16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AUTHORIAL MANUSCRIPT IMAGE ANALYSIS USING MARKOVIAN MODELS: THE BOVARY PROJECT

Stéphane Nicolas, Laurent Heutte, Thierry Paquet

Université de Rouen, LITIS EA 4108
BP 12, 76801 Saint Etienne du Rouvray, FRANCE
phone: + (33) 2 32 95 52 04, fax: + (33) 2 32 95 50 22, email: laurent.heutte@univ-rouen.fr
web: www.litislab.eu

ABSTRACT

In this paper, we present our recent work on the Bovary project, a manuscript digitization project of the famous French writer Gustave Flaubert, which aims at providing an online access to a hyper textual edition of "Madame Bovary" draft sets. We first describe the global context of this project, its main objectives, and then focus on the document image analysis problem for which we have developed effective and powerful page layout extraction algorithms based on markovian generative and discriminative models such as Hidden Markov Random Fields and Conditional Random Fields. We show with some experiments that these stochastic and contextual models are able to cope with local spatial variability while taking into account some prior knowledge about the global structure of the document image, thus being well suited to the segmentation of very complex document images like authorial manuscripts or historical documents.

1. INTRODUCTION

Libraries and museums contain collections of a great interest which can not be shown to a large public because of their value and their state of conservation, therefore preventing the diffusion of knowledge. Today with the development of the numerical technologies, it is possible to show this cultural patrimony by substituting the original documents by numeric high quality reproductions allowing sharing the access to the information while protecting the originals. These last years, numerous libraries have started digitization campaigns of their collections. Faced to the mass of the numerical data produced, the development of digital libraries allowing access to these data and the search for information becomes a major stake for the valuation of this cultural patrimony. The valorisation of our cultural heritage using digital technologies requires robust document image analysis methods allowing to recognize the structure and to detect areas of interest facilitating the indexing of large corpora of ancient documents and the production of digital libraries. These last years improvements have been made in the field of handwriting recognition, especially in the context of industrial applications such as check reading, postal address recognition or form processing. These applications have mainly focused on word or phrase recognition [1]. Recently, the advance in digital technologies has motivated numerous institutions to move towards the use of digital document images rather than traditional paper copies of the original documents. This situation raises new needs for indexing and accessing to these numerical sources [1]. This is why we explore in this paper the field of document image analysis solutions for computer aided indexation of large corpora of patrimonial documents.

We first present in section 2 the Bovary Project, a digitization project of modern manuscripts concerning especially FLAUBERT's manuscripts and we discuss the underlying principles, difficulties and technical aspects related to such a project. Then in section 3 we discuss about document image analysis problem and we propose a method based on markovian models. An application to the analysis of the manuscripts of FLAUBERT is then presented and the obtained results are discussed in section 4.

2. GENERAL PRESENTATION OF THE PROJECT

Although they have a great interest for the study and the interpretation of literary works, modern manuscripts have been addressed by few digitization programs because of the complexity of such documents and the lack of adapted tools. In 2003 the municipal library of ROUEN has begun a program for digitizing its collections. For this purpose an efficient system of digitization allowing a high resolution display of the digitized documents has been purchased. One of the first aims of this program is the digitization of a manuscript folder compound of almost 5 000 original manuscripts issued from "Madame Bovary", a well known work of the French writer Gustave FLAUBERT. This set of manuscripts constitutes the genesis of the text; it means the successive drafts which highlight the writing and rewriting processes of the author. This digitization task is now completely finished. The final aim of this program is to provide a hyper textual edition allowing an interactive and free web access to this material. Such an electronic edition will be of great interest for researchers, students, and anyone who wants to see FLAUBERT's manuscripts, especially because there is no critical edition of a full literary work of this author available on the web. This project called "Bovary Project" is a multidisciplinary project which implies people from different fields of interest: librarians, researchers in literary sciences and researchers in computer science.

In literary sciences the study of modern manuscripts is known as genetic analysis. This analysis concerns the graphical aspect of the manuscripts and the successive states of the textual content. In fact the nature of a manuscript is dual. A manuscript should be considered as a pure graphical representation or as a pure textual representation. A manuscript is a text with graphical interest [2].

As modern manuscripts reflect the writing process of the author, they may have a complicated structure and may be difficult to decipher. This is the case of Flaubert's drafts which contain several blocks of text not arranged in a linear way, and numerous editorial marks (erasures, words insertion,...) as shown in Fig. 3. For this reason, in a genetical edition, transcriptions are generally joined to the facsimile of manuscripts. A transcription allows an easier reading of the manuscript. One can distinguish two transcription types: the linear one and the diplomatic one. A linear transcription is a simple typed version of the text, which uses an adapted coding to transcribe, in a linear way, complex editorial operations of the author (deletion, insertion, substitution) sometimes located over one or several pages, even though the diplomatic one has to respect the physical aspect of the manuscript, it means the disposition of graphical elements in the page (text line, erasure, insertion,...) as shown in Fig. 1 and 2.

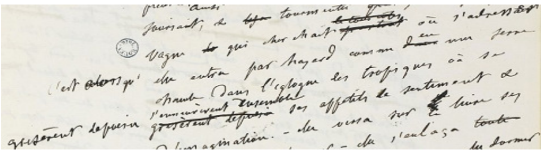
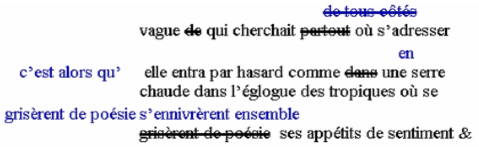


Figure 1 - A manuscript fragment



~~de tous côtés~~
vague ~~de~~ qui cherchait ~~partout~~ où s'adresser
en
c'est alors qu' elle entra par hasard comme ~~dans~~ une serre
chaude dans l'églogue des tropiques où se
grisèrent de poésie s'ennivrèrent ensemble
~~grisèrent de poésie~~ ses appétits de sentiment &

Figure 2 - Associate diplomatic transcription

Manual indexation of large corpus is a very tedious task, so a partial automation would be very helpful for researchers and scholars. Considering the lack of reliability of OCR systems when dealing with fully unconstrained handwritten documents, the automatic recognition of the textual content of the manuscripts is not conceivable because such a solution would require a fastidious checking task of recognition results. For assisted transcription purpose the aim of document analysis is more to extract automatically some regions of interest in the manuscript image in order to facilitate the extraction of textual content by human operator, and to provide an automatic coupling between parts of the image and corresponding textual transcriptions. Therefore we are more interested in locating and extracting text lines, paragraphs and editorial marks like erasures, which implies to recognize the geometrical layout of the document.

3. DOCUMENT IMAGE ANALYSIS USING MARKOVIAN MODELS

In a document image analysis process, the segmentation is one important task because it is the process that allows to locate and to extract the entities to be recognized. While the use of the handwriting recognition approaches developed so far for industrial applications may provide interesting results on historical documents, segmentation of these documents prior to the recognition is still a challenging task.

Due to the variability of ancient and handwritten documents, traditional analysis methods are not adapted and a formal description of the layout is not possible. Stochastic models are well adapted to cope with ambiguities. Markov models are usually used for sequential data segmentation and recognition. In the case of images, Markov Random Fields (MRF) are powerful stochastic models of contextual interactions for bidimensional data. In a previous work we have used MRF for document image labelling [3]. This approach has given interesting results but MRF models exhibit some limitations. Recently, Conditional Random Fields (CRF) have been proposed in order to avoid the limitations of the generative models.

The CRF were initially introduced in the field of information extraction by Lafferty and others [4] for part-of-speech tagging and syntactical analysis. Up to now they have been mainly used for sequential data modelling. Few works concerning 2D CRF models for image analysis have been proposed very recently [5][6], but to the best of our knowledge, except the work in [7] on diagram recognition and the work in [8] on handwritten word recognition, no application of the CRF to document image analysis have been proposed. The superiority of CRF models compared to MRF models has been generally reported for sequence modelling. Contrary to MRF and other generative models which define a joint probability over observation and label configurations, that requires in theory the enumeration of all possible observation configurations for the calculation of the normalization constant, discriminative models like CRF directly model the conditional probabilities of label configurations given observations. Furthermore CRF models do not require any independence hypothesis about the observations and are particularly efficient for discriminative tasks like segmentation, labelling or recognition. In this paper we propose to adapt and apply CRF to document image analysis.

3.1 Proposed model

We consider the document image analysis as a labelling problem. Each document image is considered to be produced by implicit layout rules used by the author.

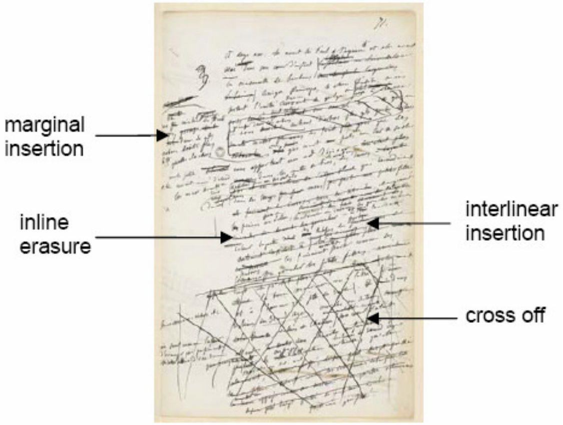


Figure 3 - One example of Flaubert's manuscript layout

For example in the case of Flaubert's manuscripts even if these rules cannot be formally justified, it is however experimentally verified by literacy experts that Flaubert's manuscripts exhibit some typical layout rules characterized by an important text body occupying two thirds of the page and containing a lot of erasures; and a marginal area with some text annotations, as can be seen on figure 3.

As there are some local interactions between these rules, Random Fields appear to be adapted to model the layout of a document. The image is associated with a rectangular grid G of size $n \times m$. Each image site s is associated with a cell on the grid defined by its coordinates over G and is denoted $g(i, j), 1 \le i \le n \quad 1 \le j \le m$. The set of sites is denoted $S = \{s\}$. Following the stochastic framework of Hidden Random Fields, the image gives access to a set of observations on each site of the grid G denoted by $Y = \{y(i, j), \quad 1 \le i \le n \quad 1 \le j \le m\}$.

Furthermore, considering that each state X_s of the field X is associated with a label l corresponding to a particular layout rule or class pattern, the problem of layout extraction in the image can be formulated as that of finding the most probable label configuration of the field X among all the possible labelling E that can be associated with the image, i.e. finding:

$$\hat{X} = \arg \max_{X \in E} (P(X/Y))$$

Whereas the MRF model gives access to the posterior probability indirectly using the Bayes rule decomposition $P(X/Y) \propto P(Y/X)P(X)$, a CRF model does not use this decomposition and therefore provides a direct formulation of the discriminative task i.e. discrimination between the labels. The general form of CRF model is given by the following formula:

$$P(X = x/Y = y) = \frac{1}{Z} \prod_{s \in S} \exp\left(\sum_k \lambda_k f_k(x, y, s) \right)$$

A CRF model is defined as the product on a set of sites of the exponential of a linear combination of k functions called feature functions, depending on the observations y, the label configuration x and the current site s. Z is a normalization factor traditionally called the partition function.

By considering the negative logarithm we can introduce the global conditional energy:

$$\begin{aligned} U(X = x/Y = y) &= -\log(P(X = x|Y = y)) \\ &= -\left(\sum_{s \in S} U_s(X_s = x_s / X_{S-\{s\}} = x_{S-\{s\}}, Y = y) \right) - \log Z \end{aligned}$$

where $X_{S-\{s\}}$ is the configuration on the remainder of the label field X except in s, and $U_s(X_s = x_s / X_{S-\{s\}} = x_{S-\{s\}}, Y = y)$ is the local conditional energy of the label x_s at the site s given the label configuration on the remainder of the label field, and the observation. This local energy (or local potential) U_s is defined as a linear combination of feature functions:

$$U_s(X_s = x_s / X_{S-\{s\}} = x_{S-\{s\}}, Y = y) = \sum_k \lambda_k f_k(x, y, s)$$

Where k stands for discriminative components taken into account in the model, and the λ_k are the weights associated with these components.

The main advantage of CRF models as opposed to MRF models is that they do not decompose the posterior probability into a data model and a prior model. In the context of MRF models these two sub-models are known to be difficult to estimate. Furthermore generative models such as Gaussian mixtures are known to be limited to using low dimensional observations vectors. CRF models do not suffer from these drawbacks.

The modelling and the solving of problems using conditional random fields require one to define the feature functions and to choose a parameter learning method and an inference method. We explain now the model we propose and our choices for these points.

3.2 Feature functions

Like Kumar *et al.* [5], and more recently He *et al.* [6], we have chosen to model the feature functions $f(x, y, s)$ by discriminative classifiers. We use Multilayer Perceptron (MLP) for this task because they are fast in decision and provide good generalization properties even in high dimensional spaces. Our CRF model can be seen as a network of interdependent classifiers taking their decision using image features as well as contextual information by incorporating the decisions of the neighbouring classifiers.

We have considered three levels of analysis ($k = 3$) and we have therefore defined three feature functions: a local feature function f_L, a contextual one f_C and a global one f_G. The f_L function takes only into account the features extracted from the observations y on a local window of analysis. The f_C feature function takes into account contextual information i.e. the label probabilities over a neighbourhood defined by a window of analysis. The f_G feature function allows integrating more contextual information by taking into account the label configuration at a coarser resolution. The local conditional energy $U_s(X_s / X_{S-\{s\}}, Y)$ at each site s is defined by a combination of the three feature functions f_L, f_C and f_G (see Figure 4):

$$P(X_s / X_{S-\{s\}}, Y) = h(f_L, f_C, f_G)$$

where h is a combination function (typically an MLP) of the three information sources.

This formulation combines a local discriminative model and two contextual discriminative models that allow capturing both the information issued from the observation field and the information issued from the label field X in a limited neighbourhood as well as in a coarser neighbourhood. This model makes it possible to capture a rich context and thus allows a good regularization and homogenization of the label field X, while taking into account the observed information. Furthermore by considering a discriminative framework, it is possible to relax the conditional independence hypothesis of the observations. Then we can take into account correlated features on a wide neighbourhood.

We now describe these three sources of information.

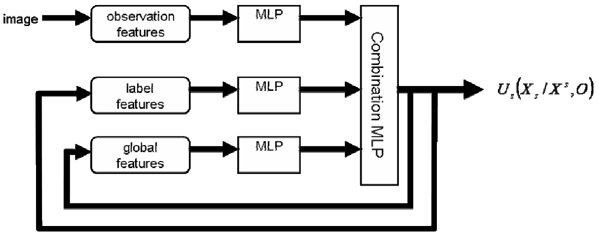


Figure 4 - Combination of the local, contextual and global information

Local features

The local feature function takes only into account features extracted on the observed image, at a given site (local image features). This function models the data association that is the adequacy between the label associated with a given site and the local observation at this site s. We take into account the same feature sets as those we used with the markov random field model we have presented in [3], i.e. multi-resolution pixel density features and site relative position. These features are extracted on each site and form a feature vector which feeds the input of the MLP which models the local feature function. The scores returned by the MLP are the values of the feature function for the different possible labels $l_i \in L = \{l_1,...,l_q\}$ which can be associated with the current site s.

Contextual features

The contextual feature function takes only into account the local conditional energies $U(X_s = l_i, i = 1,...,q / X_N, Y)$ on the label field X in a neighbourhood N around the current site. This neighbourhood is determined by defining a sliding window the size of which depends on the quantity of contextual information we wish to integrate. For example, using a window of size 3×3 and considering a label set of size $q = 3$ we can define 27 conditional energies, that is a vector of 27 contextual features applied as input of the contextual MLP.

Global feature function

An analysis of the label field X at a more global level is taken into account by a third feature function called global feature function. This global analysis is carried out using a third MLP. This classifier estimates the posterior probabilities $P(X_s = l_i, \forall l_i \in L / \mathfrak{F}_G(X))$ of associating the label l_i to the current site s given a set $\mathfrak{F}_G$ of global statistical features extracted on the global label configuration over a larger neighbourhood than that taken into account by the contextual feature function. The global classifier is also a contextual classifier that takes into account the label configuration at a coarse resolution. At this resolution, the label field is divided into several zones by superposing a grid H larger than the initial grid G. Each cell of this grid gives access to a set of sites. Statistical parameters are computed on these cells. We construct the co-occurrence matrix of the labels on each cell for different orientations. More precisely, four co-occurrence matrices are calculated for the orientations 0°, 45°, 90° and 135°. From these four co-occurrence matrices, five Haralick parameters are computed. Here the originality of our approach is that we determine these features not directly on the image, but on the label configuration.

Combination of the information sources

The use of a Multilayer Perceptron as a combination function of the different information sources is quite natural. Indeed the theoretical and mathematical definition of a MLP is rather close to that of a conditional random field since it acts as a non-linear combination of features. The values of the output of the three feature functions for the different possible labels, feed the input of one single MLP. If the label set contains q labels, the dimension of the feature vector applied at the input of the combining MLP will be $3q$. Using this type of combination the three information sources are combined in parallel.

3.3 Parameter learning

Learning the parameters of the model consists in training the four MLP. We use a supervised approach considering we have complete manually labelled data as it is the case for the CRF framework: for each image of the training database, we have the corresponding ground truth labelling. In this study, the labelling has been entered manually using a simple image editor and using a particular lookup table so as to associate a particular label to each colour. All MLPs are trained using the back propagation algorithm. The local MLP is trained first considering only the features extracted from the image. The output of the MLP is used to estimate the data association conditional probabilities $P(X_s / Y_s)$ at each site s of the image. Then these conditional probabilities are used as input features for the training of the contextual and global MLPs. All the MLPs we use have one hidden layer, with a number of neurons determined empirically according to the following general formula:

$$number\ of\ neurons = \frac{number\ of\ inputs + number\ of\ outputs}{2}$$

For each MLP the number of inputs corresponds to the number of features taken into account, and the number of outputs corresponds to the number of labels.

3.4 Model inference

The inference consists in minimizing the global posterior energy $U(X = x / Y = y)$. The techniques used for conditional random field model inference are similar to those proposed for Markov random field inference. In the 2D case, the model has a general graph structure. As a consequence, there is no exact inference method for such structure, and only a sub-optimal solution is guaranteed. As for markov random field models, we have chosen to use ICM (Iterated Conditional Modes) and HCF (Highest Confidence First) algorithm for the inference because they are known to be fast and efficient. The principle of the inference is the following. We proceed to a first labelling using only the local classifier and the intrinsic image features. During this first process, the contextual and global information about the labelling on the neighbouring sites are not taken into account. This process allows one to initialize the label field and to compute at each site the values of the local feature function. These features are then used as inputs for the contextual and global classifiers. For the next iterations the contextual and global feature functions are also taken into account to evaluate the local potential function (log of the conditional probability) at each site of the image. The inference then consists in visiting all the sites and to evaluate for each of them the score of the potential function for each possible label l_i of the set L, by combining the outputs of the local classifier and the outputs of the contextual and global classifiers. This score can be seen as the probability of assigning the label l_i to the site s given the local observations and the probabilities of the labels on the vicinity. The label providing the highest score is assigned to the current site, but all the others conditional probabilities are memorized. These probabilities are iteratively updated during the inference of the label field. This updating process is repeated until convergence of the label configuration.

4. EXPERIMENTS AND RESULTS

For the experiments we have used a dataset of 69 images of Gustave Flaubert's manuscripts. We consider a labelling task at a block level which consists in detecting large area of interest. A model with six states is defined for this task, and the parameters are learned on manually labelled images. The states are "text body", "text block", "page number", "margin", "header" and "footer". The label "text block" stands for the textual parts located anywhere in the document except in the body. The size of the grid is fixed empirically to obtain a good compromise between complexity reduction and quality of the labelling for the considered task. We choose a size of 50*50 pixels which corresponds roughly to the width of the inter-word spaces and to the height of the ascending or descending letters. As for the size of the context is concerned, we use 5*5 window.

Tab. 1 compares the average pixel labelling rate (ALR) provided by our CRF model and the average labelling rate obtained when using our previous MRF model described in [3] and local generative (Gaussian mixtures) and discriminative (MLP) classifiers, using the following formula:

$$ALR = \frac{\sum_{i=0}^{q-1}\left(\frac{number\ \ of\ \ sites\ \ correctly\ \ labeledl_{\ i}}{total\ \ number\ \ of\ \ sites\ \ labeled\ \ l_i}\right)}{q}$$

$$where\quad q\ \ is\ the\ number\ \ of\ labels$$

These results show that discriminative CRF models outperform the traditional generative MRF models considering a general image labelling task. We can see on Figure 5 an example of labelling result.

	Gaussian mixture	local MLP	MRF	CRF
ALR (%)	83.70	87.50	90.56	94.16

Tab. 1 - Comparison of labelling rates obtained with different models

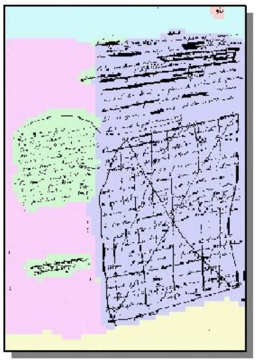


Figure 5 - Example of labelling result on a Flaubert's manuscript

5. CONCLUSION AND FUTURE WORK

In this paper we have proposed a Conditional Random Field model for 2D data labelling, in particular for document image segmentation and we have shown that document image analysis and machine learning procedures can be very useful to help the scholars to index patrimonial documents and to produce numerical version of these sources.

One of the main advantages of the approach we propose is that it relies on machine learning procedures, so no manual parameter setting is necessary. This allows an easy adaptation to different types of documents and different analysis tasks. The results we have obtained on Flaubert's manuscripts show that the proposed model provides better results than MRF generative models. These results are similar to those presented in other recent work on conditional random fields. Future work concerns the integration of more intrinsic and contextual features in the model, the replacement of MLP by logistic classifiers faster to train, and the definition of hierarchical CRF models for document image analysis.

REFERENCES

[1] H. Baird, "Digital Libraries and Document Image Analysis", 7th International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, Scotland, pp. 2-14, 2003.

[2] J. André, J.D. Fekete, H. Richy. Mixed text/image processing of old documents. Congrès GuTenberg, pp. 75-85, 1995.

[3] S. Nicolas, T. Paquet and L. Heutte, "A markovian approach for handwritten document segmentation", 18th IEEE International Conference on Pattern Recognition, (ICPR 2006), vol 3, pp. 292-295, 2006.

[4] J. Lafferty, F. Pereira and A. McCallum, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data", International Conference on Machine Learning (ICML'01), pp. 282-289, 2001.

[5] S. Kumar and M. Hebert, “Discriminative random fields: A discriminative framework for contextual interaction in classification”, 9th IEEE International Conference on Computer Vision (ICCV'03), vol. 2, pp. 1150-1159, 2003.

[6] X. He, R.S. Zemel and M.A. Carreira-Perpinan, "Multiscale Conditional Random Fields for Image Labelling", IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, pp. 695-702, 2004.

[7] M. Szummer and Y. Qi, "Contextual Recognition of Hand-drawn Diagrams with Conditional Random Fields", 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), Tokyo, Japan, pp. 32-37, 2004.

[8] S. Feng, R. Manmatha and A. McCallum, "Exploring the Use of Conditional Random Field Models and HMMs for Historical Handwritten Document Recognition", 2nd International Conference on Document Image Analysis for Libraries (DIAL'06), Lyon, France, pp. 30-37, 2006.