2011 International Conference on Document Analysis and Recognition

Continuous CRF with multi-scale quantization feature functions Application to structure extraction in old newspaper

David Hebert, Thierry Paquet and Stephane Nicolas
Laboratoire LITIS EA 4108 Universite de Rouen, France
Email: FirstName.Name@univ-rouen.fr

***Abstract*—We introduce quantization feature functions to represent continuous or large range discrete data into the symbolic CRF data representation. We show that doing this convertion in a simple way allows the CRF to automaticaly select discriminative features to achieve best performance. This system is evaluated on a segmentation task of degraded newspapers archives. The results obtained show the ability of the CRF model to deal with numerical features similarly as for symbolic representation thanks to the use of quantization feature functions. The segmentation task is achieved by the definition of a horizontal CRF model dedicated to pixel labelling.**

Keywords- L-CRF, quantization feature functions, document images labelling

I. INTRODUCTION

Automatic structure extraction of document images is the process that allows accessing document content for recognition of textual entities by OCR. At a higher level, automatic structure analysis is the process that gives access to the logical organization of the document so as to ease the information retrieval process that can then operate on various descriptors of the document e.g. Titles, Sub-Titles, Chapter, Articles, Paragraphs, Captions, etc... Generally, these two processes of document structure extraction (known as physical and logical layout analysis) operate separately and sequentially, one after the other. This is justified by the fact that most of the time physical segmentation of document images can be performed without the need for any additional knowledge, whereas logical layout extraction is generally performed thanks to the use of a document model (e.g. a style sheet) that express the relations between physical and logical entities of the two representations of the document. However, it is now well known that difficult segmentation tasks must incorporate a recognition stage so as to improve their performance. This is particularly true in the case of cursive handwriting recognition [9] [11], for which significant improvement have been observed when carrying both segmentation and recognition in conjunction. Similar schemes have been proposed already in image analysis for the detection of objects in natural scenes [5]. Some attempts have also been proposed for handwritten document segmentation [4].

In this paper we explore the use of Conditional Random Fields to achieve document image segmentation into functional entities such as titles and sub-titles, graphical separators, text lines, columns. Experiments are carried out on a task of article detection in old newspapers, for which the physical layout is changing over a range of nearly 180 years. Conditional Random Fields (CRF) introduced in 2001 by Lafferty et al. [1] have opened a new way for sequence analysis. Previously, Hidden Markov Models (HMM) were the preferred method for this type of task and have been applied in various ways for many tasks such as detection, segmentation, classification... Since then, CRF have also been used for image segmentation, as this is the case for HMM.

In its original form a CRF is a stochastic process that models the dependencies between a set of discrete observations made within a discrete sequence (originally a word sequence) and a set of labels that can be associated to these observations (originally Part Of Speech tags). Compared to HMM models, CRF do not rely on strong independence assumptions (which cannot be satisfied experimentally) between labels and neighboring observations. Another advantage of CRF compared to HMM is that they do not account for local conditional probabilities, so being free of biased estimations of these quantities when too few labeled data are available. Instead, potentials weights account for positive as well as negative contributions of the observations to the labels.

The application of CRF to image segmentation requires some adaptation. This adaptation is most of the time a preprocessing step which is dedicated to providing the CRF with discrete observations extracted from the whole image. He et al. [5] design a multi-layer CRF based approach for natural image segmentation and use the outputs of a neural network to feed the CRF. In [7], SVMs are used to model the relationship between a pixel and its label. CRF have also been used in several works for document structure extraction. Nicolas et al. [4] use a 2D-CRF based approach but also combined with MLPs. Another example can be found in [3] were the CRF model is explicitly defined as a second stage after a pixel classification stage.

In opposition to most of these studies which introduce a discretization stage prior to the CRF, we propose to use some quantization feature functions that are directly optimized

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.105

during the training phase, thus providing an efficient one shot training algorithm for continuous CRF. The rest of this paper is organized as follows. In section II we briefly recall the original CRF model. Section III describes the proposed quantization feature functions. An experimentation setup dedicated to old newspaper image segmentation is presented in section IV. Experimental results are given in section V. Finally, we discuss these results and conclude in section VI.

II. Linear chain-CRF

L-CRF, for Linear chain Conditional Random Fields have been defined for the first time by Lafferty et al. in 2001 [1]. They have introduced a discriminative model that does not assume independence between labels and neighboring observations, as an answer to the label bias problem in MEMM (Maximum Entropy Markov Model) or HMM for language analysis.

A. CRF model

Following [6] we recall the main properties of CRF.
For the remainder of the paper, we define some notations. $X = x_1, x_2, ..., x_T$ will be a sequence of T discrete observations. $Y = y_1, y_2, ..., y_T$ will be the sequence of T discrete labels attached to X. L is the set of all possible labels (all possible values for y_t) and O the observations (all possible values for x_t, e.g. the discrete lexicon in case of text). A l-CRF is defined as

$$p(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_{t}^{T}\sum_{k}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right)$$

As can be seen, the probability of having a particular label sequence associated to the observation sequence is derived from a linear combination of weighted binary functions over the observation sequence. The weights λ_k are model parameters and can be interpreted as the relevance of the binary functions f_k. $f_k(y_{t-1}, y_t, x, t)$ is the general notation of binary functions named feature functions, which accounts for the occurrence of a particular combination of observation(s) and label(s). For example,

$$f_k(y = l_i, x = o_j) = \begin{cases} 1 & if \quad y_t = l_i \ \texttt{and} \ x_t = o_j \\ 0 & else \end{cases}$$

Feature functions are user-specified. They reflect the user knowledge on the application domain. By leaving out the independence assumption between labels and neighboring observations, CRF models give rise to the possibility to define many contextual features, which is not possible with traditional HMM. Contextual features are defined by pattern templates. A template is a pattern that defines a specific contextual combination of observations and labels. For example, the template $f(y_t, x_t)$ accounts for all possible couple of observations with their label at each position t in the sequence. This single template can generate up to $Card(O) \times Card(L)$ binary features.

B. Training a l-CRF

Training a CRF is the optimization of the parameters λ_k taking account of the ground truth labels associated to the observed data. The aim is to maximize the likelihood on the training data set composed of N couples of observed data together with their labeled sequence:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(y^{(i)}/x^{(i)}) \ \texttt{with} \ \theta = \{\lambda_k\}$$

Optimization of $\ell(\theta)$ is a convex optimization problem that has only one global optimum, thus leaving place for using multiple optimization algorithms. However, the number of features grows rapidly with the size of the observation set and the number of feature templates. Therefore, even simple problems may exhibit a huge amount of parameters to optimize. This practically reduces the possible algorithms that can be used. Since 2003 LBFGS algorithm is the most commonly used to train a CRF model [2], see [8] for more details on this algorithm.

C. Decoding process

Decoding is the process of finding the best label sequence that can be associated to an observation sequence. This consists in finding the label sequence $y*$ that maximizes $p(y*/x)$. The best label sequence is computed by a Viterbi-like algorithm.

$$y* = argmax_y(p(y/x))$$

III. Defining quantization feature functions

In this paragraph we address the problem of adapting the CRF formalism to continuous observations. Indeed, as apposed to text analysis for which observations are made of words, image features are generally numerical continuous real values. To overcome this difficulty most of the studies devoted to image analysis so far have introduced a pre-processing step which consists of a classification stage. The output labels of the classifiers are then fed to the CRF as input discrete observations. To use the CRF as an entire classifier system, we propose to use quantization feature functions. Let us define a linear quantization function with quantifier q that quantizes the continuous observation o as follows:

$$\begin{aligned} Q(o, q): \quad & 0 \longmapsto X \\ & o \longmapsto x = round(\tfrac{o}{q}) \end{aligned}$$

If we assume that o is ranging within the interval $[o_{min}, o_{max}]$ then the quantization function can take only N discrete values, with $N = (o_{max} - o_{min})/q$

As we see, the parameter q cannot be chosen without the knowledge of the range of the continuous observation features. Furthermore, most physical features have a distribution over the range of values that is not constant. For example, the smaller values could be more discriminative than the

larger ones. In this case, having too large values of q may remove some discriminative information; whereas too small values of q may spread the discriminative information over too many features.

To avoid this difficult choice of q we propose to introduce multiple quantization functions. Each of them will define a set of binary features. Let $q_1, q_2, , q_N$ be a set of quantifiers, each defining a quantization function $Q_i(o) = Q(o, q_i)$, then by choosing a dyadic law of quantifiers as follows $q_i = 2 * q_{i-1} = q_1 * 2^{i-1}$, it is possible to build a multi-scale quantization scheme with the ability to keep most of the original information contained in the continuous features without any assumption about the distribution of these features. Finally, we expect the CRF to select the discriminative quantifiers by weighting them accordingly.

Now, the general model of a discrete CRF can be rewritten using the set of N quantization function as follows. This proposed model is evaluated on a document archive segmentation task.

$$p(Y/X) = \frac{1}{Z(X)} \exp\left(\sum_{t}^{T}\sum_{k}^{K} \lambda_k f_k(y_{t-1}, y_t, Q_1(o), ..., Q_N(o))\right)$$

IV. Experiments

CRFs are a powerful tool to label data and structures using contextual information. The CRF model we propose is particularly adapted and dedicated to image analysis tasks, for which continuous observations are necessary. In the context of a project concerning the indexing of old newspaper archive, we are interested in the analysis of newspaper structures, in order to allow an automatic extraction of the articles in newspaper pages like depicted in fig. 2, hence facilitating the indexing task. Newspaper archives are an important source of information for historians and people interested in common history and genealogy. They reflect the evolution of the social context during a large period of time. An efficient way to share this information is to put archives on the Internet so as to make them readable by people without the risk of degrading the original fragile paper documents.

Based only on physical attributes computed from the document image, our task is to segment and identify text lines, titles, horizontal and vertical separators of articles, and differentiate these elements from noisy parts due to digitization and paper artefacts.

We have chosen to use very simple image features relevant to the manhattan X-Y structure of newspapers. These image features we consider are run length of same colour. These features are relevant for horizontally structured documents and easy to use. Each pixel is characterized by its horizontal and vertical run lengths. These two numerical values are discrete but with large range. The average size of theses images are 1200×1550, therefore horizontal runs range within $[1, 1200]$, while vertical runs range within $[1, 1550]$. Due to the natural horizontal orientation of most of the relevant information in the document image, we have come to define one horizontal CRF as a model of a line of pixels. This means that contextual information between labels is only introduced along the horizontal direction, whereas horizontal and vertical features are attached to each local label by using first order contextual features. By doing so, we ensure the existence of an optimal and fast, one dimensional, decoding procedure very close to the Viterbi algorithm. Two dimensional extension of the method would require the use of a sub-optimal decoding algorithm such as loopy belief propagation, or graph-cut.

We now give more details on the contextual features used in this experiment. Five horizontal templates are introduced as depicted in fig. 1 (F_1 to F_5). They account for the horizontal dependencies of the current label and each of the horizontal runs over a window of five neighbours. Five vertical templates (F_6 to F_{10}) are defined to account for the vertical dependencies between the current label and the vertical runs over an horizontal window of five neighbours.

$$\begin{array}{ll} F_1(y_t, o^h_{t-2}) & F_6(y_t, o^v_{t-2}) \\ F_2(y_t, o^h_{t-1}) & F_7(y_t, o^v_{t-1}) \\ F_3(y_t, o^h_{t}) & F_8(y_t, o^v_{t}) \\ F_4(y_t, o^h_{t+1}) & F_9(y_t, o^v_{t+1}) \\ F_5(y_t, o^h_{t+2}) & F_{10}(y_t, o^v_{t+2}) \end{array}$$

Figure 1. Horizontal (F_1 to F_5) and vertical (F_6 to F_{10}) templates.

These templates defined on the continuous observed features now give quantized feature functions (Q-functions) as follows. As small runs provide discriminative information on textual information, it is important to have a small initial value of quantifier. For our experiments $q_1 = 2$ has been chosen. Similarly, large runs are not representative of any specific label. Therefore, in order to limit the number of quantized feature functions we limit the quantifier values within the range $[2, 512]$. This provides a multi-level quantization process with 9 levels. Therefore each first order horizontal template F_i generates 9 first order discrete templates f_i^j as follows:

$$\begin{array}{c} f_1^1(y_t, o^h_{t-2}) = F_1(y_t, Q_1(o^h_{t-2})) \\ f_1^2(y_t, o^h_{t-2}) = F_1(y_t, Q_2(o^h_{t-2})) \\ ... \end{array}$$

These discrete templates generate binary feature functions as defined in the original CRF definition. In addition, one first order transition label function (configuration of two consecutive labels) is also intoduced. Finally, the model is trained using the l-BFGS algorithm on a database of labelled images. The decoding process runs a Viterbi like algorithm to search for the optimal sequence of labels for each line of the image.

V. RESULTS

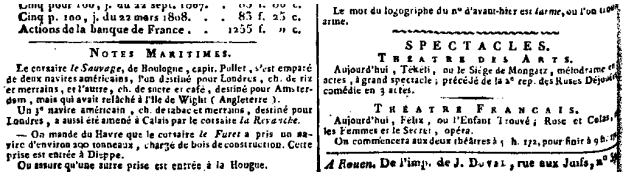
The evaluation of the segmentation results is a difficult process, depending on the segmentation task. The difficulty is to quantify the quality improvement at a pixel level. Numerous papers show typical image results to illustrate their system performances. However, segmentation competitions have to quantify results to allow comparisions of the methods. Some tasks can be evaluated computing precision/recall statistics on pixels classification but other tasks need a weighted approach which does not give the same relevance for all kind of errors. Indeed, according to post-processings of segmentation results, some errors can be critical. For example, text line segmentation are usefull before any OCR operation but the OCR performances are linked to the quality of line detection. For our experiments, we decided to quantify quality results using the Jaccard similarity coefficient which is a ratio between the number of correctly labelled pixels and the sum of the number of correctly labelled, wrong labelled and missing pixels, according to the ground truth.This coefficient is computed on each label.The CRF model is trained on 11 images representing 16978 sequences and evaluated on 23 images. All these 34 images are completely labelled at a pixel level. These images are decomposed in 10 labels to characterize all physical entities of documents. Note that some labels are combined to define higher labelling level. For example, a text line is the combination of characters, inter-characters and inter-words labels. All these labels and their combinations are:

- Vertical separator
- Horizontal separator
- Titles (composed of "title characters", "title inter-characters" and "title inter-words")
- Text lines (composed of "characters", "inter-characters" and "inter-words")
- Noise
- Background

Jaccard coefficient is computed on enclosing polygons of titles, text lines, separators and noisy parts.
To illustrate the improvement provided by our quantization feature functions, we compare the results obtained with two CRF models, one trained with quantization feature functions and the second one without these features, but using only original feature functions on large range discrete physical observations (e.g. integer run length).

A. Training of CRF

Tab. I shows the number of feature functions really generated during the training phase, using the templates described in section IV and the number of all possible features that can be generated in theory. We show that in practice, only a fraction of all possible feature functions


Cinq pour 100, j. du 22 sept. 1807. . . 83 f. 00 c.
Cinq p. 100, j. du 22 mars 1808. . . 83 f. 25 c.
Actions de la banque de France . . 1255 f. » c.

NOTES MARITIMES.

Le corsaire le Sauvage, de Boulogne, capit. Pollet, s'est emparé de deux navires américains, l'un destiné pour Londres, ch. de riz et merrains, et l'autre, ch. de sucre et café, destiné pour Amsterdam, mais qui avait relâché à l'île de Wight (Angleterre).

Un 3e navire américain, ch. de tabac et merrains, destiné pour Londres, a aussi été amené à Calais par le corsaire la Revanche.

— On mande du Havre que le corsaire le Furet a pris un navire d'environ 300 tonneaux, chargé de bois de construction. Cette prise est entrée à Dieppe.

On assure qu'une autre prise est entrée à la Hougue.

Le mot du logogriphe du n° d'avant-hier est *larme*, où l'on trouve arme.

SPECTACLES.

THÉATRE DES ARTS.

Aujourd'hui, Tekeli, ou le Siège de Mongatz, mélodrame en 3 actes, à grand spectacle; précédé de la 1re rep. des Muses Déjeunes, comédie en 3 actes.

THÉATRE FRANÇAIS.

Aujourd'hui, Félix, ou l'Enfant Trouvé; Rose et Colas, les Femmes et le Secret, opéra.

On commencera aux deux théâtres à 5 h. 1/2, pour finir à 9 h.

A Rouen, De l'imp. de J. Duval, rue aux Juifs, n° 8.

Figure 2. Focus on a document structure, with horizontal and vertical separators, titles, text and noisy parts

is used. Only the feature functions which appear at least once in the training database are generated. However, 51% of parameters λ_k have a value between 1.10^{-3} and -1.10^{-3} meaning that at least 51% of these $84,760$ feature functions do not bring a significant contribution in opposition to most significant values close to 4 and -4. This number of low weighted feature functions illustrates the selection process of the most discriminative feature functions performed during the CRF training phase, over the amount of quantized data. This figure illustrates also the processing time of the training performed on a 2.4 GHz processor.

Table I
NUMBER OF FEATURE FUNCTIONS AND TRAINING TIME

	Without Q-functions	With Q-functions
Possible feature functions	275200	102000
Effective feature functions	90520	84850
Process Time	$9382s$	$13976s$

B. Segmentation results

The Jaccard similarity coefficient is computed using the following formula: $\frac{TP}{(TP+FP+FN)}$. TP means True Positive, FP is for False Positive and FN means False Negative. This coefficient is computed for each label for the two CRF models. These results are presented in tab. II and show the improvement provided by quantization feature functions. The number of correctly segmented entities also increases using the quantization feature functions. Note that even if horizontal separators are almost correctly detected without Q-function, the quality of this segmentation is improved with the quantization feature functions, giving separators without artifacts.

As mentioned previously, a small error can have a great impact on the following post-processing stages. For our task, missing an horizontal separator means missing the beginning of a new article but the more significant error is when the vertical separator label is confused with the characters label, meaning that the lines on each side of the vertical separator will be concatenated. A good indicator for this type of error is the confusing rate between characters and vertical separator. Quantization feature functions decrease this confusing rate from 3.15% to 1.47%, hence reducing the number of critical errors for a text line segmentation task.

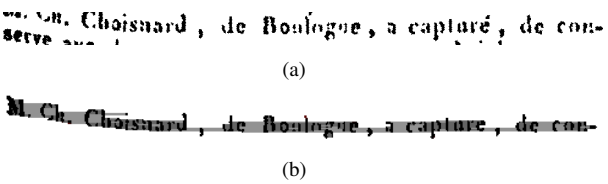
Table II
JACCARD COEFFICIENTS COMPUTED ON HORIZONTAL SEPARATORS, VERTICAL SEPARATORS, TITLES, TEXT LINES AND NOISY PARTS FOR THE TWO CRF MODELS AND NUMBER OF TITLES AND HORIZONTAL SEPARATOR CORRECTLY SEGMENTED

Jaccard coefficients		
Label	Without Q-functions	With Q-functions
H sep	0.8223	0.8919 (+0.07)
V sep	0.9136	0.9641 (+0.05)
Title	0.7503	0.8317 (+0.08)
Text Lines	0.9789	0.986 (+0.007)
Noise	0.2876	0.4078 (+0.12)
Number of entities		
H sep	204/212(96.23%)	211/212(99.53%)
Title	106/127(83.46%)	120/127(94.49%)

Decoding time is quite fast and depends on the number of feature functions defined. For 84850 feature functions, the model needs 2.7 seconds on average per image. One million features need 5.2 seconds to decode one image.

C. Text line extraction task

Most methods used for text line extraction works without any training stage and can achieve good results on digital documents but have some difficulties to work on degraded scanned documents. Indeed, many of our documents present text lines which are curved at their beginning or at their end as depicted in fig. 2. The main segmentation methods such as the RAST method [10] integrated in the OCR system Ocropus, are not able to deal correctly with this type of curved or fluctuating text lines. A typical result is shown on fig. 3(a). These errors are critical if an OCR is applied on these lines. Our training method needs some labeled images but can decompose a text line as a sequence of words, it is thus able to adapt the segmentation results for curved lines (fig. 3(b)).



(a)

(b)

Figure 3. Typical result on line with curved beginning obtained by the RAST method (a) and the CRF (b)

VI. CONCLUSIONS

In this paper, we have introduced multi-scale quantization feature functions in a discrete CRF, thus avoiding the use of a pre-classification stage on continuous features. The CRF model is used as a pixel line model and experimented on document image segmentation task.
We show that using these kind of feature functions allows better segmentation on degraded newspaper archives and can also label entities in documents. We also evaluated this method on a text line extraction task and obtained better results for curved/degraded text lines.
The system presented can be easily adapted to other segmentation tasks by using other physical features.

REFERENCES

[1] J. Lafferty, A. McCallum and F. Pereira, *Conditional random fields: probabilistic models for segmenting and labeling sequence data* , Proc. 18th International Conf. on Machine Learning. San Francisco, 2001.

[2] F. Sha and F. Pereira, *Shallow parsing with conditional random fields* , Proc. NAACL '03. Stroudsburg, USA, 2003.

[3] S. Chaudhury , M. Jindal1 and S. Dutta Roy *Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field* , Proc. PReMI '09. Berlin, Germany, 2009.

[4] S. Nicolas, J. Dardenne, T. Paquet and L. Heutte *Document image segmentation using a 2D conditional random field model* , Proc. ICDAR '07. Curitiba, Brazil, 2007.

[5] X. He, R. S. Zemel and M. A. Carreira-Perpinan *Multiscale conditional random fields for image labeling* , Proc. In CVPR. Washington DC, USA, 2004.

[6] C. Sutton and A. Mccallum *Introduction to conditional random fields for relational learning* , In "Introduction to statistical relational learning", chapter 1. 2006.

[7] C.-H. Lee, S. Wang, A. Murtha, M. R. G. Brown and R. Greiner *Segmenting brain tumors using pseudoconditional random fields* , Proc. MICCAI '08. New York City, USA, 2008.

[8] J. Nocedal and S. J. Wright *Large scale quasi-Newton and partially separable optimization* , in "Numerical Optimization", Chapter 9. Springer, 2006.

[9] T. M. T. Do and T. Artires *Conditional random fields for online handwriting recognition*, IWFHR. La Baule, France, 2006.

[10] T. M. Breuel *High performance document layout analysis*, in Symposium on Document Image Understanding Technology, Greenbelt. USA, 2003.

[11] S. Feng, R. Manmatha, and A. McCallum *Exploring the use of conditional random field models and HMMs for historical handwritten document recognition*, DIAL. Washington, DC, USA, 2006.