Feature Approach for Printed Document Image Analysis

Jean Duong[1,2], Myrian Côté[1], and Hubert Emptoz[2]

[1] Laboratoire d'Imagerie Vision et Intelligence Artificielle (LIVIA)
Ecole de Technologie Supérieure, H3C 1K3 Montréal, (Québec) Canada
{duong,cote}@livia.etsmtl.ca
[2] Laboratoire de Reconnaissance de Formes et Vision (RFV)
Institut National des Sciences Appliquées (INSA) de Lyon
Bâtiment 403 (Jules Vernes)
20 Avenue Albert Einstein, 69621 Villeurbanne CEDEX
{duong,emptoz}@rfv.insa-lyon.fr

Abstract. This paper presents advances in zone classification for printed document image analysis. It firstly introduces entropic heuristic for text separation problem. Then a brief recall on existing texture and geometric discriminant parameters proposed in a previous research is done. Several of them are chosen and modified to perform statistical pattern recognition. For each of these two aspects, experiments are done. A document image database with groundtruth is used. Available results are discussed.

Introduction

In spite of the wide spread use of computers and other digital facilities, paper document keeps occupying a central place in our everyday life. Conversely to what was expected, the amount of paper produced presently is larger than ever. Important institutions like administrations, libraries, archive services, etc. are heavy paper producers and consumers. To some point of view, paper is one of the most reliable information supports. Unlike numerical records, it is not constrained by format compatibility question, or device needs.

On the other side, document storage for safety or accessibility considerations is a very tricky problem. Research is presently done in such a direction. The primary goal of document analysis and recognition is to transform a paper document into a digital file with as less information loss as possible. Many successive tasks are needed to achieve this purpose. A document image has to be produced and processed for graphic enhancement. Then physical regions of interest have to be found, labelled according to their type (text, graphic, image, etc.), ordered (hierarchically and spatially). Finally, various kinds of information may be retrieved in different ways within certain regions. For example, text can be found via optical character recognition (OCR) in text regions and stored as ASCII data while images may be compressed.

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 159–167, 2002.
© Springer-Verlag Berlin Heidelberg 2002

Here we are concerned with printed document images. We assume that preprocessing is done and zones of interest are found. We focus on document zone classification task. This paper introduces some entropic heuristics (section 1) to achieve it. A recall about some features proposed by different authors in order to label regions physically (section 2) is done. For the most commonly used ones, a relevance study based on statistical considerations is conducted. Experiments are done on the MediaTeam Document Database and the UWI document database to validate our views.

1 Entropic Features

1.1 Text/Non-text Separation

In previous works [5,4], we introduced entropy heuristics to separate text zones from non-text ones in a black and white printed document image. As stated in [6,7], text areas will have rather regular horizontal projections while non-text elements will give projections more like random distributions (see Fig. 1). These projections are commonly stored as histograms. Thus, it is possible to compute their entropy values. Let H be the histogram representing the horizontal projection for a given region. Its entropy will be

$$E(H) = \sum_{i=1}^{n} \frac{H[i]}{n} \ln\left(\frac{H[i]}{n}\right) \tag{1}$$

assuming the index for histogram entries runs from 1 to n. If entropy is computed for every zone of interest in a given document image, this will result in low and high values for text and non-text areas respectively. Exploiting this last remark, we have been able to discriminate rather efficiently text elements from other regions of interest in various documents. Thus, entropy on horizontal projection is considered as a potentially valuable feature. To validate this assumption, we have performed some experiments which will be discussed in section 3.1.

1.2 Extensions

We have developed an adaptative binarization method to be performed on greyscale document images: within each zone of interest, we gather grey levels in two groups for low and high values via a deterministic variant of the *k-means* algorithm. Pixels with moderate or large grey scales are respectively set to black or white. Our binarization procedure is implicitly based on grey levels distribution. This histogram may carry useful information for region labelling: a text area is likely to have its grey values distribution more regular (in general bimodal) than a graphic zone. Thus, its entropy is estimated and may be considered as an interesting feature for further experiments.

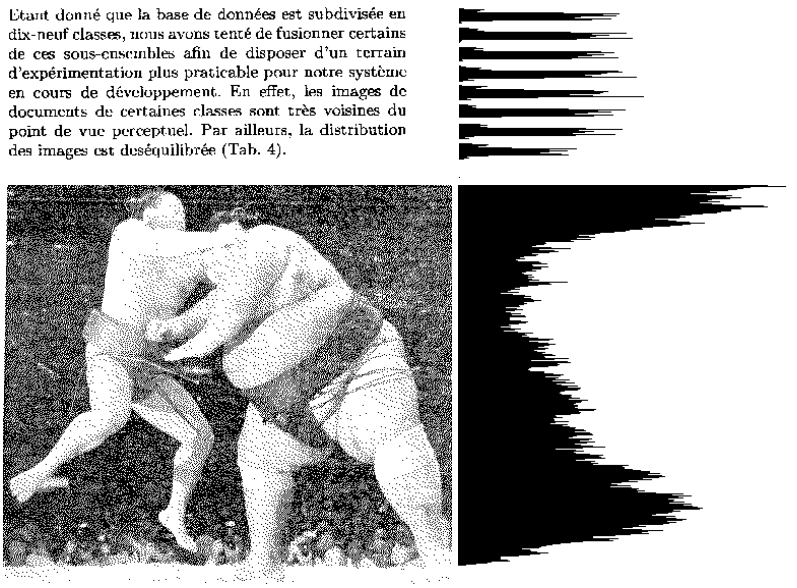

Fig. 1. Example of horizontal projection histograms for a text block and an image

More generally, entropy calculus is a convenient way to measure approximately the information conveyed in a distribution. It allows us to map a vector of a priori unknown size to a scalar. For this reason, we also use it to "compress" vertical projection, north south east and west profiles.

2 Document Zone Classification

Several "classical" features are well known in handwriting recognition. Namely, concavities, surface, profile, etc. Conversely, in printed document analysis, such a common research background does not exist. Early use of features for document zone classification can be found in [16]. A set of commonly used characteristics can be derived from recent surveys [8,10,13,11]. Most of the systems compute values for a certain set of features and perform a rule-based labelling of document zones. Many thresholds appearing in classification rules are set empirically or experimentally and tuned separately from each other. Thus, a global qualitative study remains to be done. To achieve this, we have selected the features that most frequently appear in publications.

A given document area is defined by its bounding box. Values for the entropic features introduced in precedent section are estimated. After binarization, the following measures are also computed for each region of interest.

- Eccentricity (fraction of the width to the height).
- Black pixels (ratio of black pixels population to the surface of the region).
- Horizontal relative cross-count (number of "white to black" and "black to white" transitions in horizontal direction divided by the surface of the region).
- Vertical relative cross-count.
- Mean length of horizontal black segments normalized by the region's width.
- Mean length of Vertical black segments normalized by the region's height.
- Connected components population to the region's surface.

Except for the two first ones, all these features were actually found in coarser versions (i.e. without normalization) in various works.

From now, regions of interest are represented as real vectors with fourteen components. Features can then be compared in term of relevance. Data analysis procedures are particularly well suited for this task.

3 Experiments

We ran most of our experiments with the MediaTeam Document Database[12]. This database is a collection of color or greyscale printed document images with groundtruth for physical segmentation.

3.1 Validation for Entropy Heuristic

The purpose in this first set of experiments is to test the relevance of entropy on horizontal projection histogram as a feature to separate text areas from the others. Regions of interest are retrieved via a coarse segmentation based on gradient image. Given a document image, entropy on horizontal projection is computed for each zone. Areas with low or high entropy values are labeled text or non-text respectively. To achieve this separation, a deterministic variant of the *k-means* algorithm (see Appendix) is performed over the entropy values for the image. Experimental results are presented in [5,4].

3.2 Feature Analysis

Using only one feature, we have been able to discriminate text from non-text in noisy and complicated printed document images with decent performance. Actually, the classification procedure worked locally and assumed the existence of one text zone and one non-text zone in every document image at least. This hypothesis is not fulfilled for all document images of MediaTeam. We now explore the feasability of text/non-text separation involving many characteristics. In this set of experiments, all the regions of interest (given by the database groundtruth) over all the images in MediaTeam are mapped to fourteen dimensional pattern

Table 1. Patterns distribution in MediaTeam Document Database

Pattern type	Samples
Text	4811
Graphics	735
Image	161
Composite	219

vectors. We obtained 5926 of such vectors distributed in four classes, as shown in Table 1. We try to improve the accuracy and the generality of our classification in term of text/non-text separation.

Our data are obviously insufficient and too badly distributed to train and test a neural network or a markovian process [9,3]. Thus we have decided to use classical data analysis and support vector machines for our experimental purposes. These tools are well suited to deal with the kind of data we dispose of. Due to unbalanced classes (see Table 1), classical learning machines as neural networks may lead to overfitting problems for certain classes. SVM classifiers, on the other hand, have shown robust behavior against overfitting phenomena caused by unbalanced data distribution.

Support Vector Machines Let us consider the following set of data for a two class problem: n feature vectors are called X_i with $i \in \{1, \ldots, n\}$ and $X_i \in \mathbb{R}^d, \forall i \in \{1, \ldots, n\}$. Each d-dimensional vector X_i is labelled y_i where $y_i \in \{-1, +1\}, \forall i \in \{1, \ldots, n\}$. According to its label, a vector will be said to be a negative or a positive example. A support vector machine (SVM) works seeking for optimal decision regions between the two classes. In the original formulation, the SVM searches for a linear decision surface by maximizing the margin between positive and negative examples.

Unfortunately, in most of the real-life classification problems, data are not linearly separable. They are mapped in higher dimension space via a non-linear application ϕ called *kernel* (see Table 2 for most commonly used kernels). With an appropriate kernel operating from the original feature space to a sufficiently high dimension space, data from two classes can always be separated [15,14]. The final decision function will be of type $f(X) = \sum_{i=1}^{n} \alpha_i y_i k(X_i, X)$ where k is a given kernel, $\alpha_i \in [0, C], \forall i \in \{1, \ldots, n\}$. Vector X_i is said to be a *support vector* if the corresponding α_i is non-null. C is a cost parameter. Allowing C to tend to infinity leads to optimal separation of the data, at the price of increasing processing time.

Table 2. Classical kernels for support vector machines

Kernel	**Formula**
Linear	$k(X,Y) = X.Y$
Sigmoïd	$k(X,Y) = tanh(\alpha X.Y + \beta)$
Polynomial	$k(X,Y) = (1 + X.Y)^d$
Radial Basis Function (RBF)	$k(X,Y) = e^{-\alpha\|X-Y\|^2}$
Exponential RFB	$k(X,Y) = e^{-\alpha\|X-Y\|}$

Table 3. Results for text/non-text separation using a SVM with RBF kernel. Vectors not used in training set are all considered test patterns

Learning set		**Cost**	α	**Support**	**Learning**	**Classification**
Text	Non_text			**vectors**	**accuracy**	**accuracy**
2000	500	1	$\frac{1}{14}$	816	90.44%	**8**9.87%
		10		586	94.88%	93.11%
		100		450	96.76%	93.81%
		1000		383	98.12%	93.49%
		100	0.1	447	97.36%	93.46%
		100	0.01	556	93.68%	92.38%
3000	500	100	$\frac{1}{14}$	513	96.94%	91.67%
2000	750	100	$\frac{1}{14}$	559	95.60%	**9**4.89%

Two-Class Separation Improvement Here our purpose is to improve the text/non-text separation using fourteen features. "Graphics", "Image" and "Composite" patterns are gathered in one "Non_text" class. To choose the most fitted approach, we first have to estimate the hardness of our task.

We perform a linear discriminant analysis (LDA) with all the pattern vectors for both training and validation. Observed classification accuracy is 67.09%. This leads us to conclude that our problem may be non-linearly separable (Theoretically, the problem can be considered linearly separable if the obtained accuracy is 100% with LDA.). A trial with a linear support vector machine (SVM), supposed to be the most powerful linear classifier [2], confirms this assumption: obtained classification accuracy is only 87.54%.

Since many types of SVMs exist, different sets of experiments have to be done to determine the best suited classifier. Finally, the SVM with RBF kernel shows the best performances. A preliminary collection of experimental results presented in Table 3 is available. Some subtle tradeoff between the size of the learning set, its distribution, values for kernel parameter, cost threshold remains to be found.

Table 4. "*Separability* estimation using a support vector machine with RBF kernel

Cost parameter	α	Support vectors	Accuracy
100	$\frac{1}{14}$	945	95.83%
500	$\frac{1}{14}$	861	97.11%
1000	$\frac{1}{14}$	804	97.43%
10000	$\frac{1}{14}$	705	**98.48**%

Comments When considering figures presented in Table 3, one must take into account the context of the classification task. As we did to examine the linearity of the problem, we use all the pattern vectors to train a support vector machine with RBF kernel. Different experiments (see Table 4) show that our problem is everything but trivial.

3.3 Recent Advances

Many problems arise, while using the MediaTeam document database. The corpus is not sufficiently large to deploy most of the statistical learning techniques. Moreover, proposed documents are from very different types (nineteen document classes are found in the database, some of them with less than a dozen of images). Thus, we have decided to perform another set of experiments with a more specific document database.

We have computed the above-presented characteristics over the regions of interest proposed in UWI document database. This collection consists of 1000 pages from different english journals. Since the document images are binary in this database, we dropped the grey distribution entropy characteristic. Our calculus resulted in 10573 patterns vectors. These 13-dimensional vectors are distributed as following: 9307 samples for text regions and the other 1266 ones for non-text regions.

We used SVM classifier with the KMOD kernel, newly designed by Ayat et al. [1]. This kernel's specification is given by the equation

$$kmod\,(x,y) = a\left[exp\left(\frac{\gamma}{\|x-y\|^2+\sigma^2}\right)-1\right] \tag{2}$$

$\sigma = 0.01$ and $\gamma = 0.001$ are two parameters that jointly control the behavior of the kernel function. σ is a space scale parameter that define a gate surface around zero whereas γ controls the decreasing speed around zero. In other words, σ measures the spread of the kernel function, and γ describes the shape of this function within this domain. We set empirically $\sigma = 0.01$ and $\gamma = 0.001$ [1]. The normalization constant a is defined as

$$a = \frac{1}{e^{\frac{\gamma}{\sigma}}-1} \tag{3}$$

Table 5. Results using SVM with KMOD kernel. Parameters σ and γ (in formula 2) are set to 0.01 and 0.001 respectively

Cost	0.1	1	10	100
Accuracy	91.10	97.06	97.34	97.34

UWI is both more homogeneous and more voluminous than MediaTeam. To avoid the problem of designing training and test sets, we performed a five-fold cross-validation on our data set: we divided the data into five subsets of (approximately) equal size. We trained the classifier five times, each time leaving out one of the subsets from training, and using it for test. Accuracy is defined as the mean value over the five obtained performance score for tests. Table 5 shows results for such experimental setting. Statistical examination should be conducted on data to establish an efficient preprocessing. The features we chose were interesting since they are frequently used by different authors. But some of them may be strongly correlated (redundant), due to a lack of standardization. Experiments are currently in progress to perform feature selection and extraction. Results will be presented in further publications.

Conclusion

This paper was intended to show some recent developments in printed document image analysis and several gaps in related features normalization. It also proposes a way to fill these lacks. Application to document zone classification is presented. Some basic features are selected for their simplicity and a statistical examination is performed.

The use of common statistical tools and support vector machines has been proved to be adequate for this kind of problem. Other experiments are in progress to optimize learning parameters. SVM paradigm is still under development in machine learning research community [1]. As a consequence, better results for document region classification may be obtained in a near future. The following step will be to investigate multiple class discrimination for thinner document zone classification. This should help document logical labelling process. Many other characteristics have to be jointly tested. Some will surely have to be dropped. This will be part of our further work.

Acknowledgements

We wish to thank our colleague N.E. Ayat and Professor M. Cheriet for very instructive discussions about SVM paradigm and helpful advice for experimentations.

References

1. Nedjem E. Ayat, Mohamed Cheriet, and Ching Y. Suen. Kmod-a two parameter svm kernel for pattern recognition, 2002. To appear in ICPR 2002. Quebec city, Canada, 2002. 165, 166
2. C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20, 1995. 164
3. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley Interscience, 2001. 163
4. Jean Duong, Myriam Côté, and Hubert Emptoz. Extraction des régions textuelles dans les images de documents imprimés. In *Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, Angers (France), Janvier 2002. 160, 162
5. Jean Duong, Myriam Côté, Hubert Emptoz, and Ching Y. Suen. Extraction of text areas in printed document images. In *ACM Symposium on Document Engineering (DocEng)*, pages 157–165, Atlanta (Georgia, USA), November 2001. 160, 162
6. K.C. Fan, C.H. Liu, and Y.K. Wang. Segmentation and classification of mixed text/graphics/image documents. *Pattern Recognition Letters*, 15:1201–1209, 1994. 160
7. K.C. Fan and L.S. Wang. Classification of document blocks using density feature and connectivity histogram. *Pattern Recognition Letters*, 16:955–962, 1995. 160
8. Robert M. Haralick. Document image understanding: Geometric and logical layout. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4, pages 384–390, 1994. 161
9. Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):4–37, Januray 2000. 163
10. Anil K. Jain and Bin Yu. Document representation and its application to page decomposition. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 20(3):294–308, March 1998. 161
11. George Nagy. Twenty years of document image ananlysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):38–62, January 2000. 161
12. University of Oulu (Finland). Mediateam document database, 1998. 162
13. Oleg Okun, David Dœrmann, and Matti Pietikäinen. Page segmentation and zone classification: The state of the art, November 1999. 161
14. B. Scholkopf, C. Burges, and A. Smola. *Advances in Kernel Methods: Support Vector Learning*, chapter 1. MIT Press, 1999. 163
15. Vladimir Vapnik. *The nature of Statistical Learning Theory*. Springer Verlag, New-York (USA), 1995. 163
16. Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. Document analysis system. *IBM Journal of Research and Developpment*, 26(6):647–656, November 1982. 161