Document Segmentation using Relative Location Features

Francisco Cruz Fernández, Oriol Ramos Terrades
Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain
{*fcruz, oriolrt*}*@cvc.uab.es*

Abstract

In this paper we evaluate the use of Relative Location Features (RLF) on a historical document segmentation task, and compare the quality of the results obtained on structured and unstructured documents using RLF and not using them. We prove that using these features improve the final segmentation on documents with a strong structure, while their application on unstructured documents does not show significant improvement. Although this paper is not focused on segmenting unstructured documents, results obtained on a benchmark dataset are equal or even overcome previous results of similar works.

1. Introduction

In the last years, the interest on retrieving information registered from huge data collections, either typed or handwritten, has risen as digitalization technology has improved. The great variety of documents makes that the method that works properly on a specific type of documents offer worst results for other types. In addition, the degradation of ancient documents makes the task even more complicated. All in all, document segmentation is an active and challenging research field within the Document Image Analysis community.

Over the years this problem has been addressed from many different perspectives. The success or failure of a segmentation task largely depends on the type of features used. Some methods are based on connected components [1], while others are based on multi-resolution features [9]. Texture features, like Gabor is, have been proved to obtain good results working in both spatial and frequency domains [7, 8].

In [7], a method based on the Gabor transform applied to text region segmentation in different types of documents is proposed. Also, several page segmentation contests have been organised [1, 2]. There, a number of page segmentation methods have been evaluated. One of the methods achieving good results in one of these competitions was the Océ method. It performs text segmentation by classifying connected components using decision trees based on morphological features. Recent works have proved that the inclusion of contextual information, like relationships between different entities of documents, significantly helps to improve the segmentation results [4, 10]. In [9] the authors propose a segmentation method for documents and scene images based in matched wavelets specially adapted for text and no text region segmentation using Conditional Random Fields (CRF). However, the use of the CRF in these works is limited to model pairwise relationships between pixels, without taking into account further relationships between entities. In [5] RLF are introduced in order to model spatial relationships between classes in a segmentation task on real scene images. Our working hypothesis is that the use of RLF together with CRF will improve the final segmentation on documents.

We have structured the rest of the paper as follows: the CRF framework and the RLF are respectively introduced in section 2 and section 3. Then, we discuss the results of experiments in section 4. Finally, in section 5 we present conclusions and the future work.

2. Conditional Random Field Framework

Image segmentation, and in particular document image segmentation, can be seen as a pixel labeling problem. We want to find the optimal labeling configuration maximizing the *a posteriori* probability $P(c|I)$, where I is the image we want to segment, and $c = \{c_j\}$ is the set of label values corresponding to each pixel p in the image. Since the label of a given pixel also depends on the values of neighboring pixels, CRF permit us to model the pairwise dependencies between adjacent pixels in terms of energy minimization as follows:

$$P(c|I) \propto \exp\left\{ -\sum_j D_j(c_j) - \sum_{\{i,j\}\in N} V_{i,j}(c_i, c_j) \right\} \quad (1)$$

where N is the set of all the interacting pair of pixels. Thus, D_j represent the unary potentials, or how well the label c_j fits in the pixel p_j. We model D_j as the log of the *a posteriori* probability $p(c_j|x_j)$, where x_j is a feature vector given by the multi-scale Gabor filters introduced in [6]. Then, using Bayes rules, we compute $p(c_j|x_j)$ by means of the likelihood probability density function $p(x_j|c_j)$, which is estimated by a Gaussian mixture model (GMM).

$V_{i,j}$ model the pairwise pixel interaction. Values of this term represent the penalty of assigning pixel label c_i and c_j to the pixels p_i and p_j respectively, and define how smooth the final labeling will be. In the proposed approach, we have manually fixed the values of these potentials for each pair of labels according to the class configuration in the training datasets. Regular minimization methods can not be applied to minimize these energy potentials since it is a NP-hard problem. To overcome this problem, the Graph Cut algorithm was proposed in [3] to perform energy minimization of the CRF in an efficient way.

Working at pixel level will increase the processing time required to segment each document. To reduce time complexity, we define a grid $S_k \in \{S_1, \ldots, S_K\}$ of K rectangular cells over the image including all pixels. So, the goal of finding the optimal pixel labeling is replaced by the goal of finding the optimal labeling at cell level. Then, we obtain the document segmentation at pixel level by nearest neighbor interpolation.

As the GMM gives a pixel-level labeling for each of each pixels inside the cell, we define the probability of a label c_k assigned at cell level as the count of times in which the most probable label at each pixel is c_k, over the total number of pixels in a cell S_k:

$$P(c_j|S_j) = \frac{\#\{p \in S_j | c_j = \arg\max_c p(x_j|c)\}}{\#S_j} \quad (2)$$

Then, the unary potential is computed as: $D_j = -\log P(c_j|S_j)$.

3. Relative Location Features (RLF)

In this section, we explain RLF and how we can use them in a CRF framework for document segmentation tasks. RLF were introduced in [5] in the context of a multi-class segmentation problem in natural scene images. They model spatial relationships between classes, which allows them to include information about relative position of one class with respect to the other classes. For instance, on structured documents, we can encode that a text entity labeled as *header* should appear above another text entity called *footnote*.

The RLF, $v_c(S_k)$, model the probability of labeling a cell S_k by the class c by taking into account the class label and the position of other cells. $v_c(S_k)$ has been split in two features $v_c^{other}(S_k)$ and $v_c^{self}(S_k)$ (see Eq. (3)), to distinguish between the information given by cells having the same label from cells having different labels, and thus to provide a different weight to each RLF when combined in the CRF model.

First of all, the estimation of these features is obtained from probability maps which encode, for each pair of classes, the spatial relationship between them. Specifically, given two classes: c_i and c_j, and a pixel p_j, the map $M_{c_i|c_j}(u, v)$ encodes the probability that a pixel p_i at offset (u, v) from p_j belongs to the class c_i.

Secondly, we can use the initial prediction given in Eq. (2) to obtain a first image labeling composed of class labels $\hat{c}_j$ and the corresponding probability $\alpha_j = P(\hat{c}_j|S_j)$. Using this information and the pre-computed probability maps, each cell votes for the location where it is expected to find cells of other classes including its own. In that way, each cell will receive $K-1$ votes from the rest of cells, resulting in the following features:

$$\begin{aligned} v_c^{other}(S_k) &= \sum_{j\neq k:\hat{c}_j\neq\hat{c}_k} \alpha_j M_{c_k|\hat{c}_j}(x_k - x_j, y_k - y_j) \\ v_c^{self}(S_k) &= \sum_{j\neq k:\hat{c}_j=\hat{c}_i} \alpha_j M_{c_|\hat{c}_j}(x_k - x_j, y_k - y_j) \end{aligned} \quad (3)$$

where (x_k, y_k) and (x_k, y_j) are the coordinates from the centroids of the cells S_k and S_j, respectively. Using the computed RLF in Eq. (3) we propose a second approach to set the unary term D_j in Eq. (1). More precisely, RLF are linearly combined with Gabor features as follows:

$$\begin{aligned} D_j(c_j) = w^{app} \log P(c_j|S_j) + \\ + w_{c_j}^{other} \log v_{c_j}^{other}(S_j) + w_{c_j}^{self} \log v_{c_j}^{self}(S_j), \end{aligned} \quad (4)$$

where weights w^{app}, $w_{c_j}^{other}$ and $w_{c_j}^{self}$ are learned using logistic regression techniques as it was explained in [5], where more details of these features can be found.

Finally, we have normalized the features to check $\sum_c v_c^{other}(S_j) = 1$ in order to define a proper probability distribution, and similarly for the values of v_c^{self}.

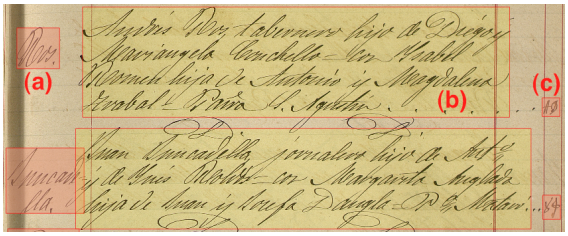

Figure 1. Two records from 5CofM database volume 208 showing the structure of the classes: (a) Name (b) Body (c) Tax

Class	Gabor features: Eq. (2)		
	Precision	Recall	F-measure
Body	0.89	0.87	0.88
Name	0.27	0.82	0.40
Tax	0.35	0.94	0.51
Class	Gabor features + **RLF**: Eq. (4)		
	Precision	Recall	F-measure
Body	0.88	0.96	0.92
Name	0.77	0.73	0.74
Tax	0.69	0.66	0.66

Table 1. Results of applying the method with and without RLF on the 5CofM dataset

4. Experiments

We have performed the experiments to evaluate the difference of applying this method on structured and unstructured documents. We present results in terms of precision and recall measures, see Eq. (5). We also include the weighted harmonic mean (F-measure), for each class of the corresponding dataset.

$$\begin{aligned} \text{Recall rate} &= \frac{\text{\# of pixels correctly identified}}{\text{\# of pixels in ground truth}} \\ \text{Precision rate} &= \frac{\text{\# of pixels correctly identified}}{\text{\# of pixels detected}} \end{aligned} \tag{5}$$

We have used two datasets to evaluate the proposed method: the 5CofM dataset (for structured documents) and the PRImA dataset (for unstructured documents). The 5CofM, or "Five Centuries of Marriages", dataset includes a set of books written between 1451 and 1905 called *Llibre d'Esposalles*. This corpus records marriages and the corresponding fees paid according to the social status of the families. It is composed of 244 books with information on approximately 550,000 marriages celebrated in over 250 parishes. Two kinds of documents are found, those regarding indexes of family names and those with the marriage license records. We have considered marriage records books for this experiment. Each page of the dataset consists of a variable number of marriage license records highly structured in three fields, from now onwards our classes: *family name, body of the record* and *paid tax* organized as seen in Figure 1. The goal of this experiment is to segment these three classes. We have used the Volume 208 which contains 512 pages. The 80 first pages of the total 512 are tagged, 75% for training and 25% for testing.

The PRImA dataset is a benchmark dataset composed of contemporary documents, like magazines and technical/scientific publications. Documents in this dataset are more loosely structured than in 5CofM, so not significant improvement is expected when RLF are used but results are useful for comparing purposes to other related methods. Although there are a bigger number of possible classes in this dataset, we have only considered the *text* and *image*. We have downloaded the dataset for training from the PRImA dataset repository. While, the test partition is provided by the ICDAR Page Segmentation Competition [2].

For both datasets, we have computed a 36-dimensional Gabor feature vector, using 9 orientations in 4 different scales. The parameters have been chosen manually but ensuring a circular Gabor support and an overlapping degree of 0.5 in the frequency domain. The highest frequency covered by the Gabor functions is approximately 0.35 (see [6] for further details on the implementation of the Gabor bank filter used). The cell size has been fixed to 50 pixels width. Finally, we have estimated the GMM, needed in Eq. (2), and the probability maps, needed to compute relative location features in Eq. (3), from the training datasets. Similarly, the regression weights used in Eq. (4) are also learnt from the training partition of each dataset.

Results on the 5CofM dataset confirm our working hypothesis. The use of RLF combined with Gabor features improves the performance of text region segmentation. We can appreciate in Table 1 how the use of RLF on the 5CofM gets twice the value of precision in classes *name* and *tax*. Also, we observe some improvement in the results on the class *body*. To assert whether differences are significant or not from a statistical viewpoint, we have computed a two-sided Wilcoxon rank sum test with a significance level of 5%. The test proves that including RLF into the CRF model significantly improves the detection results of each of the three classes.

	Gabor features: Eq. (2)		
Class	Precision	Recall	F-measure
Text	0.87	0.93	0.90
Image	0.74	0.88	0.79
	Gabor features + **RLF**: Eq. (4)		
Class	Precision	Recall	F-measure
Text	0.89	0.96	0.92 (0.0539)
Image	0.80	0.85	0.82 (0.7007)

Table 2. Results of applying the method with and without RLF on the PRImA dataset (p-value from Wilcoxon test in brackets).

As expected, the improvement after including RLF are not statistically significant in the PRImA dataset, see Table 2. The values obtained by the Wilcoxon test show that the inclusion of Relative Location Features do not produce significant improvements. We notice that the p-value in the case of detecting text regions is not far from 0.05, which means that even for this kind of loosely structured documents, RLF can help in segmenting text regions. However, for the class *image*, the obtained p-value (0.7) show that these features are useless, an expected outcome since images are located in any part of the document. Moreover, although this paper is not focused on segmenting these type of documents, the obtained results , are quite similar to or even overcome some previous works [2, 9] (see Table 3).

Method	Non-text	Text	Overall
DICE	0.66	0.92	0.90
Fraunhofer	0.75	0.95	0.93
REGIM-ENIS	0.67	0.92	0.88
Tesseract	0.74	0.93	0.91

Table 3. F-measure of the submitted methods in the ICDAR Page Segmentation Competition [2].

5. Conclusions and Future work

In this paper, we have proposed to use RLF in a CRF framework to perform document segmentation tasks. We have proved that using RLF on historical documents we improve the final document segmentation. We have also applied our method on a benchmark dataset obtaining results at the same level of the state of the art in document segmentation. Being an ongoing work, we are currently working in several aspects to improve the reported results. First of all, detection rate of *name* and *tax* classes is expected to increase using a smaller cell size. Secondly, the inclusion of more appropriate features in addition to the Gabor-based ones may reduce the classifier error rate and therefore should produce more accurate RLF. Finally, we also plan to apply learning methods to estimate the values of the pairwise potentials of the CRF.

Acknowledgment

This work has been partially supported by the Spanish projects TIN2009-14633-C03-01/03, 2010-CONES-00029 and the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018).

References

[1] A. Antonacopoulos, B. Gatos, and D. Bridson. Icdar 2005 page segmentation competition. In *8th ICDAR*, volume 1, pages 75 – 79, 2005.

[2] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos. Icdar 2009 page segmentation competition. In *10th ICDAR*, pages 1370 –1374, july 2009.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222 –1239, 2001.

[4] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *12th ICCV*, pages 670 –677, 2009.

[5] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *Int. J. Comput. Vision*, 80(3):300–316, 2008.

[6] J. Ilonen, J.-K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kalviainen. Image feature localization by multiple hypothesis testing of gabor features. *Image Processing*, 17(3):311 –325, 2008.

[7] A. Jain and S. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Machine Vision and Applications*, 5:169–184, 1992.

[8] Z. Kato and T.-C. Pong. A markov random field image segmentation model for color textured images. *Image and Vision Computing*, 24(10):1103 – 1114, 2006.

[9] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. *Image Processing*, 16(8):2117–2128, 2007.

[10] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte. Document image segmentation using a 2d conditional random field model. In *9th ICDAR*, volume 1, pages 407 –411, 2007.