

A Robust Front Page Detection Algorithm for Large Periodical Collections

Iuliu Konya Christoph Seibert Sebastian Glahn Stefan Eickeler
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Iuliu.Konya, Christoph.Seibert, Sebastian.Glahn, Stefan.Eickeler} @iais.fraunhofer.de

Abstract

Large-scale digitization projects aimed at periodicals often have as input streams of completely unlabeled document images. In such situations, the results produced by the automatic segmentation of the document stream into issues heavily influence the overall output quality of a document image analysis system. As a solution to the issue segmentation problem, this paper introduces a robust, two-step front page detection algorithm. First, the salient connected components from the front page of the periodical are described using a multi-dimensional Gaussian distribution based on discrete cosine transform (DCT) features. Second, a graph model is computed by applying Delaunay triangulation on the selected set of components. A specialized, error-tolerant graph matching algorithm is used to compute the distance score between the model and each candidate page. Experiments on a large, real-world newspaper data set demonstrate the generality and effectiveness of the proposed method.

1. Introduction

A large number of scientific articles have been published on topics concerning specific areas of document image analysis (see surveys [1, 11]). In comparison, the problem of front page detection for issue separation is still largely unexplored, although it is indispensable when working with large collections of unlabeled images of periodicals (such as magazines or newspapers).

Two topics closely related to front page detection are logo recognition and structural pattern recognition. The detection and recognition of logos from document images are still being actively pursued in the document analysis community [3, 21]. In the vast majority of papers dealing with logo recognition, logos are handled as unitary entities and various statistical classifiers (e.g. support vector machines, neural networks, Fisher

classifiers, etc.) are used for the classification task. In contrast to the common approaches for logo recognition, in the area of structural pattern recognition objects are described by their topology or shape, most commonly encoded as different types of graphs. The graphs can afterward be matched exactly or error-tolerantly using existing methods, many of which have a strong theoretical basis (e.g. [13, 19]). Such approaches offer a great deal of flexibility, as proved by their successful application in domains as distinct as object recognition via skeletons or shapes [4, 17] and on-line graphics recognition [20]. Their main drawback however, is that structural approaches are inherently sensitive to noise, therefore specialized solutions are necessary for dealing with this problem. Noise robustness is of particular importance when working with real-world, error-prone document scans. Also, one must note that graph and subgraph isomorphism in general is an NP-complete problem, thus matching is only feasible for small graphs (exact speed/graph size measurements can be found in [13]). For special cases, such as closed contours, algorithms exist which are at the same time fast and guarantee the finding of optimal solutions [17]. In most other cases, in order to circumvent the graph size issue, different heuristics for approximating the optimal matches have been proposed [8, 20].

The front page detection algorithm introduced in this paper combines the advantages of both statistical and structural pattern recognition methods. As presented in detail in section 2, statistical models are used for describing and detecting the salient parts of a front page, which are in turn connected into a structural model. In this way, the amount of noise (i.e. false graph vertices or edges) which must be dealt with during the creation and matching of the structural model is greatly reduced. This allows us to use a simple and efficient matching algorithm, while still achieving a high degree of robustness. A large newspaper image data set was used to validate and find the optimal parameters for the proposed method, as shown in section 3.

2. Methodology

The proposed algorithm works by computing and assigning a reference weighted graph to a front page model. As vertices in the constructed graph we use the salient connected components from the front-page specific elements, such as the title and the logo of the periodical. The exact methodology for the choice of salient components is described in section 2.2. Each of the salient components is subsequently described by a Gaussian distribution, provided several training samples. For all candidate pages, a weighted graph is computed in a similar fashion from those connected components which fit one of the trained Gaussian distribution models. Finally, the graph associated to each candidate page is matched against the front page graph. A correspondence is found if the matching score for the best correspondence exceeds the threshold value associated to the front page model. The methodology for calculating the edge weights and the threshold for each model graph is presented in section 2.4.

2.1. Pre-processing

We assume as input to the algorithm a deskewed and binarized document image. Many methods for page skew detection and document image binarization have been described in the literature. According to recent surveys and comparisons [1, 7], there exist several algorithms capable of accurately determining the document skew, as well as some which are able to adequately binarize complex and/or degraded document images. It is important to note that most existing techniques for geometric and logical layout analysis of documents make the same assumption about the input document image [1]. This means that a regular document image analysis system will not require any additional (computational or implementation) effort in order to satisfy the precondition of our front page detection method.

2.2. Identification of Salient Components

From the binarized image, the connected components (both black and white) must be labeled. This can be accomplished efficiently using any standard algorithm (e.g. [5]). In the case of noisy input images, it is advantageous at this point to apply certain hole-filling algorithms or morphological operators so as to remove a significant part of the white/black noise. This additional step is very helpful for subsequently obtaining more consistent feature descriptors for the salient components. Note that the rest of the current section is only relevant for training the front-page specific salient component models.

Next, we have to choose the salient components from among the set of labeled connected components. Most commonly the chosen salient components will be characters, connected character groups or logos specific to the front page model being considered. The choice of the salient components must currently be performed at least in part manually, as we are not aware of any fully automated selection method satisfying (most of) the criteria presented in the following. Some general guidelines for choosing the salient components are: salient components are not likely to be merged with other connected components even in the presence of noise, the selected set of salient components should span over as much of the title section area as possible and each component should be large enough so that the danger of it being mistaken for noise is minimized. For practical purposes, it is relatively easy to implement (as was done in our case) a semi-automatic salient component candidate selection method. This method can simplify the manual task considerably by filtering out those components which do not satisfy a set of simpler criteria, such as a minimum area, a certain aspect ratio and a high enough proximity to other candidate components.

As an alternative for automating the choosing of the salient components, one may resort to one of the existing affine region detectors [14]. A good repeatability and accuracy is provided for example by maximally stable extremal regions [12], i.e. those parts of an image where local binarization is stable over a wide range of thresholds. The density of the detected stable regions contained in or partially overlapping a certain connected component can be used as a clue about the stability of the respective component. Although any salient region detector may be used, a very important factor to consider at this point is its runtime performance, which varies widely [14].

2.3. Feature Extraction and Matching

As features for describing a salient component, we have chosen to use the coefficients of the discrete cosine transform (DCT) applied to the component's bitmap. The DCT is known to be very suitable for decorrelating the pixels of images. It is used in the JPEG image compression standard [16]. In our case, the feature descriptor for the image of a component contains only the coefficients in the upper left triangle (low-pass) of the DCT transformed image. Early experiments showed that using the first 36 coefficients generally give good results, but decreasing or increasing the number of coefficients to 21, 28 or 45 produces very similar results. It is worth noting that the number of 36 coefficients was also determined to be the most appropriate from the visual experiments performed during the development of

the MPEG-7 standard [10]. The feature extraction is similar to the one used in [6] for face recognition, where it was also shown that the probability density function of the descriptors can be accurately modeled by a Gaussian distribution with a diagonal covariance matrix.

Note that it is entirely feasible to use a different set of features instead, such as those commonly employed for character recognition [18] or for generic content-based image retrieval [2]. As observed in [18], there exist no features which perform best in all situations and therefore the best feature type is in general highly dependent on the application domain.

An important reason for selecting the DCT was its scaling invariance property and the efficient algorithms available for its computation. The scaling invariance is important because many publishers frequently re-scale the size of their front page titles and/or logos (e.g. in order to accommodate for last minute news or advertisements). The DCT equation employed is:

$$C(u, v) = \sum_{x=0}^W \sum_{y=0}^H \frac{I(x, y)}{WH} \cos \left[\frac{(2x+1)u\pi}{2W} \right] \cos \left[\frac{(2y+1)v\pi}{2H} \right]$$

Here, W and H respectively represent the width and the height of the connected component, and $I(x, y)$ represents the image gray value at the position (x, y) .

Other types of features, such as the Zernike moments, also offer rotation- and limited distortion invariance. In our case, the rotation invariance is rendered unnecessary by the skew correction already applied on the document (as a prerequisite). The distortion invariance is of very limited practical use for real-world document images, since such pronounced distortions are extremely rare in any professional scans, as commonly employed in large-scale digitization projects.

A Gaussian distribution with a diagonal covariance matrix is now trained for each salient component using as input several of its feature descriptors (obtained from different example front pages). From the performed experiments, we have determined that between 5 and 10 samples per component are sufficient for obtaining an accurate Gaussian distribution model. In order to prevent overfitting, a relatively large floor value of 0.1 was used for the variances.

Given the set of trained Gaussian distributions, one can subsequently determine for any candidate component the distribution with the highest output probability for its feature vector. In case the output probability is higher than a certain a-priori fixed threshold value, a match is considered to have been found.

2.4. Graph Construction and Matching

For modeling the topology of the identified salient components, we construct the point Delaunay triangulation [9] using as input the centers of their bounding boxes.

As an alternative to the Delaunay triangulation we have also considered and tested the Euclidean minimum spanning tree (EMST) applied to the same set of points. Although the EMST has the advantage that the number of edges is about 3 times lower than for a triangulation (thus the graph matching problem is substantially simplified), the appearance of the resulting tree is more heavily influenced by small position changes of the tree vertices. This phenomenon can be observed easiest when (some of) the sites are almost uniformly distributed - in such case, even small perturbations of the node positions can produce very different spanning trees, whereas the Delaunay triangulation remains almost unchanged.

Even with the increased structural stability provided by the Delaunay triangulation, the obtained graphs are still sensitive to the loss of internal vertices. For enabling a robust graph matching, we additionally assign a positive real weight to each edge. Assuming that the probability of a node to be missing from the candidate graph is the same for all vertices, one can readily see that the probability of certain edges to be absent from the resulting triangulation is higher than for others. Our goal is that the edge weights are selected in such a way that for each node, the sum of the weights of the adjacent edges will always be equal to a certain constant. At the same time we want to minimize the absolute differences between the weights of each pair of edges, so that no edges will have a (near) zero weight. Unfortunately, this problem is NP-hard, because of its equivalence to the problem of finding a solution with a maximum number of zeros in an underdetermined linear system, which was shown to be NP-hard in [15]. Because of this, one must settle for a heuristic function approximating the ideal edge weights. A possible edge weight computation function which performed well in our tests is:

$$\text{Weight}(e) = \frac{1}{1 + (\deg(v_1) - \deg_{min}) + (\deg(v_2) - \deg_{min})}$$

Here, $e = (v_1, v_2)$ is an edge in the graph between vertices v_1 and v_2 , \deg_{min} is the minimum vertex degree in the graph and $\deg(v_i)$ is the degree of vertex v_i .

By taking into account the computed edge weights, graph matching can now be readily performed on an edge-by-edge basis. The matching algorithm traverses the edges of the model graph in decreasing weight order and searches for the best correspondence within the candidate graph. If a possible correspondence is found, the total distance between the two graphs is incremented by the model edge weight multiplied with the modulus of the sinus of the angle between the two edge vectors. The total distance must not exceed a certain threshold



Figure 1. Three front page graph models, each with a correctly recognized candidate, illustrating: a) limited distortion resilience; b) scaling invariance; c) occlusion.

ratio of the total sum of the edge weights of the model graph. The experiments described in the following section helped in determining a generally suitable range for the threshold ratio.

3. Experimental Results

The test were carried out on a collection of 17 572 images from 1141 newspaper issues, scanned at 200 or 300 dpi. The document images have been provided by two major German-language publishers. Five different front page models were trained and subsequently the front page detection algorithm was run on the entire document collection. We have tested two different threshold ratios for graph matching, as it was not clear what value would constitute a well-performing generic threshold.

Table 1. Test results on a collection of 17 572 real-world newspaper images

Model	θ	Pages	Issues	Rec. rate [%]	Recall [%]	Pr. [%]
LV 1930	50%	1516	264	99.27	95.83	100
	70%			99.93	99.62	100
LV 1970	50%	7719	589	99.97	99.66	100
	70%			99.99	99.83	100
LV 1986	50%	3991	293	100	100	100
	70%			100	100	100
DK 1940	50%	2955	74	99.86	94.59	100
	70%			100	100	100
DK 1960	50%	1391	32	99.35	71.87	100
	70%			100	100	100

In the previous table, we have included not only the

recognition rate, but also the precision and the recall, because the disparity between the number of front pages and regular pages make the recognition rate much less meaningful.

From the obtained results, one can see that a 70% graph matching threshold θ is in general the better choice. It is important to notice that the 50% threshold, experimentally determined as being too high, generally implies that significantly more than half of the model graph's edges are present in a candidate graph, due to the additional penalties applied for all imperfect vector matches. The overall results obtained are encouraging and show that the described method can be applied with good results in digitization projects on a mass scale.

4. Conclusion and Future Work

We proposed a novel approach for front page detection in periodical publications. The technique can easily be adapted to employ different features for modeling the salient components and in principle allows the use of any existing error-correcting subgraph isomorphism algorithms. These characteristics facilitate the integration of our algorithm in any existing document analysis systems. Experiments performed on an extensive real-world newspaper image set show that our method is both robust and fast enough for practical use. In our future work we will address the issue of automatic salient component identification, as well as try to experimentally determine the best features to use in the context of scanned periodical images spanning over extended time intervals.

References

- [1] R. Cattoni, T. Coianiz, S. Messelodi, and C. Modena. Geometric layout analysis techniques for document image understanding: a review. Technical Report 9703-09, ITC-irst, 1998.
- [2] A. Chalechale, A. Mertins, and G. Naghdy. Edge image description using angular radial partitioning. *IEE Proc. Vision, Image and Signal Processing*, 151(2):93–101, 2004.
- [3] J. Chen, M. K. Leung, and Y. Gao. Noisy logo recognition using line segment Hausdorff distance. *Pattern Recognition*, 36(4):943–955, April 2003.
- [4] C. di Ruberto. Recognition of shapes by attributed skeletal graphs. *Pattern Recognition*, 37(1):21–31, January 2004.
- [5] M. B. Dillencourt, H. Samet, and M. Tamminen. A general approach to connected-component labeling for arbitrary image representations. *J. ACM*, 39(2):253–280, 1992.
- [6] S. Eickeler, S. Müller, and G. Rigoll. Recognition of JPEG compressed face images based on statistical methods. *Image and Vision Computing Journal, Special Issue on Facial Image Analysis*, 18(4):279–287, March 2000.
- [7] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, March 2006.
- [8] S. Gold and A. Rangarajan. Graph matching by graduated assignment. In *Proc. 1996 Conf. Computer Vision and Pattern Recognition*, pages 239–244. IEEE Computer Society, 1996.
- [9] L. Guibas and J. Stolfi. Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *ACM Trans. Graphics*, 4(2):74–123, April 1985.
- [10] ISO/IEC JTC1/SC29/WG11/N3321. MPEG-7 visual part of experimentation model version 5, March 2000.
- [11] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: A literature survey. In *Document Recognition and Retrieval X*, volume 5010, pages 197–207. SPIE, 2003.
- [12] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Proc. 13th British Machine Vision Conf.*, volume 1, pages 384–393. British Machine Vision Association, 2002.
- [13] B. T. Messmer. *Efficient Graph Matching Algorithms*. PhD thesis, University of Bern, Switzerland, 1995.
- [14] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Computer Vision*, 65(1–2):43–72, November 2005.
- [15] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 24(2):227–234, 1995.
- [16] W. B. Pennebaker and J. L. Mitchell. *JPEG Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1993.
- [17] F. R. Schmidt, D. Farin, and D. Cremers. Fast matching of planar shapes in sub-cubic runtime. In *IEEE 11th Int'l Conf. Computer Vision*. IEEE Computer Society, 2007.
- [18] O. D. Trier, A. K. Jain, and T. Taxt. Feature-extraction methods for character-recognition: A survey. *Pattern Recognition*, 29(4):641–662, April 1996.
- [19] W. H. Tsai and K. S. Fu. Subgraph error-correcting isomorphisms for syntactic pattern recognition. *IEEE Trans. Systems, Man and Cybernetics*, 13(1):48–62, 1983.
- [20] X. Xu, Z. Sun, B. Peng, X. Jin, and W. Y. Liu. An online composite graphics recognition approach based on matching of spatial relation graphs. *Int'l J. Document Analysis and Recognition*, 7(1):44–55, March 2004.
- [21] G. Zhu and D. Doermann. Automatic document logo detection. In *Proc. 9th Int'l Conf. Document Analysis and Recognition*, volume 2, pages 864–868. IEEE Computer Society, 2007.