})p(\mathbf{c})$, with $p(\mathbf{x}|\mathbf{c})$ the likelihood of the data with labels \mathbf{c} and $p(\mathbf{c})$ as defined in Eq. (3). However, logistic regression estimates posterior probabilities $p(c_i|\mathbf{x}_i)$ for pixel i instead of the likelihood $p(\mathbf{x}_i|c_i)$. This is resolved by using Bayes formula in the unusual direction [5]: $p(\mathbf{x}_i|c_i) = p(c_i|\mathbf{x}_i)p(\mathbf{x}_i)/p(c_i)$, or if further assuming conditional independence and discarding the constant term $p(\mathbf{x}_i)$:

$$p(\mathbf{x}|\mathbf{c}) \propto \prod_i \frac{p(c_i|\mathbf{x}_i)}{p(c_i)}.$$

If we further assume equal class probabilities, the denominator can also be omitted. This way we will end up with the definition of the MAP segmentation:

$$\begin{aligned} \hat{\mathbf{c}} &= \arg \max_{\mathbf{c}} p(\mathbf{x}|\mathbf{c})p(\mathbf{c}) \\ &= \arg \max_{\mathbf{c}} \left(\sum_i -\log p(c_i|\mathbf{x}_i) + \beta \sum_{\{i,j\} \in \mathcal{C}} \delta(c_i - c_j) \right), \end{aligned}$$

whose exact minimization can be done using the graph cut.

3. EXPERIMENTAL RESULTS

We study the performance of the proposed method for two cases: segmentation of subcellular objects [15] and segmentation of connection pads in printed electronics [16]. These cases were selected because they represent challenging segmentation cases where the supervised framework, enabling automated analysis through incorporating expert knowledge in the training phase, would be beneficial. In the first case, we show how the sparse logistic regression reduces the number of features in the model, and the second case highlights the benefit of generating MAP segmentation result from the regression probabilities by using spatial priors.

3.1. Subcellular spot detection from simulated images

As the first dataset we use simulated images of cell populations. The main benefit of using simulated images stems from the availability of ground truth information of the number and locations of subcellular objects. The simulated image set, publicly available at <http://www.cs.tut.fi/~sgn/sgb/simcep/benchmark> contains 20 images with a fixed number of cells and subcellular objects per cell. A part of a simulated image is shown in Fig. 1(a), the probability values given by the logreg model as a heatmap in Fig. 1(b), and the result after classification in Fig. 1(c).

Coordinates for positive training samples were given by the user and negative samples were randomly picked from the input image with minimum distance to any positive sample at least 10 pixels. Here, 70 positive samples were given and the number of negative samples was set to be $10 \times$ that of positive samples. Using these 770 training samples each having 106 features, a sparse solution with only five nonzero feature weights was obtained. The simplicity of the model can be seen from the actual formula for the probability of the foreground class for the pixel with coordinates (i, j) :

$$p(c_{i,j}|\mathbf{x}_{i,j}) = \frac{1}{1 + \epsilon(i, j)}, \quad (4)$$

with

$$\begin{aligned} \epsilon(i, j) &= \exp(-5.23 + 0.04 H_{63}(i, j) + 0.003 H_{68}(i, j) \\ &\quad + 0.0007 E_8(i, j) + 0.011 T_8(i, j) + 0.30 W_3(i, j)), \end{aligned}$$

where H denotes the result of h-maxima masking, T top-hat filtering, E edge enhancement and W wavelet decomposition at pixel (i, j) and the subscripts denote the parameter values (see Table 1).

Since the ground truth information is available for simulated images, we can use direct measures of detection accuracy. Thus, we determine the number of true positive (tp), false positive (fp), and false negative (fn) detections, as well as the precision ($p = tp/(tp + fp)$) and recall ($r = tp/(tp + fn)$). Finally, using these measures the F-score is defined as

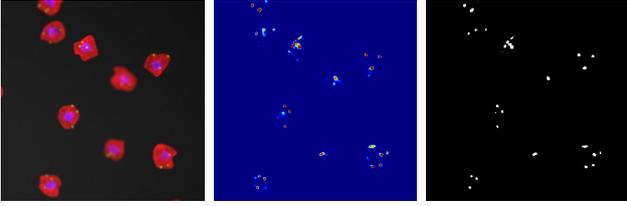


Fig. 1. (a) Close-up of a simulated image. (b) Probability values by LR. (c) Classification result.

Table 2. Subcellular object detection results for simulated images. The results are averaged over the set of test images.

tp	fp	fn	p	r	F-score
121.11	17.26	20.00	0.876	0.858	0.867

$F_{\text{score}} = 2pr/(p+r)$. The results for simulated images are given in Tab. 2. The values are averaged over a set of 19 test images - one of the 20 images is used for training and in order to keep the test data independent it is excluded from the test set. For comparison, the same image set was used for evaluating unsupervised spot detection algorithms in [15], and based on the results reported for simulated images, the supervised framework outperforms the methods used in the earlier study.

3.2. Detection of Connection Pads in Printed Electronics

As the second example case, we consider segmentation of the connection pads of integrated circuits (IC) on an electronics module. Such an application is used, e.g., as a preprocessing step in printed electronics, where the components need to be detected from the image in order to be able to print the connection layer on the right location on top of the module. In such an application, the connection pads are considered the only dimensionally accurate features in the images, which makes segmentation a proper approach for this problem [16].

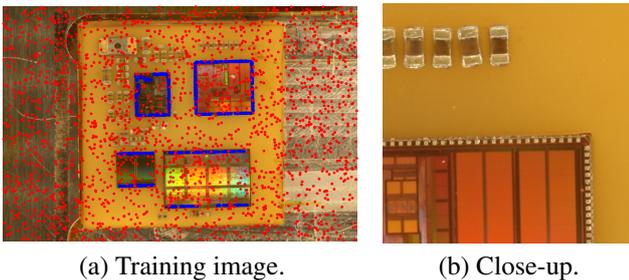


Fig. 2. An electronics module with 4 ICs used in training a classifier for IC connection pad segmentation. The positive and negative training pixels are shown with blue and red dots, respectively. A closeup of a corner of one of the ICs shows connection pads located on the sides of the IC.

Table 3. Connection pad detection results. The results are averaged over the 2 CV folds.

Method	tp	fp	fn	p	r	F-score
LR-MRF	24484	87133	460	0.219	0.982	0.359
LR	24683	131573	261	0.158	0.990	0.272
SVM	24498	154422	447	0.137	0.982	0.240
10-NN	24566	156066	378	0.136	0.985	0.239

We use 12 Mpix color images taken from 4 IC modules as shown in Figure 2 (a). There are a total of 409 connection pads located at the sides of each IC. A ground truth segmentation for the whole image has been acquired by manually fitting the connection pad layout as given in the design data on top of the image and annotating each pixel at most $30 \mu\text{m}$ away from any connection pad as positive and the rest of the pixels as negative samples. The amount of training data per image has then been reduced by randomly picking 1000 positive and 2000 negative training samples as indicated by the blue and red dots, respectively. The actual connection pads can be seen in a closeup of one of the ICs in Figure 2 (b).

In this example, 4 different supervised methods are tested in the segmentation of the connection pads. In addition to the proposed regularized logistic regression combined with the graph cut post processing, also a regularized logistic regression classifier without graph cut as well as support vector machine (SVM) and a 10 nearest neighbors (10-NN) classifiers are tested. The features used are those listed in Table 1. In the case of SVM and 10-NN, an additional sequential forward selection step is used to reduce the amount of intentionally redundant features. In SVM, we use linear kernel in order to avoid optimizing kernel parameters. The resulting model is exactly the same as in logistic regression. In 10-NN, ℓ_1 norm is used as the distance metric and the number of nearest neighbors is selected through experimenting. In all the methods, the features are first normalized to unit variance.

The performance of the different classifiers was assessed by using a simple 2-fold cross-validation (CV) such that the classifiers were trained with a single module image similar to that shown in Figure 2 (a) and tested with another one. The results are given in Table 3. There are plenty of false positives in each case, which is due the high amount of small bright spots other than connection pads in the images. Logistic regression seems to give slightly less false positives than SVM and 10-NN resulting in better precision and F-score. Using MRF together with the class probabilities estimated by the logistic regression remarkably reduces the number of false positives giving it by far the most accurate result. Figure 3 shows the part of the segmentation results that corresponds to Figure 2 (b).

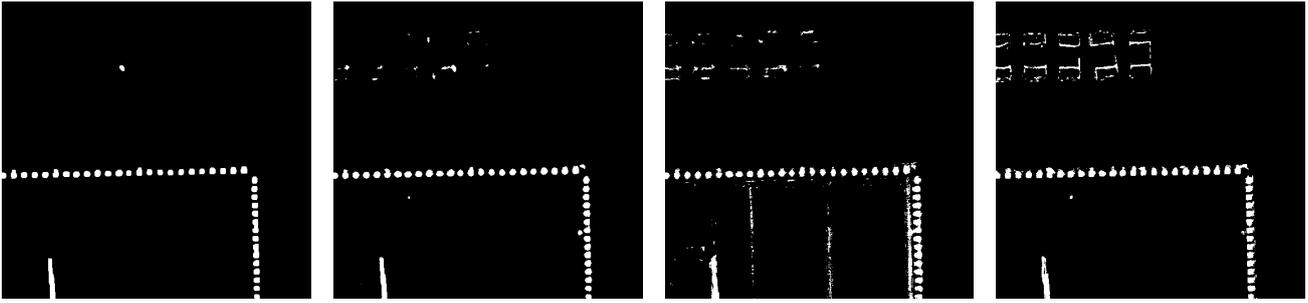


Fig. 3. The segmentation result corresponding to the closeup seen in Figure 2 (b) as given by different classification methods. (a) LR-MRF, (b) LR, (c) SVM, (d) 10-NN.

4. CONCLUSIONS

In this paper a framework for segmenting images using sparse logistic regression based classification was proposed. The framework calculates a large set of features with multiple support areas from the images and feeds them into the regularized logistic regression process, which weights the features based on their usefulness for the classification task, giving a sparse solution where non-informative features are excluded from the model. The class probabilities given by the logistic regression model together with spatial priors may be further processed by using graph cuts for generating a MAP segmentation result. In this case we have used detection of subcellular objects and connector pads as application cases to demonstrate the framework. The results for both simulated and real images confirm that the logistic regression based classification is a powerful tool for segmentation of small low contrast spots from images. Further studies will be taken in order to explore the full potential of the proposed framework.

5. REFERENCES

- [1] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
- [2] C. Sagiv, N.A. Sochen, and Y.Y. Zeevi, "Integrated active contours for texture segmentation," *Image Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1633–1646, June 2006.
- [3] R. Rigamonti, M.A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 1545–1552.
- [4] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
- [5] J.S. Borges, J.M. Bioucas-Dias, and A.R.S. Marcal, "Bayesian hyperspectral image segmentation with discriminative class learning," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 6, pp. 2151–2164, June 2011.
- [6] P. Paclik, R.P.W. Duin, G.M.P. van Kempen, and R. Kohlus, "Supervised segmentation of textures in backscatter images," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, vol. 2, pp. 490–493 vol.2.
- [7] I. Smal, M. Loog, W. Niessen, and E. Meijering, "Quantitative comparison of spot detection methods in fluorescence microscopy," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 282–301, 2010.
- [8] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.
- [9] J-C. Olivo-Marin, "Extraction of spots in biological images using multiscale products," *Pattern Recogn.*, vol. 35, pp. 1989–1996, 2002.
- [10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [11] B. Krishnapuram, L. Carin, M.A.T. Figueiredo, and A.J. Hartemink, "Sparse multinomial logistic regression: fast algorithms and generalization bounds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 957–968, June 2005.
- [12] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
- [13] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society.*, vol. 51, no. 2, pp. 271–279, 1989.
- [14] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, Sept. 2004.
- [15] P. Ruusuvuori, T. Äijö, S. Chowdhury, C. Garmendia-Torres, J. Selinummi, M. Birbaumer, A.M. Dudley, L. Pelkmans, and O. Yli-Harja, "Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images.," *BMC Bioinformatics*, vol. 11, pp. 248, 2010.
- [16] T. Manninen, V. Pekkanen, K. Rutanen, P. Ruusuvuori, R. Rönkkä, and H. Huttunen, "Alignment of individually adapted print patterns for ink jet printed electronics," *J. Imag. Sci. and Tech.*, vol. 54, no. 5, pp. 050306, 2010.