Bayesian mixture models on connected components for Newspaper article segmentation

Giorgos Sfikas Georgios Louloudis Nikolaos Stamatopoulos
Basilis Gatos

Institute of Informatics and Telecommunications
NCSR "Demokritos"
Athens, Greece
{sfikas, louloud, nstam, bgat}@iit.demokritos.gr

ABSTRACT

In this paper we propose a new method for automated segmentation of scanned newspaper pages into articles. Article regions are produced as a result of merging sub-article level content and title regions. We use a Bayesian Gaussian mixture model to model page Connected Component information and cluster input into sub-article components. The Bayesian model is conditioned on a prior distribution over region features, aiding classification into titles and content. Using a Dirichlet prior we are able to automatically estimate correctly the number of title and article regions. The method is tested on a dataset of digitized historical newspapers, where visual experimental results are very promising.

1. INTRODUCTION

Digitization of machine-printed and handwritten documents of various kinds has led to a subsequent demand for tools to handle more effectively the digitized content. Automatic document understanding techniques have been employed to ease access to scholar, students, or casual users alike. Page layout analysis, optical character recognition and keyword spotting [10] are techniques that are widely used today, and attract considerable interest from the part of end-users as well as researchers.

Newspapers are a special class of documents, from the point of view of their layout as well as their content structure. Content is organized into a set of articles, presented typically in a small number of columns. Article tracking, or otherwise called article identification or article segmentation [7], is the process of segmenting a scanned input page into a set of semantically-coherent regions that would correspond to the area covered by each newspaper article in the scanned image. Article tracking typically is built on top of a page segmentation step that clusters the page into sub-article areas containing article body text, titles, supertitles, or other content.

Newspaper segmentation can be considered as a special form of generic document or page segmentation. Segmenting newspaper pages can be aided if the used algorithm encodes successfully the prior knowledge that newspaper layout is in practice constrained. For example, newspapers present their material as a series of semantically-coherent articles, which are usually presented in a number of columns. On the other hand, newspapers can be also more difficult to segment than the average document, due to factors that may hinder the performance of generic algorithms. A typical factor that may hinder performance is the usually close contact of regions in newspapers [7].

In this work we present a novel algorithm for segmentation of a newspaper image into articles. Our algorithm identifies articles after performing newspaper segmentation at a lower-level, where the page is clustered into sub-article components. We experiment with using a Fully Bayesian Gaussian mixture model (FBGMM) [5] as a classifier in the current context of newspaper image data. Connected component analysis provides input to the FBGMM, used here to cluster connected components into region types and title groups. We take advantage of the fully probabilistic structure of the used mixture model to include parameter priors and detect numbers of components automatically. Tests on a number of digitized historical newspapers show promising results.

The remainder of this article is organized as follows. In section 2 we provide the reader with a brief overview of related work in newspaper segmentation. We briefly present the Bayesian model we use as a classifier in section 3. In section 4 we present the proposed algorithm in detail. In section 5 we present the dataset of newspaper pages we used for experiments, and present experiment visual results. Finally, we conclude the paper and discuss future work in section 6.

2. RELATED WORK

Works in newspaper article segmentation in general aim first to segment the input page into sub-article layout components, before merging these into output articles. Sub-article components can be tagged with regard to the content of the region they refer to. Possible choices are mainly tagging a region as text-containing or image-containing, though other considerations are also possible. Such a consideration would be using a separate class for titles and another one for article body text, or to an even finer class taxonomy. In [8] for example, a total of 12 different classes to tag text regions are possible.

A considerable number of works propose methods to segment the newspaper page into sub-article layout components, without proceeding to use these as a base step to identify articles. We shall briefly refer to a number of such works here. In [19] a neural network is used to classify a newspaper image into image and text

DocEng '16, September 12 - 16, 2016, Vienna, Austria

ACM ISBN 978-1-4503-4438-8/16/09…$15.00

DOI: http://dx.doi.org/10.1145/2960811.2967165

regions. The neural network is trained on high-probability output data of a preliminary segmentation step. In [1], a neural network is trained to identify text and non-text regions, using a manually created training set and a variety of hand-crafted features. In newspaper segmentation competitions [6] and [2], the competitor systems were required to cluster the image into sub-article components. In [13], the winning system of the competition [6], regions are clustered by merging together connected components, after estimating region boundary positions.

Segmentation into articles, or otherwise referred to as article tracking or article identification, has been addressed in [7]. In this work, a rule-based method is used to group text, images and titles into articles. Sub-article components are produced in the first stage of the method. First vertical/horizontal line extraction is performed, followed by image/drawing region identification and text/title extraction. A DFT-based algorithm is used to identify image areas, and RLSA with adaptive parameters is used to identify text areas. Regions are grouped into articles with a rule-based decision stage that includes over 40 empirical rules. Subsequent works [8, 15, 16, 14] are closely related to this base model for article tracking. Recent works in article tracking [11, 18] also use sets of hand-crafted rules to segment pages into articles.

Extraction of horizontal and vertical lines is recognized as an important step for the success of the whole algorithm in a multitude of works [7, 12, 17]. This is due to the fact that lines, either made up of foreground pixels or by background pixels (i.e. virtual separating lines), are typically used to demarcate region and article areas. In [12], the problem of line extraction for degraded documents in particular is studied. In [13] and [17], lines are extracted before merging connected components into regions. In the current work, we also use line separator detection as a step in our processing pipeline, and also use connected components as a base entity that is merged into sub-article regions and articles.

More complex models have been proposed to encode the newspaper hierarchy of articles and regions. A Conditional Random Field is used in [18] to model the newspaper page structure into coherent regions. In [3] a learning-based method, using a fixed-point model, has been proposed to identify articles. In the current work, we have experimented incorporating a Fully Bayesian Gaussian Mixture Model (FBGMM) [5] as a classifier for an article tracking pipeline. FBGMM has the advantage of being able to automatically estimate the number of classes of the data to be modeled, in contrast to the typical paradigm where the user has to specify the number of classes beforehand. In the context of newspaper segmentation, we use FBGMM to handle parameter prior knowledge and model region features, where the number of regions is a priori unknown.

3. BAYESIAN MODEL

We have used a Fully Bayesian Gaussian Mixture Model as a CC classifier, as part of the proposed segmentation pipeline. The Bayesian model is presented in detail and solved [1] in [5, ch. 10]. Input X is assumed to be a set of N real-valued vectors of dimension d, $X = [x^1, \cdots, x^N]$. These are modeled as being generated by a set of (maximum) K classes. For each class j, a one-zero vector z^n keeps track of the class responsible for generating the datum. Responsibility vectors z^n follow a multinomial distribution, $z^n \sim Mult(w)$ where w is a K-sized vector that encodes *a priori* responsibilities. All vectors are grouped in $N \times d$ matrix Z. Class emissions are modelled as Gaussians of class-specific mean μ_j and covariance Σ_j, with $x_j^n \sim N(\mu_j, \Sigma_j)$, $\mu = [\mu_1, \cdots, \mu_K], \Sigma = [\Sigma_1, \cdots, \Sigma_K]$. Model parameters μ, Σ, Z are assumed to follow prior distributions that are conjugate to the Gaussian and Multinomial distributions (these are respectively Gaussian/Wishart and Dirichlet priors); this enables us to solve the otherwise intractable model with approximate inference [5, ch. 10]. Solution of the model returns posteriors given observations, over model parameters. For the current problem, we are specifically interested in the posterior for Z, i.e. $p(Z|X)$. Input data are clustered according to the class that maximizes their responsibility in the posterior for Z.

As part of the proposed algorithm, we use the FBGMM for clustering in two different parts of the pipeline. We cluster CCs into text, content, and non-text/non-content elements. We exploit the Bayesian formulation of the FBGMM to specify a prior mean for each component. Prior means are specified in terms of a base mean $\hat{\mu}$, with priors given as $[.25\mu, 4\mu, 20\mu]$.

In the second use of the FBGMM, the Dirichlet prior hyperparameter is set to a small value $\alpha = 10^{-6}$ and maximum $K = 50$ to enable the model to automatically estimate the true number of clusters. The clusters in this case correspond to title regions.

4. SEGMENTATION MODEL

The proposed algorithm for newspaper segmentation is made up of two main stages. In the first stage regions containing text, titles or other content are detected. These sub-article regions are merged in the second stage of the algorithm to create article-level regions.

We first run Connected Component Analysis on a binarization of the input page image [9]. Connected Components (CCs) are used throughout the algorithm as the element of base that is used to form regions of all levels. Convex regions of CCs with a major axis dominant over their minor axis are marked as separators. The CCs minus separators and CCs that are too small (compared to the average CCs) are clustered with FBGMM as described in the previous section (section 3). We have used 3 features for each CC. These are CC width, CC height and CC thickness. CC thickness is computed as the maximum value of each CC's distance transform. The resulting clustering is smoothed by a max-voting step.

Regions, title or content, are split when overlap with a separator is detected. Content CCs are then dilated in an RLSA-like step in both dimensions. We exclude title and separator pixels from the resulting dilation, and CCA is run to form content regions.

Article tracking is performed on the basis of the previous segmentation step of the image into title and content regions. Title regions are first grouped into title groups. Title regions are dilated in the vertical direction and CCA is performed on the image, excluding all pixels marked as non-title. The resulting CCs form our title groups. Afterwards content regions are assigned each to a single title group. We assign each content region to the nearest title group that is found above the region. Groups of title groups and assigned content regions form the required article segments.

5. DATASET AND EXPERIMENTS

In order to test our article tracking system, we used samples from a collection of digitized historical editions of the regional greek newspaper "*Tharros*". *Tharros* started being published as far back as 1899, and is still today in circulation in Greece. In this paper we show results of our method for 5 pages of *Tharros*, coming from editions published in different dates. While the test set is admittedly small, we feel that these pages are representative, to an extent, of most of the core variations in layout of the newspaper throughout its years of publication. We used pages with publication dates: February 9, 1901; February 18, 1917; June 4, 1948; March 1, 1975; June 23, 1989. Results of application of our method on

[1] An implementation can be found at http://www.cs.uoi.gr/~sfikas

Algorithm 1 Proposed newspaper segmentation algorithm

Segmentation into sub-article components
- Connected Component Analysis
  - Compute CC features
  - Detect separator components
- Cluster CCs into content, title, other components
  - Use FBGMM with appropriate priors
  - Smooth clustering with max voting
- Detect title regions
  - Estimate number of regions with FBGMM
  - Cluster into title regions with FBGMM
- Post-processing
  - Split regions when overlap with separators

Article tracking
- Merge neighbouring titles into title groups
- Assign content regions to a unique title group
- Merge titles and content into articles

these images can be examined in fig. 1. Intermediate stages and the segmentation end result are shown. Comparison of the article tracking result with the input images suggests that our method in general identifies the number and position of articles correctly.

6. CONCLUSION AND FUTURE WORK

In this paper we have presented a novel system for newspaper segmentation into article and sub-article components. We used a Bayesian GMM, solved with Variation inference, to cluster the page more accurately.

In this work we have presented qualitative/visual results. Due to lack of annotated ground truth, we have not presented numerical results. Numerical methods for evaluation of region and article tracking have been considered elsewhere [7, 3, 4]. We envisage presenting numerical evaluation figures for our method, with tests on a dataset of larger scale in the future.

Samples from a single newspaper were used in the current work. While our algorithm is in principal not dependent to the specific layout of the *Tharros* newspaper, and throught the years there is considerable variance in the paper layout, we would like to test the method on other/different editions. Discriminating into more subclasses of content, and providing an estimate of reading order is also left as future work.

Finally, we envisage working on a unified model for newspaper document understanding, where layout analysis in multiple levels and optical recognition would work jointly, and not as independent tasks.

7. REFERENCES

[1] T. Andersen and W. Zhang. Features for neural net based region identification of newspaper documents. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 403–407, 2003.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. ICDAR 2013 competition on historical newspaper layout analysis (HNLA 2013). In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1454–1458, 2013.

[3] A. Bansal, S. Chaudhury, S. D. Roy, and J. Srivastava. Newspaper article extraction using hierarchical fixed point model. In *Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS)*, pages 257–261, 2014.

[4] R. Beretta and L. Laura. Performance evaluation of algorithms for newspaper article identification. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 394–398, 2011.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] B. Gatos, S. Mantzaris, and A. Antonacopoulos. First international newspaper segmentation contest. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1190–1194, 2001.

[7] B. Gatos, S. Mantzaris, K. Chandrinos, A. Tsigris, and S. J. Perantonis. Integrated algorithms for newspaper page decomposition and article tracking. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 559–562, 1999.

[8] B. Gatos, S. Mantzaris, S. Perantonis, and A. Tsigris. Automatic page analysis for the creation of a digital library from newspaper archives. *International Journal on Digital Libraries*, 3(1):77–84, 2000.

[9] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern recognition*, 39(3):317–327, 2006.

[10] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou. A survey of document image word spotting techniques. *Submitted to Pattern Recognition*, 2016.

[11] A. Jain, V. Sahasranaman, S. Saxena, and K. Chaudhury. Segmenting printed media pages into articles, Oct. 16 2012. US Patent 8,290,268.

[12] A. Lemaitre, J. Camillerapp, and B. Couasnon. Approche perceptive pour la reconnaissance de filets bruités, application à la structuration de pages de journaux. In *Colloque International Francophone sur l'Ecrit et le Document*, pages 61–66. Groupe de Recherche en Communication Ecrite, 2008.

[13] F. Liu, Y. Luo, M. Yoshikawa, and D. Hu. A new component based algorithm for newspaper layout analysis. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 1176–1180, 2001.

[14] S. Mantzaris, B. Gatos, and N. Gouraros. Creating a digital library from newspaper archives. In *SDIUT*, page 285, 2001.

[15] S. Mantzaris, B. Gatos, N. Gouraros, and S. Perantonis. Linking article parts for the creation of newspaper digital library. In *RIAO*, pages 997–1005, 2000.

[16] S. Mantzaris, B. Gatos, N. Gouraros, and P. Tzavelis. Integrated search tools for newspaper digital libraries. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, page 389, 2000.

[17] P. E. Mitchell and H. Yan. Newspaper layout analysis incorporating connected component separation. *Image and Vision Computing*, 22(4):307–317, 2004.

[18] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, and T. Paquet. Logical segmentation for article extraction in digitized old newspapers. In *Proceedings of the ACM symposium on Document engineering (DocEng)*, pages 129–132, 2012.

[19] P. S. Williams and M. D. Alder. Generic texture analysis applied to newspaper segmentation. In *IEEE International Conference on Neural Networks*, volume 3, pages 1664–1669, 1996.

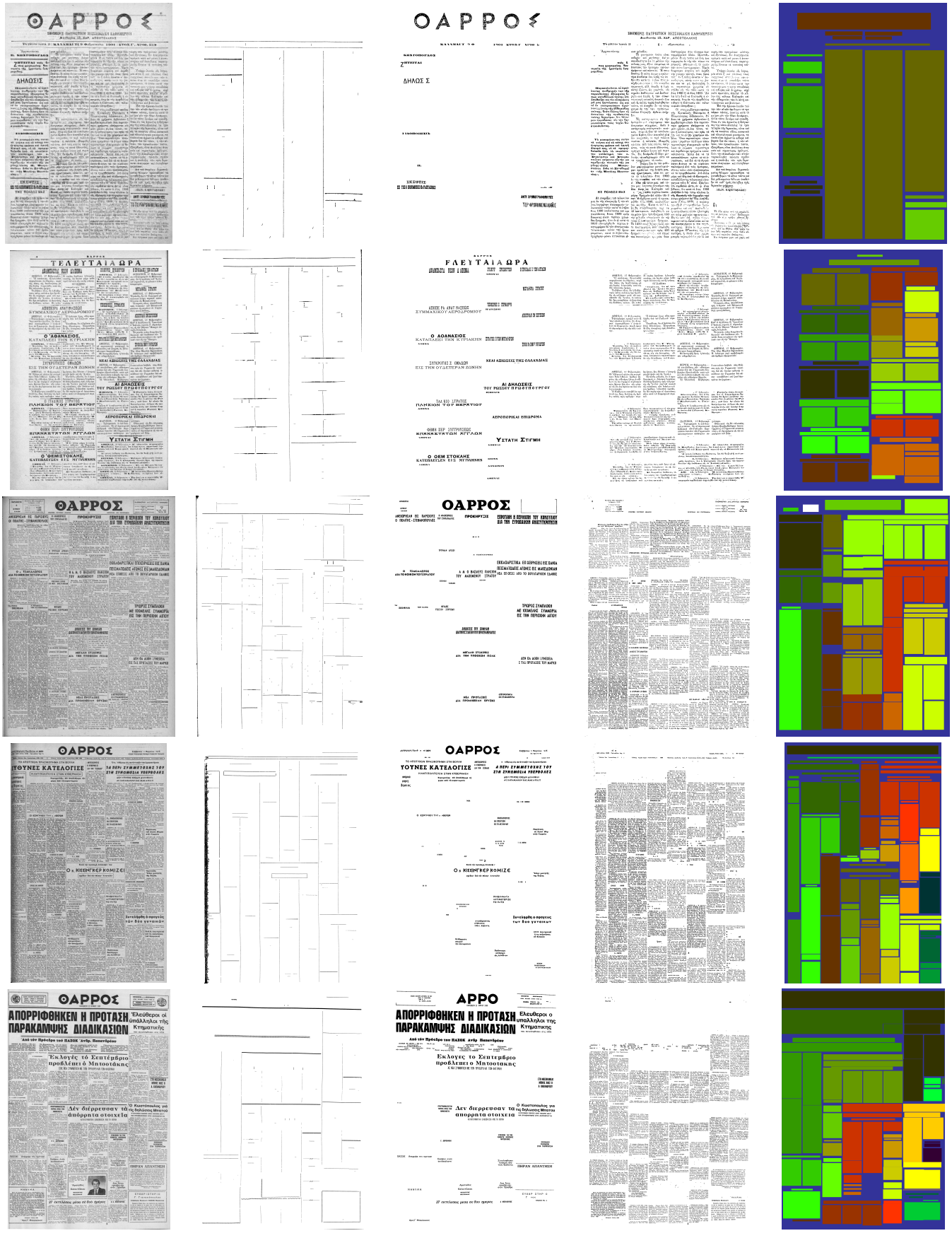

Figure 1: Sample of newspaper pages we used to test the proposed method and corresponding results using our method. All samples are historical editions of the greek newspaper "Tharros". From top to bottom, years of publication for the depicted pages are: 1901, 1917, 1948, 1975, 1989. From left to right, we show: the original image; detected separating lines; titles only; content only; segmentation into articles. Regions under the same article are painted with the same colour.