EM-based Layout Analysis Method for Structured Documents

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

***Abstract*—In this paper we present a method to perform layout analysis in structured documents. We proposed an EM-based algorithm to fit a set of Gaussian mixtures to the different regions according to the logical distribution along the page. After the convergence, we estimate the final shape of the regions according to the parameters computed for each component of the mixture. We evaluated our method in the task of record detection in a collection of historical structured documents and performed a comparison with other previous works in this task.**

I. Introduction

In many areas of computer vision, the segmentation of an image on its semantic classes suppose a challenging problem due to the variability of types of objects, textures, shapes, and many other features that influence in the segmentation methods. In addition, it is common that this task represent a previous step for many other processes, so the effort to produce good segmentation methods is increasing over the years. Document analysis is not an exception facing this problem. The variability in types of documents can be as complex as in other types of images. We can think in scientific papers, magazines, maps, administrative documents, handwritten letters, historical documents and an endless of possibilities. Given this complexity, efficient and global methods are required to produce a good segmentation of the regions of interest and allow subsequent tasks to be properly applied.

In document analysis, we refer to the process of identifying and categorizing these regions as layout analysis. Depending of the type of entities that we want to detect we can make a distinction between physical and logical layout [1]. We understand by physical layout the parts of a document that can be identified without further knowledge about the semantics of the document, for instance the different text blocks, illustrations or other types of regions. On the contrary, when we refer to logical layout we assume a semantic information about the regions, for instance that will include logical categories as headline, title or footnote. The distinction between this two categories can be easily done by a human, but it is still an open problem in document analysis.

With respect to the organization of the different entities within the document, we can distinguish between two types of documents depending of the structure that they present. On the on hand there are documents that follow a predefined structure. Between these documents we include the documents with a Manhattan style and also other defined structures as could be administrative forms, registers or some other documents subjected to a specific template. On the other hand we find documents written with any trace of structure, as documents with non-Manhattan layouts.

When we deal with structured documents there are some clues that we can use to guide the search of the logical entities. For instance, it is common to find some regions from the same class located in similar regions of the page, in some cases this knowledge can be deduce by definition of the own region, as in the case of the the class *title* in newspaper pages or the *signature* in some kind of administrative documents. However, in other more generic tasks the definition of these classes may be unknown, although the distribution of the different entities may still follow a logical distribution within the page.

In this paper we propose a method to detect and categorize the regions in structured documents where the different classes of regions are located following a logical structure, and therefore, they are usually located in similar regions on each page. Our working hypothesis is that we can approximate these regions by means of a set of Gaussian mixtures and then, use the EM algorithm to fit each component to the most likely areas on the page. In addition, we propose a representation of the document using only a small sampling of pixels, which reduces the complexity of the model and consequently the computational time needed to process a page, since the computation of superpixels or other clusters is not needed. Besides, this representation allows to define the final regions at pixel level, which results in a more accurate definition of these regions in comparison to merge superpixels.

For evaluation purposes we tested our method in a record detection problem on the *BH2M* dataset. This collection of documents comprises a set of marriage licenses books where each of the pages is composed of several records distributed along the page. Each of these records are composed at the same time of three entities following a logical structure. Two segmentation problems arise from this structure, first there is the need to detect the different records, and second we want for each of them to detect the three parts that compose them.

The rest of the paper is organized as follows: Section II describes previous works in the field of layout analysis and segmentation of document images. Section III describes the different parts of the proposed method. In section IV we describe the set of features used and how are included in our model. Section V details the performed experiments and the obtained results. Finally, we conclude this paper and describe the future open lines of research in Section VI.

II. Related Work

During the last decades many methods have been proposed to deal with the problem of logical layout analysis [2], [3]. The goal on these methods is to find logical labels for the physical regions that compose the documents and is usually tackled following two main lines: model-driven and data-driven.

On the one hand, model-driven methods are characterized by the need of a previous knowledge about the documents to interpret the input data. Between these methods is common the use of grammars [4], [5] or other rule-based algorithms. Also there are some works that try to dynamically learn this structure, as the work of Rangoni *et.al.* in [6].

On the other hand between the data-driven approaches we can identify methods that take into account features computed at a more physical level level without further knowledge about the classes. As example there are methods based in the analysis of the texture [7], [8], and other that make use of the morphological properties of the document as the contours or the distribution of text lines [9].

Other works have been proposed in the past to solve the task of record detection in the *BH2M: Barcelona Historical Handwritten Marriages* database. In [10] the authors proposed a study of the inclusion of relative location prior in this problem following a CRF-based approach. In this work, although the method performed well in the detection of the parts of the records, it was not able to segment them properly. In posterior works [11], the authors treated to exploit the logical structure of the documents using 2-dimensional stochastic context free grammars with promising results in the detection of the different records. In addition, each page in these works was represented by regular cells of pixels that produced some regions to be missed in the detection process reducing the accuracy of the results. In this work we propose a method to jointly identify each record and its parts, and a representation in a set of random pixels to obtain tighter regions.

III. Proposed Method

The proposed method is devised to detect document regions and its labels. We describe each region by a rectangle, eventually rotated that we can approximate by a set of 2-dimensional Gaussian distributions. Then, by means of the EM algorithm [12], we update the parameters of these distributions to fit them to the shape and location of the regions. Once the EM algorithm converges we compute the rectangular regions represented by each Gaussian component to give the final segmentation.

In the remainder of this section we will describe each of the previous steps in detail, as well as the graphical model defined to represent the relation between the different variables taken into account for our method.

A. Model

We model the problem of assigning to each pixel of a document image a label corresponding to the number of region and the corresponding semantic class by a Conditional Random Field (CRF) [13]. In this model we combine a set of random variables (g, u), which represent local information on each of the selected pixels of the image given by a descriptor g at a particular location u. And a set of categorical random variables s, y and r representing the detected region, it class and the number of regions, respectively. The class *background* will be treated separately as is detailed in the next subsection. Therefore, our model can be expressed as the following Maximum a Posteriori inference problem:

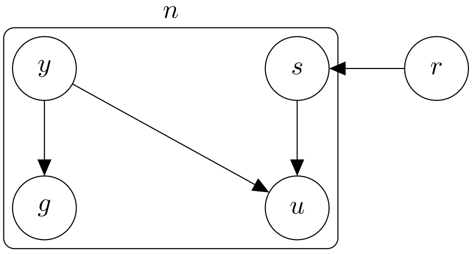

Fig. 1. Bayes net representing dependencies between model variables

$$\begin{aligned} \arg\max_{y,s,r} \quad & P(y,s,r|g,u) = \\ &= \arg\max_{y_n,s_n,r} \prod_n \frac{p(y_n,s_n,g_n,u_n|r)p(r)}{p(g_n,u_n|r)} \end{aligned} \tag{1}$$

where we reduced the problem to estimate the join densities of $p(y_n, s_n, g_n, u_n)$ and $p(g_n, u_n)$ for each pixel n. We learned the value of $p(r)$ for each possible r computing its relative frequency from the training set. We represented the variable dependencies in Fig 1 which correspond to the following hypothesis. First of all descriptors g depend on the type of regions y (incoming arrow in the node g). Secondly, not all positions u in the document image are equally probable for all the semantic labels y and regions s (incoming arrows in the node u). Finally, the values that can take s depend on the numbers of regions in the document, that we have denoted by r (incoming arrow in node s). Thus, we factorize $p(y_n, s_n, g_n, u_n|r)$ as:

$$\begin{aligned} p(y_n,s_n,g_n,u_n|r) &= p(g_n|y_n)p(u_n|y_n,s_n)p(y_n)p(s_n|r) \\ &= p(g_n,u_n|y_n,s_n)p(y_n,s_n|r) \end{aligned} \tag{2}$$

We observed that we can apply the EM algorithm to the factorization above being $p(y_n, s_n|r)$ the mixture weights of the probability $p(g_n, u_n|y_n, s_n)$. To estimate $p(g_n, u_n|y_n, s_n)$ we rely again on the graphical model of Fig. 1. On the one hand we have estimated $p(g_n|y_n)$ by estimating an GMM on the feature space of descriptors g for each semantic class y during the training step. On the other hand, we have defined $p(u_n|y_n, s_n)$:

$$p(u_n|y_n,s_n) \propto \exp\left\{-\frac{1}{2}(u_n-\mu_{y_n,s_n})^t\Sigma^{-1}_{y_n,s_n}(u_n-\mu_{y_n,s_n})\right\} \tag{3}$$

where we define the first term in terms of the probability density function of the Gauss distribution from the mixture associated to the region (y_n, s_n) centered in $\mu_{y,s}$ and variance $\Sigma_{y,s}$ and u_n represents the coordinates of the random pixel

n. Finally, we define $p(g_n, u_n|r)$ by summing conditional probabilities:

$$p(g_n, u_n|r) = \sum_{y_n, s_n} p(g_n, u_n|y_n)p(y_n)p(s_n|r) \tag{4}$$

Once we have defined the equations of our model, we need to establish how to update the parameters of the mixture to find the corresponding regions.

B. EM region estimation

As the original EM algorithm, this version follows an iterative approach to estimate the parameters of the model consisting of two main steps. During the first step (*Expectation*) we compute the join expectation of the variables (y, s) conditioned to a number of regions r. Then, in the *Maximization* step, these values are updated according to the computed probabilities in the expectation step. Initially the algorithm is started with a coarse estimation of the parameters $\{\mu, \Sigma\}$ for each of the components of the mixture as described in Section IV-A, then we run the iterative steps until the convergence of the algorithm. This process is repeated for each possible values of r previously learned from the training set. We use the usual updating equations for GMM to estimate the mixture parameters:

$$p_{new}(y, s|r) = \frac{1}{N} \sum_n p_{old}(y_n, s_n|g_n, u_n)$$

$$\mu_{y,s}^{new} = \frac{\sum_n x_n p_{old}(y_n, s_n|g_n, u_n)}{\sum_n p_{old}(y, s|g_n, u_n)}$$

$$\Sigma_{y,s}^{new} = \frac{\sum_n p_{old}(y_n, s_n|g_n, u_n)(x_n - \mu_{y,s}^{new})(x_n - \mu_{y,s}^{new})^t}{\sum_n p_{old}(y_n, s_n|g_n, u_n)}$$

$$p_{new}(y_n, s_n|g_n, u_n) = \frac{p_{new}(g_n, u_n|y_n, s_n)p_{new}(y_n, s_n|r)}{p_{new}(g_n, u_n|r)}$$

$$p_{new}(background|g_n, u_n) = \frac{1}{1 + \sum_{y_n, s_n} p_{new}(y_n, s_n|g_n, u_n)}$$

where the computation of $\mu_{y,s}^{new}$, $\Sigma_{y,s}^{new}$ and $p_{new}(y_n, s_n)$ correspond with the *Maximization* step, and the computation of $p_{new}(y_n, s_n|g_n, u_n)$, $p_{new}(background|g_n, u_n)$ and the involved probabilities with the *Expectation* step. An update on these parameters will result in a displacement and a rescaling of the components according to the configuration of the page. After some iterations these components are supposed to be centered on the regions that we wanted to detect.

C. Final labeling

We approximated the regions of a given document by a set of 2-dimensional Gaussian distributions. Now, given the distribution of each component along the page, we estimate for each of them the rectangular bounding box that comprises it. The last step of our method is to compute these rectangular regions that according to our assumptions should match up with the different regions of the page.

We defined each of these rectangles according to the parameters of each of the components. On the one hand, the center of a region will be defined by the mean of its component $\mu_{y,s}$. On the other hand we factorized the covariance matrix as $\Sigma_{y,s} = ADA^t$, where A represents the rotation matrix and D is a diagonal matrix that will define the dimensions of the region as: $[dx, dy] = 2\sqrt{3 \cdot diag(D)}$. This estimation will define a rectangular area on the image that will be labeled as the region $\{y, s\}$. Repeating this computation for every component of the mixture will give the final segmentation of the image.

IV. Features

We defined a set of features to represent the structure of the document as well as the local details at pixel level. With this aim we propose three set of features: structural features to encode the location of each of the regions for each class, texture features and relative location features to model inter-class relationships.

A. Structural features

We assumed that the different regions of a given class are usually located in similar areas within the page. Consequently we defined a set of features x_s for each of the s instances of this class found in the training set according to its shape and location. Once we defined the components of the mixture we are able to provide an initialization for the EM process described in the previous section.

Shape features are defined by computing the height $h_{y,s}$ and width $w_{y,s}$ for each region s of the class y. Note that these features are invariant to the number of regions within a page, since just give us information about the dimensions of each entity of a class. In what regards to the location we made a distinction between the different values of s, and compute the distance $u_{y,s}$ from the center of each region to the origin of the page for each of the observations. Finally, we define structural feature vectors as $x_s = (u_{y,s}, w_{y,s}, h_{y,s})$ for the possible values of s.

We define Normal distributions for each pair of elements (y, s). Thus, the mean of the resulting Gaussian is defined as $\mu_{u_{y,s}}$. To define the covariance matrix $\Sigma_{y,s}$ we took into account that the variance of the size of a region is independent from the location of it. Since each of the covariance matrices will represent the dimensions of each region we propose to approximate each covariance by two Uniform distributions, one for each of the dimensions, so according to the definition of these distributions and the relation between the variances of Normal and Uniform distributions, the final form of this matrix is defined by:

$$\Sigma_{y,s} = \begin{pmatrix} \frac{\mu_{h_{y,s}}^2}{12} + \sigma_{h_{y,s}}^2 & 0 \\ 0 & \frac{\mu_{w_{y,s}}^2}{12} + \sigma_{w_{y,s}}^2 \end{pmatrix}$$

where $\mu_{w_{y,s}}$ and $\mu_{h_{y,s}}$ are the mean of the widths and heights computed from the training set, and $\sigma_{w_{y,s}}$ and $\sigma_{h_{y,s}}$ the corresponding variances.

B. Texture features

Texture descriptors are regularly used in segmentation tasks to extract information from local areas of an image. In this work, following the outline in previous works [10], [11] we used a multi-resolution filter bank to compute a set of filter responses for several orientations and frequencies ensuring invariance to position or orientation changes of the elements on the image. We defined a set $g \in \mathbb{R}^{f \times q}$ which corresponds with the set of Gabor filter responses computed for f frequencies and q orientations at pixel level.

To compute the probability $p(g_n \mid y_n)$ required by our model in Eq. (2), we trained a Gaussian mixture model over the feature vectors provided by the filter bank.

C. Relative Location Features

We defined features to obtain local descriptors from each one of the classes considered in our problem. However, those features focus just in a particular class without taking into account any relation between them. One way to represent this information is to add relative location prior to our model. In [14] the authors describe how to include this prior in the context of real scene segmentation. In previous works [10] we encoded this information into Relative Location Features (RLF) and included them into the segmentation process of structured documents. We proved that the improvement produced by these features was significant enough to affirm its validity to include inter-class spatial relationships between the different classes. Each of these features model the probability of assigning the class label of the region y to a pixel u taking into account the information provided by the rest of image pixels about their location and its estimated label predictions.

To encode this features in the first place we computed a set of probability maps for each pair of classes. These maps give us the most probable areas of the image where to find elements from one particular class with respect to the others. For example, in a newspaper segmentation problem one map may represent that is more common to find regions of the class *title* above elements from the class *footnote*. In the second place, we computed an initial labeling for each of the pixels using the GMM trained with the texture features described before. Finally, combining both probability maps and the initial labeling we obtained the values for the RLF. Further information about this features can be seen in the original paper [14] or in previous works [10], [11].

Once computed the values of RLF for each pixel we compute:

$$\log p(y_n \mid g_n) = w^{tex} \log p(y_n|g_n) + \\ + w_{y_n}^{other} \log v_{y_n}^{other}(u_n) + w_{y_n}^{self} \log v_{y_n}^{self}(u_n) \quad (5)$$

where $v_{y_n}^{self}(u_n)$ and $v_{y_n}^{other}(u_n)$ are the different sets of RLF for the class y and $w_{y_n}^{tex}$, $w_{y_n}^{self}$ and $w_{y_n}^{other}$ are the corresponding weights learned from a logistic regression model. Then, using Bayes rules we are able to compute the probability $p(g_n \mid y_n)$ required by our model in Eq. (2).

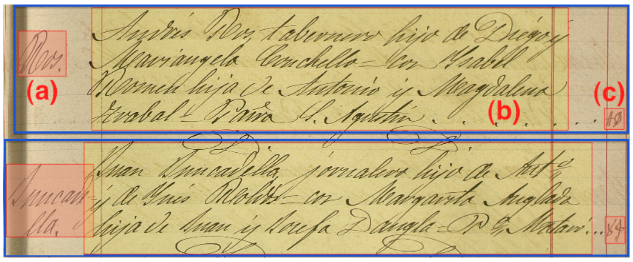

Fig. 2. The figure shows two records cropped from a page of the *BH2M* dataset framed in blue bounding boxes. Each record is composed by the three regions framed in the red bounding boxes. (a) Name (b) Body (c) Tax.

V. Experiments

We performed several experiments to evaluate the proposed method in the task of logical layout analysis on structured documents. In this paper we presented an application of our method to the task of detect marriage licenses from the *BH2M* dataset, on which some works have been published before and we are able to perform an exhaustive qualitative and quantitative comparison of the results.

A. Database

The *BH2M: Barcelona Historical Handwritten Marriages* dataset includes a set of handwritten documents written by ecclesiastical institutions for centuries that register licenses of marriages. The entire collection is composed of 244 books with information on approximately 550,000 marriages celebrated along five different centuries. Each page of the collection consist of a variable number of marriage license records structured in three fields: the husband surname's block , the central body of the license and the tax block distributed as can be seen in Fig. 2. From now onwards we will refer to these elements as *name, body* and *tax*, respectively. The goal of this experiment is to detect the different records s as well as the three classes of entities y that compose them. We used in these experiments the documents from the volume 208 of the collection, and following the approach on previous work we used the 80 first pages, 75% of them used for training and the rest 25% for testing.

B. Settings

We performed two main experiments using two different combination of the features described in Section IV. In both experiments we used the described location features to initialize the Gaussian mixtures for the EM process, however we made a distinction in the manner in which we include local information of the image. First we test our method using only texture features, and in the second place we will use instead the relative location features to evaluate the possible improvement produced by them.

We defined two stop criteria for our algorithm. On the one hand we fixed the maximum number of iterations of the EM to 50, we empirically checked that iterations over this value do not produce significant changes in the result. On the other hand we established the convergence criterion according to the Kullback-Leibler divergence between two consecutive iterations for the distribution defined by $p(y, s, r|g, u)$, we

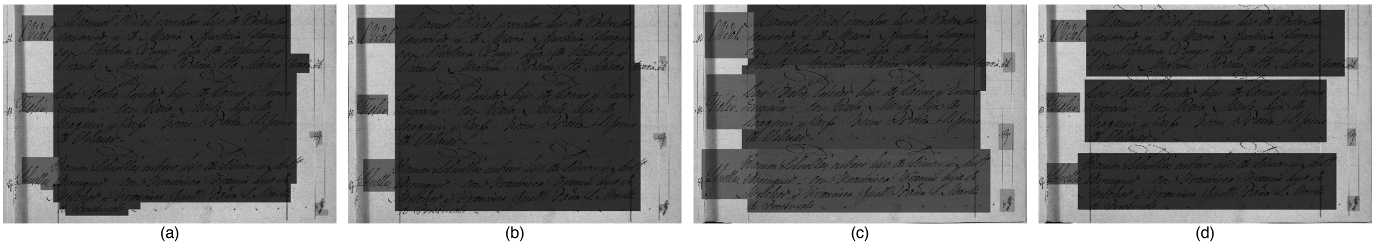

Fig. 3. Comparison between the results obtained by previous methods with respect to the proposed for a specific page. The loss of region portions can be appreciated in images (a) and (b), as well as the fusion of the different records by the central column. The proposed method (c) splits the central region and distinguishes between the different records assigning different labels to them. (a) CRF (b) 2D-SCFG (c) Proposed (d) GroundTruth.

Model	Class	Gabor		
		Precision	Recall	F-measure
CRF	Body Name Tax	0.89 0.27 0.35	0.87 0.82 0.91	0.88 0.40 0.50
2D SCFG	Body Name Tax	0.88 0.72 0.39	0.97 0.93 0.94	0.92 0.81 0.54
Proposed	Body Name Tax	0.83 0.27 0.20	**0.93** **0.99** **0.95**	0.88 0.41 0.32

TABLE I. PAGE SEGMENTATION RESULTS FOR DIFFERENT MODELS AND TEXTURE CLASSIFICATION FEATURES.

Model	Class	RLF		
		Precision	Recall	F-measure
CRF	Body Name Tax	0.88 0.77 0.69	0.96 0.73 0.66	0.92 0.74 0.66
2D SCFG	Body Name Tax	0.88 0.82 0.63	0.97 0.83 0.69	0.92 0.82 0.64
Proposed	Body Name Tax	0.85 0.39 0.27	**0.93** **0.94** **0.90**	0.89 0.54 0.41

TABLE II. PAGE SEGMENTATION RESULTS FOR DIFFERENT MODELS AND RELATIVE LOCATION FEATURES.

fixed a convergence value for $\epsilon \ll 0$. With respect to the parameters of the method, we fixed the number of random points in 20,000 to ensure an acceptable computational time and a proper coverage of the page. The mean size of an image in this dataset is $4,000 \times 2,000$ pixels, which means that we are able to process a page with only the 2% of the total information on it. We learned the value of the number of records r from the training set, and run our algorithm for the interval from 5 to 7 records. Regarding to the features, we used the training partition to learn the probability maps and the weights to build the RLF and the GMM for texture and location. In the case of the Gabor filter bank we computed a 36-dimensional feature vector, using 9 orientations at 4 different scales.

We considered two criteria to evaluate the quality of the results. On the one hand we are interested in the correct detection of the logical layout of the page. For that we will require the three regions in a record to be correctly identified. On the other hand we want to perform quantitative and qualitative comparison of our approach with respect to previous works. With that purpose we show results in terms of precision and recall of the detection of the three regions computed at pixel level, as well as the F-measure for each of them.

C. Discussion

We show in Tables I and II the results obtained for the three different classes compared with the ones obtained in previous works. Overall results show a similar trend in the detection of each class. Regions from classes *Body* and *Name* are detected with higher F-measure rates, whereas the class *Tax* still represent the most challenging part.

We can see as in terms of F-measure rates the proposed method does not produce significant improvements in the detection of the three classes, however we think that this value do not fairly reflects the real contribution of our method. The contribution for this task can be seen in the Recall values with respect to the other methods. With the proposed method we are able to detect more than 90% of all the regions using any set of features, which in the domain of our task should be considered as a great advantage. In addition, given that the objective of our method is to obtain a logical layout of the page, we evaluated the obtained results by the number of pages properly segmented (*i.e., the number of detected records equal to the number of records in the page.*), getting an 80% of accuracy, which represent an improvement with respect previous works.

In previous works based in the inference of CRFs the area between two records was merged with the *Body* class. That produced an overlapping of all the records and consequently the method was not able to return a proper result of the document layout. In the case of 2D-SCFG we defined the grammar to find a set of records within a page. That definition allowed to extract the implicit region of a record from the output of the grammar. However this method was sensitive to the detection of all the regions from a record, since the case of an undetected region will suppose a fail in the detection of the whole record. Besides, because of the noise present in some areas of the image, sometimes the grammar forced the detection of inexistent records affecting to the analysis of the rest of the page. These situations were improved by the proposed approach. Our method is designed to detect the three elements of a record, so that in the case of detecting a record, all three parts will be always included on it. Besides, the components of the mixture are initialized taking into account the structure of the page, so we avoid the problem of overestimate the number of records.

Another problem of previous approaches was the representation of the page in a grid of cells. This representation produced that some of the small regions from classes *name* and *tax* were missed and therefore the detection of record was not correct. The proposed representation based in a small sample of random pixels avoid the loss of small regions in comparison with the approach based in cells, since some pixel will be always included in these small areas.

All the previous situations are illustrated in Fig. 3. Analyzing this image it is also possible to understand the loss in the precision rate. We can see that using the proposed approach the regions obtained match with the expected results, although the obtained area for the classes *name* and *tax* is usually wider that the area in the GroundTruth. That effect is produced mainly by two reasons. On the one hand the area around these regions usually presents some noise that the EM tries to fit into the region, also in most of the records the *name* and *body* regions are overlapped, so it is common that some of the overlapped pixels influence the corresponding component of the mixture. Nevertheless, we think that providing a better definition of the probabilities in $p(g_u|y_u)$ this situation can be smoothed producing better overall results. This could be achieved either using other classifiers or other set of features more suitable to the task that we want to tackle. On the other hand the variance of the size and location of these classes is higher than in the case of the class *body*. That provokes that at the beginning the components are initialized with larger values in the covariance matrix and because of the noise in the region is difficult to fit to the proper region.

In what regards to the use of RLF with respect to texture features, we proved in previous works that the use of these features improved the results with methods that do not take the structure into account, as in the case of CRFs. In the presented method the structure was already implicit in the definition of the classes, so the use of RLF in this case do not produce any remarkable improvement.

VI. Conclusion

We presented a method based in the EM algorithm to perform logical layout analysis in structured documents. Our proposal is that we can learn about the structure of the document and approximate the different regions that compose it by mixtures of 2-dimensional Gaussian distributions. For that we defined EM-based method where the components of the mixture are initialized according to a configuration learned from a training set. Then in each iteration the parameters of the components are updated to be adjusted to the most likely areas of the image. We also defined a representation using a small portion of the pixels on the image, which speed up the process without loss of precision.

We tested our method in the task of detecting marriage records in historical structured documents with positive results, either in the detection of the different records as the regions that compose them. However, since the development of this method is an ongoing work we were not able to compare our method with any other layout analysis works. We are currently working in the application of the method to the benchmark PRImA dataset, used in the ICDAR Page Segmentation Contest, as well as in other tasks of logical layout analysis on structured documents.

Another ongoing work is being carried out in order to avoid the possible overlapping between components. It could happen that two or more components try to fit in the same region producing a fragmentation within the region. In this case we are working in interpose restrictions of separability between components into the EM process, as well as the option of combine two or more of them.

In addition we have other open research lines to improve the proposed approach in this paper. On the one hand, one of the drawbacks of our method was the need to know an estimation in advance of the number of regions for each of the classes. In this paper we proposed to compute the mos probable configuration and select the one that maximize a defined criterion. The problem is that in other types of documents with more variable layouts this criterion may result computationally very expensive. In future versions we plan to tackle this problem by adding to the EM process the ability to set null probability values to the surplus components, becoming the process more dynamic.

Acknowledgment

This work have been partially supported by the European Research Council Advanced Grant (ERC-2010-AdG 20100407-269796) and the project TIN2012-37475-C02-02.

References

[1] R. M. Haralick, “Document image understanding: Geometric and logical layout,” in *ICPR*, 1994, pp. 385–390.

[2] S. Mao, A. Rosenfeld, and T. Kanungo, “Document structure analysis algorithms: A literature survey,” 2003.

[3] F. Montreuil, S. Nicolas, E. Grosicki, and L. Heutte, “A new hierarchical handwritten document layout extraction based on conditional random field modeling,” in *ICFHR*, 2010, pp. 31–36.

[4] A. Conway, “Page grammars and page parsing. a syntactic approach to document layout recognition,” in *ICDAR*, 1993, pp. 761–764.

[5] M. Lemaitre, E. Grosicki, E. Geoffrois, and F. Prêteux, “Preliminary experiments in layout analysis of handwritten letters based on textural and spatial information and a 2d markovian approach,” in *ICDAR*, 2007, pp. 1023–1027.

[6] Y. Rangoni, A. Belaid, and S. Vajda, “Labelling logical structures of document images using a dynamic perceptive neural network.” *IJDAR*, vol. 15, pp. 45–55, 2012.

[7] A. K. Jain, N. K. Ratha, and S. Lakshmanan, “Object detection using gabor filters,” *Pattern Recognition*, vol. 30, no. 2, pp. 295–309, 1997.

[8] 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*, vol. 16, no. 8, pp. 2117–2128, 2007.

[9] M. Bulacu, R. Koert, L. Schomaker, and T. Zant, “Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen,” in *ICDAR*, vol. 1, 2007, pp. 23–26.

[10] F. Cruz and O. R. Terrades, “Document segmentation using relative location features,” *ICPR*, pp. 1562–1565, 2012.

[11] F. Alvaro, F. Cruz, J. Sanchez, O. R. Terrades, and J. Benedi, “Page segmentation of structured documents using 2d stochastic context-free grammars,” *IBPRIA*, vol. 7887, pp. 133–140, 2013.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[13] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *18th ICML*, ser. ICML ’01, 2001, pp. 282–289.

[14] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, “Multi-class segmentation with relative location prior,” *IJCV*, vol. 80, no. 3, pp. 300–316, 2008.