Pattern Recognition 47 (2014) 959–968

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Contextual text/non-text stroke classification in online handwritten notes with conditional random fields

Adrien Delaye *, Cheng-Lin Liu

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Available online 2 May 2013

Keywords:
Online handwritten documents
Structured prediction
Stroke classification
Conditional random fields
Text/non-text separation

ABSTRACT

Analysing online handwritten notes is a challenging problem because of the content heterogeneity and the lack of prior knowledge, as users are free to compose documents that mix text, drawings, tables or diagrams. The task of separating text from non-text strokes is of crucial importance towards automated interpretation and indexing of these documents, but solving this problem requires a careful modelling of contextual information, such as the spatial and temporal relationships between strokes. In this work, we present a comprehensive study of contextual information modelling for text/non-text stroke classification in online handwritten documents. Formulating the problem with a conditional random field permits to integrate and combine multiple sources of context, such as several types of spatial and temporal interactions. Experimental results on a publicly available database of freely hand-drawn documents demonstrate the superiority of our approach and the benefit of contextual information combination for solving text/non-text classification.

1. Introduction

Automatic interpretation of free form online handwritten documents is considered as a very challenging task due to the high diversity of contents and the lack of prior knowledge available. By fully exploiting possibilities of pen-based interfaces, one is able to input rich contents, from writing text lines and paragraphs to composing tables, sketching diagrams, realizing free drawings, or gesturing annotations and commands. In a realistic note-taking scenario, no constraint is enforced to the user who has the liberty to create a document without complying with any specific composition rule. As more and more of such rich, heterogeneous documents are acquired from digital pens, pen-enabled computers, smart phones, tablets and electronic whiteboards, there is a need for better analysis and recognition algorithms.

Without loss of generality, it is often assumed that ink documents have some textual content that is of special importance for their interpretation. Contrary to non-textual elements that can have highly variable properties, textual elements present regularities and robust features, such as a hierarchical organisation into words, lines and paragraphs, or a locally stable size. Moreover, textual elements usually convey semantically rich information and they can be analysed by efficient handwriting recognition engines [1], whereas it is difficult to define general analysers for non-textual elements. For these reasons, the task of accurately separating textual from non-textual strokes in an online document is regarded as a crucial step towards general free-form document interpretation [2].

In this paper, we present our advances for the text/non-text classification problem in online handwritten documents with conditional random fields. Contextual interactions between elements of the document are of crucial importance and the use of CRFs permits to model these complex dependencies under a well principled framework. Though many works have been conducted for stroke classification using various features and classifiers, and some considered the interactions between strokes, a thorough exploitation of contextual information was not reported. We conduct a comprehensive study on the exploitation of contextual information for supporting the stroke classification task. To compensate for the lack of prior knowledge, the rich information contained in the online documents is fully exploited: visual information from the spatial distribution of the strokes, and dynamic information from the temporal sequence of writing. Our results establish new state-of-the-art performance for text/non-text stroke classification in free form documents.

The first section of this paper reviews past research works about the separation of textual and non-textual content in online documents. This survey highlights the need for a global study on the exploitation of context and combination of various types of

* Corresponding author. Tel.: +86 15811394302.
E-mail addresses: adrien.delaye@nlpr.ia.ac.cn, adrien.delaye@gmail.com (A. Delaye), liucl@nlpr.ia.ac.cn (C.-L. Liu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.04.017

interactions for this task. In Section 3, we present the conditional random field model that allows a large flexibility in the exploitation of contextual knowledge. Different definitions for stroke interaction systems are then presented in Section 4, suggesting that complementary neighbourhood systems can be combined in the graphical model. The experimental section demonstrates the effectiveness of the approach and the benefit of combining different contextual sources by evaluating the proposed system on the public database IAM-OnDo.

2. Related works

Over the last decade, much effort has been devoted to the interpretation of online hand-drawn content, from handwritten text [1,3], symbols and gestures [4–6], to structured scientific notations [7], diagrams and sketches [8–10] or tables[11,12]. The problem of analysing unconstrained, free-form online documents has also been receiving an increasing attention lately [13,14], and methods have been proposed for processing content block segmentation [15], text lines segmentation [16] or for document structure retrieval [13,17,18].

An online note-taking document is a sequence of points acquired along the trajectory of the pen on the surface, where points are organised in strokes (portions of the trajectory realised without lifting the pen from the surface). Strokes offer a natural unit for partitioning a document, and it is generally assumed that a stroke can be affected a single type (either text or non-text), as users usually lift the pen from the surface when switching from text to non-text content [14]. In the literature, the task of separating textual from non-textual strokes is regarded as a central problem for document understanding, whether it is considered in isolation (as in [2,13,14,19,20]) or combined with the inter-related tasks of segmentation and structure analysis (as in [15,17]).

Several categories of approaches for text/non-text stroke classification can be distinguished according to how contextual information is exploited. In this section, we first present approaches for isolated stroke classification, without any contextual information or simply including a local context description. We then present structured prediction methods, where interactions between elements are exploited to support the stroke classification task. We distinguish methods relying on the temporal structure of the document and methods relying on the spatial structure.

2.1. Isolated stroke classification

The work of Jain et al. [13] introduces a strictly local approach for classifying online strokes as text or non-text in handwritten documents. Two features are extracted from each stroke (namely the stroke curvature and stroke length), without considering interactions with other strokes. A linear classifier predicts the type of each stroke in isolation with a high accuracy (97%) and homogeneous regions such as text blocks or tables are detected in a subsequent step. The same features applied with a support vector machine classifier for isolated stroke classification on the challenging IAM-OnDo database were reported to perform significantly lower, at 91.3% [21]. Other authors have designed entropy-based methods, assuming that higher entropy of the online pen trajectory distinguishes the writing of text elements from the drawing of polygons or geometrical shapes [22,23]. Spectral features (e.g. discrete Fourier transform coefficients) have also been employed with a linear classifier for isolated stroke classification [24]. Willems et al. [25,26] propose to use a set of eight features (including global and structural features) with a nearest-neighbour classifier for detecting *modes* in online signal. In that case, the non-textual elements are further classified into subcategories considered as different modes (deictic gestures, complex drawings, geometric shapes).

If it makes sense to classify strokes individually, this decision problem can usually not to be answered unambiguously without considering some contextual information. For example, the same stroke shaped as a small circle could be considered to be a textual stroke (representing the letter “o”) or a non-text stroke (representing a circle) depending if it is located within a line of text or if it is a part of a drawing. Peterson et al. [27] thus proposed to extract features not only from the stroke to be classified but also from surrounding strokes. It was shown that extracting local context features (such as the average size of neighbouring strokes, or the average distance to them) significantly improves the stroke prediction. The findings of several other researchers [2,14] confirm this observation.

This clearly highlights the importance of considering contextual knowledge when predicting a stroke label. Accordingly, beyond inclusion of local context features from surrounding strokes, researchers have formulated the problem of stroke classification as a structured prediction problem, i.e. by modelling jointly the labelling of strokes from a document.

2.2. Exploitation of temporal context

An obvious way of modelling interactions between strokes in an online document is to exploit the temporal information. The online nature of the data suggests handling stroke labelling as a sequence labelling problem, where each stroke is an observation to be labelled as text or non-text.

As a direct exploitation of the temporal structure, the document can be segmented into sub-sequences of strokes that are assumed to share the same type because their temporal distance is below an empirical threshold. Then the classification can be operated at the subsequence level, by exploiting a richer context [28]. The problem of defining an appropriate segmentation strategy is not always clearly addressed [29] and the assumption of consistency of labels over sub-sequences is often unverified in realistic datasets [21].

A probabilistic framework for sequence prediction was adopted by Bishop et al. [20], who designed a hidden Markov model for modelling interactions between successive strokes in the drawing sequence. The dependencies express the fact that two strokes written successively are more likely to be of the same type. Emission probabilities for each stroke are computed by a multi layer perceptron with 11 input features and the transition probabilities are estimated from training data. Labelling strokes with the HMM model significantly outperforms the independent labelling with MLP classifier. The structured model can be improved with a bipartite HMM formulation, where the gaps between strokes are considered as observations and additional hidden states are integrated for modelling transitions between text and non-text states. A shortcoming of HMMs is the assumption of independence between observations [30], which in practice prevents from considering local context for prediction of a stroke label. In other words, the advantage of modelling dependencies between labels of temporally adjacent strokes is mitigated by the limitation to a weaker description for each stroke label prediction.

More recently, Indermuhle et al. [31] presented a mode detection approach based on bidirectional long short-term memory (BLSTM) neural networks. BLSTM is a type of recurrent neural networks that have been applied successfully to sequence prediction in speech and handwriting recognition [32]. For text detection in online documents, BLSTM is applied on the document represented as a stream of feature vectors extracted from points of the sampled pen trajectory. Prediction of labels is influenced by

a bidirectional temporal context, since the data stream is fed to the network both forwards and backwards. The training step adjusts the parameters for contextual interactions through the mechanism of memory blocks for long short-term memory. At test time, a voting scheme is applied to the output point wise labelling, so as to determine labels at stroke level. The method yields very accurate predictions on the IAM-OnDo dataset. Its advantages are that it is trainable, does not rely on heuristics and requires no prior empirical segmentation. However, if the temporal context is efficiently considered in the model, other potential sources of context (such as interaction between spatially adjacent but temporally distant strokes) are omitted.

2.3. Exploitation of spatial context

The spatial distribution of strokes in an online document obviously provides precious information for the stroke classification task. In their work, Zhou and Liu [19] explicitly integrate spatial interactions in a Markov random field model for stroke classification in Japanese hand-drawn documents. The observation potential and interaction potential functions of the MRF model are computed from the outputs of support vector machine classifiers trained independently. Experiments demonstrate the superiority of MRF model over a hidden Markov model, thus attesting the importance of interactions between spatially adjacent strokes. As a generative model, the MRF model however does not permit sufficient exploitation of local context for prediction of stroke labels, due to the underlying independence assumption of the observations (as in the HMM formulation).

2.4. Towards integration of multiple contexts

Combining temporal and spatial information for categorization of strokes in online documents has also been proposed. Shilman et al. make use of successive temporal and spatial grouping of strokes for segmenting the document into homogeneous blocks that can be classified as text or graphics [17]. In their work, the influence of contextual information over the stroke classification cannot be trained and several parameters have to be tuned by hand, making the robustness of the method questionable when applied to different datasets.

This short review of the literature highlights the need for a method combining efficiently different useful sources of context: integration of local context description, as well as modelling of spatial and temporal dependencies between stroke labels. We propose to exploit the conditional random field framework [30] for addressing this need. The discriminative nature of CRFs permits to model and propagate label dependencies between strokes without requiring observation independence, thus enabling the combination of overlapping and long-range features [18,30]. This increased flexibility permits to combine in a single method several benefits of methods presented in the literature, while retaining the advantages of a trainable method, without the need for empirical segmentation strategies based on hand tuned parameters.

Fig. 1 gives an overview of the proposed system for classification of strokes as text or non-text in online documents. Documents are represented as graphs of strokes that reflect their structure by modelling various types of stroke interactions. The conditional random field model involves a local stroke classifier, trained independently, and a set of weights discriminatively optimised from the training document graphs. The prediction of labels for a test document requires to construct the document graph and optimize the labelling based on the full model.

In the next section, we introduce the conditional random field formulation of the stroke classification problem and present the algorithms for performing inference and training of the model

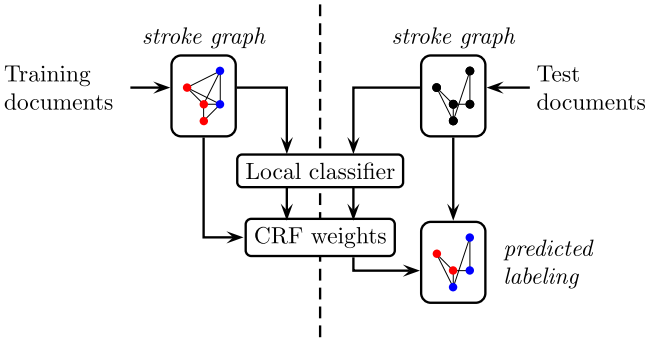

Fig. 1. System work flow. On the left is illustrated the training of the system, involving a local classifier and CRF weights trained from documents represented as stroke graphs. On the right is presented the model inference for prediction of stroke labels of a test document from its stroke graph.

parameters (weights). In Section 4, we propose a number of strategies for constructing the stroke graphs to represent the documents, by including different types of stroke interactions with different definitions of neighbourhood systems. Section 5 presents experiments and results obtained on the stroke classification task over IAM-OnDo database [21], a realistic benchmark for evaluation of free form handwritten document analysis systems.

3. Conditional random fields for online stroke classification

Conditional random fields (CRFs) were introduced by Lafferty et al. [30], as a type of probabilistic graphical models directly representing a distribution over labels conditioned on the observations as Markov random fields. Graphical models combine graph theory and probabilities, providing a framework for modelling global probability distributions based on the definition of local dependencies. The discriminative nature of CRF allows handling of arbitrarily complex interactions without assuming independence of the observations, making it an appropriate framework to overcome the limitations of existing methods for stroke labelling as a structured prediction problem. We introduce here the formulation of the text/non-text stroke labelling problem as a CRF and discuss the choice of algorithms for inference and parameter training.

3.1. CRF formulation

Let x be the observed data over a set of *sites* S: $x=\{x_i\}, i\in S$, and y be a set of random variables over the sites, taking values in the set of labels $\mathcal{L}$: $y=\{y_i\}, i\in S, y_i\in\mathcal{L}$. In our problem of stroke classification, we attach a site to each stroke and the set of labels is $\mathcal{L}=\{T,N\}$ for *text* and *non-text* classes. A CRF is based on a graph structure that models conditional dependencies between random variables. More precisely, let $G=(S,E)$ be a graph such that the variables y_i are indexed by the vertices of G. Then (y,x) is a conditional random field if, when conditioned on x, the variables y_i obey the Markov property with respect to the graph

$$p(y_i|x,y_{S\setminus i})=P(y_i|x,y_{\mathcal{N}_i})$$

where $\mathcal{N}_i$ is the set of neighbours of the node i in G, and $y_{\mathcal{N}_i}$ is the set of labels associated to nodes in $\mathcal{N}_i$. By the Hammersley–Clifford theorem, the distribution $p(y|x)$ can be expressed as a product of *potentials* defined over the cliques of the graph G:

$$p(y|x)=\frac{1}{Z(x)}\prod_{c\in G}\Psi_c(y_c,x_c), \tag{1}$$

where c denotes a maximal clique on the graph G and Ψ_c is the potential attached to the clique. x_c is the necessary information for computing Ψ_c and need not be limited to observations indexed by

c. $Z(x)$ is a normalization function that guarantees the probabilistic nature of $p(y|x)$. It is defined as a sum over all the possible labellings $\mathcal{Y}$, where $|\mathcal{Y}| = |S|^{|\mathcal{L}|}$:

$$Z(x) = \sum_{y' \in \mathcal{Y}} \prod_{c \in G} \Psi_c(y'_c, x_c) \tag{2}$$

The graph structure and clique system strongly depend on the observations x, but it is constructed as a repetition of similar structures called *clique templates* that are defined to factorise the model parameters. If $\mathcal{C}$ denotes the set of clique templates, the CRF model is expressed as

$$p(y|x) = \frac{1}{Z(x)} \prod_{C_p \in \mathcal{C}} \prod_{c \in C_p} \Psi_c(x_c, y_c; \theta_p) \tag{3}$$

where Ψ_c indicates a potential on clique c parametrised by θ_p. The expression of potentials is

$$\Psi_c(x_c, y_c; \theta_p) = \exp\left\{ \sum_{k=1}^{K_p} \theta_p^k f_p^k(y_c, x_c) \right\} \tag{4}$$

where $f_p^k(y_c, x_c)$ are feature functions and θ_p^k are model parameters for cliques in C_p. The definitions of clique templates are given in Section 4, where we investigate different ways of constructing the graph structure, in order to model interactions of different natures. Expressions of the feature functions f are also given in that section.

As it is usually done in practical implementations of CRFs, we limit the model to cliques of sizes 1 and 2 only. Thus, the two types of possible cliques are the unary cliques, involving a single site, and the binary cliques, involving the relationship between two sites. In the sequel, the potential functions associated to the two types of cliques are referred to as *association potentials* and *interaction potentials* respectively.

3.2. Inference

The problem of inference in a CRF consists in finding the labelling y^* maximizing the conditional probability $p(y|x)$ (see Eq. (3)). Over the years, many methods have been proposed for solving inference problems in Markov random fields, often formulated as an equivalent energy minimization problem [33]. Several methods permit to exactly find the optimal labelling when the model has certain properties, such as belief propagation on tree-structured graphs or graph cut algorithms with submodular interactions potentials. However, finding the optimal labelling is NP-hard in general and approximate algorithms are thus required for inference in arbitrary graphs.

In this work, the structure adopted for the CRF is necessarily irregular, since the sites are attached to strokes that are distributed irregularly over time and space. Moreover, as we aim at introducing complementary sources of contextual knowledge in the CRF framework, a flexible inference method, able to cope with loopy, more or less densely connected graphs is necessary. For these reasons, we adopt iterated conditional mode (ICM), as a simple approximate inference algorithm able to handle arbitrary graphs, by making local approximations of the global probability distribution [33]. It is a greedy algorithm that sequentially maximises the local conditional probabilities of the solution, assuming that the labelling maximizing $p(y|x)$ can be approximated by the labelling maximising local posterior probabilities $p(y_i|x_i, y_{\mathcal{N}_i})$ at each site i. At iteration t, for each site i in turn, it selects the label l maximising $p(y_i^{t+1} = l|x_i, y_{\mathcal{N}_i}^t)$ according to the current labelling of its neighbours $y_{\mathcal{N}_s}^t$. This avoids the computation of the partition function $Z(x)$, but it can only guarantee convergence to a local optimum and is sensitive to the initial labelling y^0. For initial labelling, we select at each site the label having the highest probability based on the *association potential* only.

3.3. Parameter learning

Guided by the need for flexibility with respect to different graph definitions as explained before, we chose an approximate algorithm for performing parameter learning. Values of parameters θ are selected so as to maximize the pseudo-likelihood (PL) of the ground truth labellings over a training set. The PL is a local approximation of the true likelihood, where it is assumed that the training objective only depends on conditional distributions over single variables, hence no assumptions are required regarding the graph structure. Formally, the log-PL of a labelling is the sum of the logarithm of $p_{PL}(y_i|y_{\mathcal{N}_i}, x; \theta)$ at each site, with

$$p_{PL}(y_i|y_{\mathcal{N}_i}, x; \theta) = \frac{\prod_{c \ni i} \Psi_c(y_c, x_c)}{\sum_{y'} \prod_{c \ni i} \Psi_c(y'_c, x_c)}, \tag{5}$$

where $y_c \subset \{y_i\} \cup y_{\mathcal{N}_i}$ and $y'_c \subset \{y'_i\} \cup y_{\mathcal{N}_i}$. In this formulation, computation of the denominator only involves summing over the labels of a single site y'_i, while considering as fixed the ground truth labelling of its neighbours $y_{\mathcal{N}_i}$. The maximization of PL is performed by a second order gradient descent algorithm (limited-memory BFGS [34], for which we adopt the *libLBFGS* library implementation).

4. Modelling stroke interactions with CRF

Since our goal is to combine various sources of contextual information for improving the task of stroke classification, we make full usage of the CRF flexibility for modelling interactions. In this section, we present in details our choices of features, association potentials and interaction potentials. In addition, we propose several definitions for clique systems that can be integrated simultaneously in the CRF for combining various types of interactions.

To give an overview of the contextual sources included in the model, Fig. 2 describes the conditional random field structure around the site y_k attached to the stroke x_k. The sites of the CRF are connected in an irregular graph with edges carrying pairwise interaction potentials. Different colours of edges denote different types of interactions, modelled under different clique templates. It should be noted that the association potentials involve a local

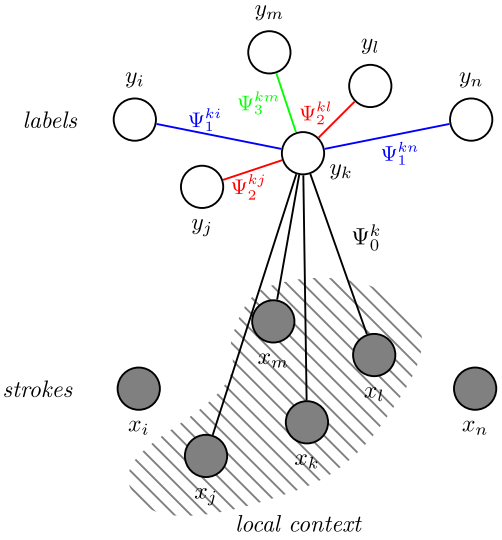

Fig. 2. Description of the context sources influencing the classification of stroke x_k in the proposed CRF model. The association potential Ψ_0^k is affected by features extracted from a local context around the stroke x_k. Several types of interactions between labels are also integrated in the model, represented with different colours. Ψ_p^{ij} denote interaction potentials between sites i and j, it is a short notation for $\Psi_{c_{ij}}(x_{c_{ij}}, y_{c_{ij}})$. Dependencies of the interaction potentials with respect to the observations are not illustrated in the model for simplicity.

context around the stroke x_k, which is made possible thanks to the non-generative formulation of CRF that allows non-independent features to be extracted from observations.

In the next paragraph, we first expose in details the definition adopted for association potentials (stroke classification model), including features from the local context. Then we introduce the general formulation of pairwise clique systems, with their associated feature functions. Finally, we present the different definitions for constructing cliques, in an attempt to model simultaneously the different sources of interaction influencing the stroke classification.

4.1. Individual stroke classification

4.1.1. Association potentials

Association potentials attached to each site of the CRF model the compatibility between a local observation (from the stroke attached to the site and its local context) and a label (either *text* or *non-text*). We decide to derive these potentials from the output of a support vector machine (SVM) classifier, trained on the task of discriminating textual from non-textual strokes. The SVM outputs are converted into posterior probabilities by fitting sigmoid functions with parameters estimated on a validation set [35]. The negative logarithms of probabilistic outputs $l_1(y|x) = -\log(p_{svm}(y|x))$ are then combined as follows to define four association features:

$$f_1^{i,\mathrm{T}}(y_s, x) = \mathbb{1}_{\{y_s = i\}} l_1(\mathrm{T}|x) \tag{6}$$

$$f_1^{i,\mathrm{N}}(y_s, x) = \mathbb{1}_{\{y_s = i\}} l_1(\mathrm{N}|x) \tag{7}$$

and associated to parameters $\theta_1^{i,\mathrm{T}}$ and $\theta_1^{i,\mathrm{N}}$, with $i \in \mathcal{L}$. In the feature expressions, indicator functions imply that different weights are applied to the outputs of the SVM depending on the labelling hypothesis y_s of the site s. A similar idea was applied before in several CRF implementations [36,37]. In addition to the above features, four features with weights $\theta'^{i,\mathrm{T}}_1$ and $\theta'^{i,\mathrm{N}}_1$ are defined from the prior probabilities of each class estimated over the training set. These eight features in total constitute the association potentials of the CRF.

4.1.2. Stroke descriptors

The classification of a stroke x as *text* or *non-text* by the SVM classifier involves a set of features that we call here *descriptors* to avoid ambiguity with the CRF feature functions f. We extract for each stroke a set of 19 descriptors as presented in Table 1.

The first 13 descriptors characterise properties of the stroke x_k itself. Most of these descriptors have also been considered for the problem of isolated text/non-text stroke prediction (see [13–15]) or mode detection [26]. In addition, as the importance of *local context* for stroke classification has been demonstrated before [2,14,27], we integrate a set of six descriptors of the spatial and temporal neighbours of the stroke.

4.2. Interaction modelling

An important advantage of the CRF framework is to allow the modelling of empirical dependencies between labels over the sites. For the problem at hand, it means that we can model interactions expressing the influence of labelling of strokes over each other, conditioned on the observations, without limitations. These interactions convey contextual information, for example expressing the intuition that strokes close to each other tend to have similar labels. We formalize here the general form of interaction potentials introduced in Section 3 and explicit the feature functions. Definitions proposed for neighbourhood systems will be presented in Section 4.3.

4.2.1. Interaction potentials

Suppose that two sites s and t, corresponding to strokes x_s and x_t, are connected in the graphical model resp. (the different choices for connecting strokes together are discussed in details in the following paragraph). The pairwise interaction potentials involve the following feature functions, if q indexes a system of cliques with shared parameters in the CRF (i.e. $C_q \in \mathcal{C}$, see Eq. (3)):

$$f_q^{ij,k}(y_{s,t}, x) = \mathbb{1}_{\{y_s = i, y_t = j\}} p_q^k(x) \quad \text{with weights } \theta_q^{ij,k} \tag{8}$$

for every $i,j \in \mathcal{L}$, and for $k \in \{1 \ldots K_q\}$. As in association potentials, we consider different weights depending on the labelling of the clique s,t by incorporating an indicator function in the feature definitions. Thus if K_q *descriptor functions* p_q^k are defined, the number of associated feature functions $f_q^{ij,k}$ is $4 \times K_q$. This design, also adopted by Ye and Viola in their work [18], permits to factorise the computation of the descriptor functions that depend only on the observations x. This is of special importance in our implementation where we propose to share the descriptor functions across the clique systems. For example, if two sites are considered to be both spatial and temporal neighbours, they will interact according to feature functions with different weighting parameters ($\theta_q^{ij,k}$ and $\theta_{q'}^{ij,k}$) but with identical underlying descriptors. Since all the clique systems share the same descriptor functions, we can drop the q indices and simply denote the descriptors p^k. We also have $\forall C_q \in \mathcal{C},\ K_q = K$. Directly extracting descriptors from observations is an important difference with respect to our previous works [2]

Table 1
About 19 descriptors extracted from stroke x_k and local context.

#	*Description*
1	Trajectory length of x_k
2	Area of the convex hull of x_k
3	Duration of the stroke
4	Ratio of the principal axis of x_k seen as a cloud of points
5	Rectangularity of the minimum area bounding rectangle of x_k
6	Circular variance of points of x_k around its centroid
7	Normalized centroid offset along the principal axis
8	Ratio between first-to-last point distance and trajectory length
9	Accumulated curvature
10	Accumulated squared perpendicularity
11	Accumulated signed perpendicularity
12, 13	Width and height of x_k, normalised by the median stroke height in the document
14	Number of temporal neighbours of x_k
15	Number of spatial neighbours of x_k
16, 17	Average and standard deviation of the distances from x_k to spatial neighbours
18, 19	Average and standard deviation of lengths of spatial neighbours

Table 2
About 20 descriptors extracted for a pair of strokes x_s, x_t.

#	7 *Description*
1	Minimal distance between x_s and x_t
2, 3	Horizontal and vertical distances of the centroids of x_s and x_t
4	Off stroke distance from x_s to x_t or from x_t to x_s
5	Off stroke distance projected on X and Y axis
6, 7	Maximal and minimal distance from any endpoint of x_s to any endpoint of x_t
8	Temporal distance between x_s and x_t
9	Ratio between off stroke distance and temporal distance
10	Ratio between areas of the largest bounding box of x_s and x_t and that of their union
11	Number of strokes in the sequence between x_s and x_t
12	Number of spatial neighbours of x_s or x_t intersecting the bounding box defined by centroids of x_s and x_t
13	Number of spatial neighbours of x_s or x_t intersecting the line joining the centroids
14	Ratio between areas of the bounding boxes
15	Ratio between the lengths of the diagonal of the bounding boxes
16	Ratio between the heights of the bounding boxes
17	Ratio between the widths of the bounding boxes
18	Ratio between the stroke curvatures
19	Ratio between the stroke lengths
20	Ratio between the stroke durations

and other systems [19,20] that make use of independently trained pairwise stroke classifiers for interaction potentials. It not only permits to factorise descriptors among the clique systems, but it also guarantees an easier and more direct process for training the CRF with different clique systems, without the need for external training of new classifiers for each choice of neighbourhood definition thresholds.

4.2.2. Interaction descriptors

A set of $K=20$ interaction descriptors is extracted from any pair of connected strokes in the model. The list of descriptors for two strokes x_s and x_t is given in Table 2.

The 13 descriptors from the first category focus on the gap between the strokes x_s and x_t to measure under different aspects the distance between them: spatial proximity, temporal gap, distance between their endpoints, overlapping degree, interleaved strokes. The seven descriptors from the second category measure the similarity between the strokes, by computing normalised ratio of some of their properties like bounding box dimensions, trajectory length, mean curvature or size. All the distance-based descriptors are defined symmetrically. The ratios are also normalised such that the result is symmetrical and between 0 and 1. For each descriptor value $v^k(x)$, the associated function $p^k(x)$ is obtained by first linearly mapping $v^k(x)$ between bounds v^k_{min} and v^k_{max} (either theoretically defined or estimated from data evidence), and by applying exponential transformation

$$p^k(x) = \exp\left\{\frac{v^k(x)-v^k_{min}}{v^k_{max}-v^k_{min}}\right\}.$$

4.3. Clique system definitions

In this paragraph, we propose a variety of definitions for the clique systems, each one describing a different source of contextual information for stroke classification. Contrary to many applications of graphical models to image analysis, where interactions can be defined over a regular grid, the irregular distribution of strokes in an online document leaves many possibilities for defining spatial neighbourhood systems. In addition, the availability of dynamic information enables definitions of temporal neighbourhoods.

4.3.1. Spatial system

The spatial neighbourhood is obviously an important source of knowledge when trying to categorise a stroke as part of text or not. We propose in this work to define a *spatial neighbourhood* (SPA) based on the minimal distance to the stroke. In the SPA system, two strokes x_s and x_t are considered as connected if their minimal distance $d(x_s,x_t)$ is below a given threshold T_1. Illustrations for spatial neighbourhoods are given in Fig. 3(a).

4.3.2. Temporal system

The online nature of the data permits to see the document as a sequence of strokes. The temporal neighbourhood system models interaction between strokes that are written successively in the document. Considering that elements (such as paragraphs, parts of diagram, tables, etc.) are often drawn with an uninterrupted sequence of strokes, the temporal interaction can be of great help for classifying strokes. Several results reported in the literature confirm this intuition [2,20].

We define the *temporal neighbourhood* system (TMP) considering that two strokes are neighbours if the temporal distance between them is below a threshold T_2 (in seconds) and they are not separated by more than T_3 intermediate strokes in the sequence. Fig. 3(b) gives examples of temporal neighbourhoods.

4.3.3. Intersecting system

In complement to the spatial neighbourhood system presented above, one might think that the strokes that intersect each other present special interaction properties, and thus should be connected under a specific neighbourhood system. Indeed, two intersecting strokes are more likely to be of the same type (both text or both non-text) than two close but non-intersecting strokes. The *intersection neighbourhood* system (INT) is then defined by connecting only strictly intersecting strokes. By definition, pairs of strokes connected in INT are also spatial neighbours (in SPA), because their distance is 0. It is, however, hoped that this additional system can model more specific interactions between labels of intersecting strokes, which can be of importance for instance in the case of ruled tables or tangled drawings.

4.3.4. Lateral system

As stated in the introduction, freely handwritten documents are especially difficult to analyse because they present no stable properties that can be easily exploited for their interpretation.

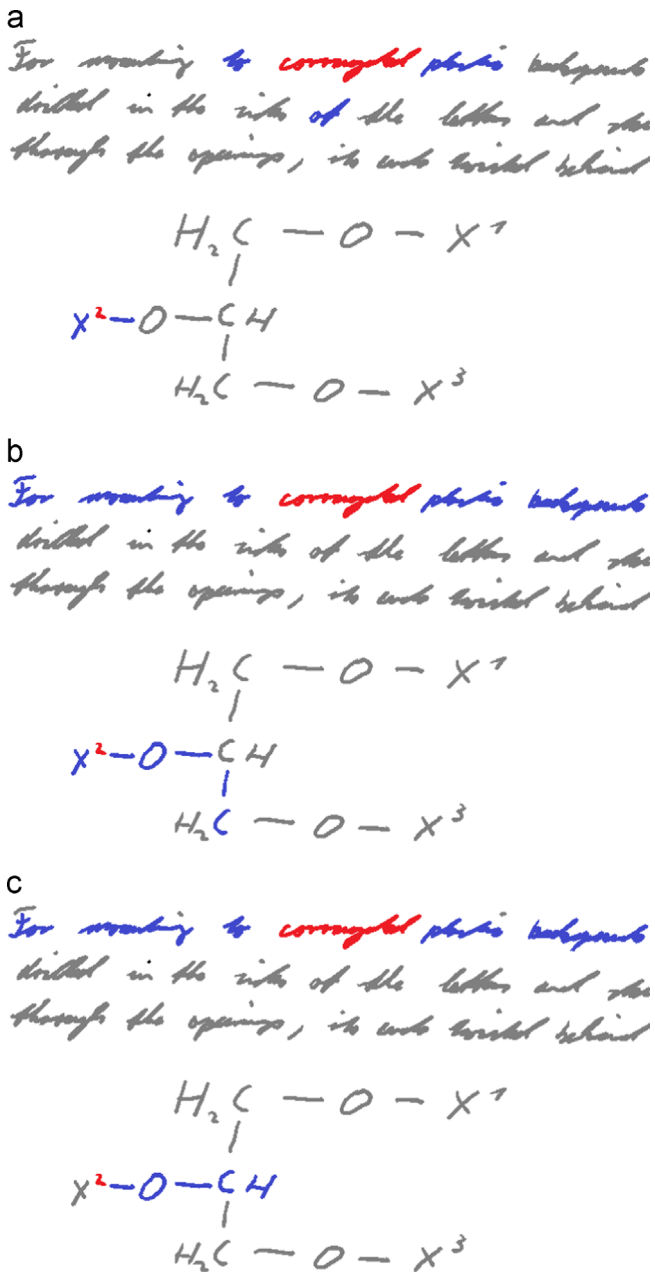

Fig. 3. Illustrations of several neighbourhood systems. Blue strokes are neighbours of the red strokes according to the different clique systems: spatial (a), temporal (b) or lateral (c). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

However, despite the general lack of prior knowledge available, the textual content can present some regularity that can be of precious help. Because text is usually organised in horizontal lines, we propose to consider interaction between a stroke and other strokes that could belong to the same (hypothetical) text line. In the *lateral neighbourhood* system (LAT) we consider strokes as neighbours if they are more or less horizontally aligned. The horizontal alignment condition checks if the vertical distance between the centres of the strokes bounding boxes is below a threshold T_4. The LAT neighbourhood system is illustrated with an example in Fig. 3(c).

4.3.5. Stroke continuation system

In his works about perception of hand-drawn sketches, Saund [38] emphasised the importance of paths formed by smoothly continuous strokes. Accordingly, we propose to define a last neighbourhood system, by considering connections between strokes that present a *smooth continuation* property. The CNT clique system connects strokes x_s and x_t if they respect basic constraints on their directions and on the distance between their endpoints. First the trajectory of the pen at the end of x_s and at the beginning of x_t should follow a similar direction, and this direction in turn should be similar to the direction from last point of x_s to the first point of x_t (off stroke direction). Two directions are considered as similar if they form an angle of less than $\pi/3$. Finally, the distance between the last point of x_s and the first point of x_t (off stroke distance) should be below a threshold T_5.

5. Experiments

The objective of our experiments is to evaluate the contribution of each type of contextual information for the task of text/non-text classification in online handwritten documents and to validate the exploitation of the CRF framework for their efficient combination. We first present the experimental database IAM-OnDo introduced by Indermuhle et al. [21] that contains realistic free-form online handwritten documents. Then the details of unary stroke classifier are given, as well as the thresholds settings determining the clique systems introduced in Section 4.3. At last, we present results obtained by integrating interaction systems and show that their combination permits to outperform the state-of-the-art result reported on this dataset.

5.1. IAM-OnDo database

We run our experiments on the documents from IAM-OnDo database, a publicly available collection of freely handwritten online documents with full ground truth content annotation and transcription [21]. The database contains 1000 documents mixing handwritten text, drawings, diagrams, formulas, tables, lists, and marking elements arranged in an unconstrained way. We adopt the same assignment of content categories to text or non-text labels as suggested by the database authors [14], where 15–20% of the strokes are considered as non-text. For our experiments, we select documents from *set*0 and *set*1 as training samples (for both SVM training and CRF parameter learning) and used *set*2 as a validation set (for optimisation of clique systems parameters). All the results reported in the following are obtained on an independent test set (*set*3), where performance is measured at the stroke level. Note that a minor category of content annotated as *garbage* is ignored. We also exclude the elements annotated as *marking*, because their distribution is very atypical as users where specifically told to write these *markings* after they have finished composing the document. We believe that marking elements should be dealt with in an interactive context and have little meaning for posterior document analysis, as in our experiment. All in all, the data partition, the text/non-text assignment and the choice of categories precisely match with the settings used by Indermuhle et al. in their work on mode detection [31]. Table 3 summarises the number of documents and strokes for training, validation and testing of the system, as well as the proportion of *text* (T) and *non-text* (N) strokes.

5.2. Local stroke classifier

We use an support vector machine (SVM) as the unary stroke classifier, trained with the 19 descriptors presented in Table 1. The definitions for *spatial neighbours* and *temporal neighbours* for *local context descriptors* match the definitions of the neighbourhood systems SPA and TMP, with thresholds values made explicit in the

Table 3
Number of documents and strokes for training, validation and test, and distribution of text/non-text classes in each set.

Set	# documents	# strokes	% N	% T
Training	403	143,348	18.52	81.48
Validation	200	68,725	15.87	84.13
Test	203	70,927	18.29	81.71

Table 4
Parameter definition for clique systems and mean number of neighbours per stroke in the validation dataset.

System	Parameters	Ngb/stroke
SPA	$T_1 = 10$	2.75
TMP	$T_2 = 3.5$, $T_3 = 4$	7.14
INT	–	0.37
LAT	$T_4 = 4$	11.58
CNT	$T_5 = 20$	0.97

next paragraph. After experiments with different kernel functions and meta parameters, we adopted the 3rd order polynomial kernel that yields best results for isolated stroke classification on the validation dataset. The same SVM classifier is used in all the experiments. It determines the *association potentials* defining the single sites cliques of all the variants of CRFs in our experiments (see Section 4). The classification rate obtained by the SVM classifier in isolation on the strokes of the test dataset is 94.44%.

5.3. Clique system definitions

The definitions of clique systems given in Section 4 involve several threshold parameters. For each clique system, we experimentally select the value of the threshold that maximizes the stroke recognition rate on the validation dataset. This is done independently for each system, i.e. by exploiting the CRF model involving only the corresponding neighbourhood system. Table 4 shows the values retained for the parameters. For each neighbourhood system, the third column displays the average number of neighbours per stroke measured on the validation dataset, as an indication of the graph density induced by the chosen threshold values. There was no attempt to control directly these values or to fix minimum or maximal limits on the number of neighbours per stroke.

5.4. Evaluation of contextual sources

We compare in this part the contribution of the sources of context for the stroke recognition task with the CRF. Several variants of the system are proposed, integrating one or several sources of context by including clique systems. All the experimental results are reported in Table 5, with recognition rates measured on the evaluation dataset. For reference, we also report the best performance obtained so far on this dataset [31].

The first five experiments highlight the importance of contextual information, since all the clique systems are beneficial for the stroke classification, in comparison with the performance of the isolated SVM classifier (94.44%). The TMP clique system is by far the most beneficial to the recognition, followed by the spatial system SPA. The INT and CNT system have the smallest impact on the classification, but this can be easily explained by the sparsity of these systems (as shown in Table 4). Note that they still provide a significant increase of performance with respect to the SVM in isolation.

Experiment 6 demonstrates the interest of combining spatial and temporal clique systems, as was reported in our previous works [2]. This performance of 96.81% is slightly better than the recognition rate of 96.66% obtained with a CRF also involving spatial and temporal neighbourhoods but with pairwise potentials modeled by SVM classifiers [2]. In addition, experiment 8 shows that the LAT neighbourhood system provides useful and complementary context in the model. The idea of introducing prior knowledge by this horizontal text line context is shown to be beneficial at least for this dataset. By comparison with experiments 7 and 9, LAT is shown to contribute more importantly to the performance than INT and CNT that only have a limited impact (but still improve the performance).

Table 5
Recognition rate for text/non-text stroke classification by combining different clique systems in the CRF model.

#	Interactions	Rate (%)
1	SPA	95.81
2	TMP	96.57
3	INT	94.93
4	LAT	95.65
5	CNT	95.38
6	SPA, TMP	96.81
7	SPA, TMP, INT	96.89
8	SPA, TMP, LAT	97.05
9	SPA, TMP, CNT	96.83
10	SPA, TMP, LAT, INT	97.21
11	SPA, TMP, LAT, CNT	97.08
12	SPA, TMP, LAT, INT, CNT	97.23
[31]	BLSTM neural network	97.01

It should be highlighted that documents from the IAM-OnDo dataset contain text lines oriented and skewed in a number of angle variations, thus a significant number of text elements are not horizontally aligned. The motivation for defining the LAT system and the reason for its effectiveness is that most text elements are indeed surrounded by horizontally aligned text elements, but it does not prevent the CRF from detecting text elements that are not (see for example detection results of Fig. 4(d)).

Our observations are confirmed by experiments 10 and 11, where a fourth interaction system is added to the system from experiment 8, allowing to improve the performance from 97.05% to 97.21% (with INT) and 97.08% (with CNT). Finally, the full system combining the five neighbourhoods systems performs the best, with an overall recognition rate of 97.23% (experiment 12). This result significantly improves over the best performance of 97.01% reported by Indermuhle et al. [31] on the same dataset.[1]

Contrary to online mode detection systems, such as [28,29] that need to meet real-time constraints, our system is applied a posteriori on completed documents, hence a longer processing time is less problematic. The time required for processing an average document from the test dataset varies from 1.26 s (with system 1) to 1.53 s (with system 12).[2] Detailed analysis shows that most of the time is spent in building the graph of the strokes (from 0.68 s to 0.76 s), computing the features and running the isolated classifier (from 0.56 s to 0.58 s), while the inference itself is comparably very fast even if its cost increases greatly as the graph gets more densely connected (from 0.02 s to 0.19 s).

5.5. Detailed results per category

In order to provide a better insight on the contribution of contextual sources to the recognition performance, we present

[1] We consider that the performance of 98.57% reported by Otte et al. [29] cannot be fairly compared to our results because their system involves an unclear pre-segmentation of the documents into sub-sequences.

[2] Processor: Intel Core i5-24003.10 GHz.

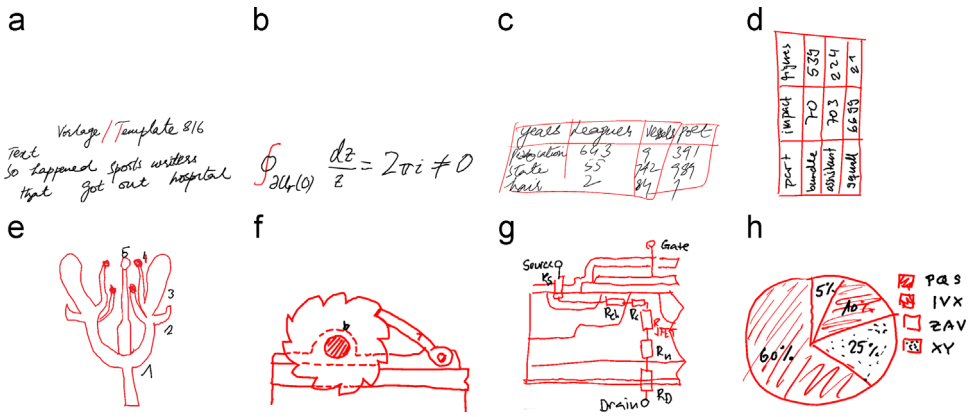

Fig. 4. Text/non-text classification results obtained with system 12 over different objects. Colours represent the class attributed to each stroke: black for text and red for non-text. (a) text, (b) text, (c) table, (d) table, (e) diagram, (f) diagram, (g) diagram, and (h) diagram.

stroke classification rates partitioned according to various categories of content from the documents. As stated in [29], the difficulty of classifying strokes largely depends on the type of content considered. For example, *tables* are elements where *text* and *non-text* strokes are mixed but are easy to discriminate because non-text strokes are usually long segments (table rules) that are very different from textual strokes. Conversely, diagrams and drawings present more difficulties because they may contain non-textual strokes that present similarities with textual strokes. Table 6 shows the recognition rates obtained by four distinct systems over three content categories: text blocks (full-text paragraphs and lists), diagrams (mixed text/non-text made of labelled charts, scatter plots, free drawings) and tabular structures (mixed text/non-text). The four systems are trained on the whole data (as before) but applied separately on each category of content.

The first line shows the performance of the isolated SVM classifier for stroke classification over the three categories. It confirms that mixed text and non-text in diagrams are much harder to distinguish than in tabular structures. The three other lines present performances of systems from experiments 1, 8 and 12 from Table 5. System 1 shows that the contextual information from SPA greatly contributes to the classification rate in the three categories. Similarly, TMP and LAT significantly boost the performance of system 8 over the three domains. Analysing results from system 12 reveals that the two interaction systems with a fewer impact, namely CNT and INT, contribute to improve the performance for *text blocks* and *diagrams* category, but this is mitigated by a decrease of performance for the *tables* category. The last line of the table shows the number of strokes in each category, suggesting to weight differently each group when appreciating the differences between two systems.

To provide further qualitative appreciation of the results and give insight into the errors made by our best system, Fig. 3 presents the results obtained in several situations. Fig. 4(a) and (b) presents objects from the *text* category. Strokes in black are correctly classified as text, while strokes in red are mistakenly labelled as non-text. In Fig. 4(c) and (d), tables are presented, where all the strokes have been correctly classified. Fig. 4(e)–(h) reflects the diversity of diagrams and the increased difficulty of separating text from graphics in them. Diagram 4(g) and (h) is especially challenging for the system. Textual strokes that overlap with graphical ones are sometimes mislabelled as non-textual (such as "R_{JFET}" label in 4(g) or a *percent* symbol in 4(h)). Conversely, some small graphical elements (as the dots in 4(h)), or strokes that are aligned with textual content (as the circle next to the word "*Drain*" in 4(g)) are misclassified as text.

Table 6
Recognition rate for text/non-text stroke classification over different categories of content.

System	Text blocks	Tables	Diagrams
SVM	98.64	96.93	81.27
1	99.13	97.96	85.34
8	99.51	99.02	88.63
12	99.58	98.95	88.88
# Strokes	46,556	7046	17,325

Most of the errors made by the system are quite understandable because they often occur in ambiguous situations. We believe directions for improvements include the exploration of better training and inference algorithms, for example the application of Loopy belief propagation to compute approximate MAP labelling at running time and to estimate marginals for approximate maximum likelihood training. However, a longer term perspective will jointly consider the problems of segmenting the document into homogeneous regions and recognizing textual elements in order to resolve ambiguities at stroke level.

6. Conclusion

This paper presents an investigation about the importance of contextual information for the classification of strokes as text or non-text in freely handwritten online documents. Under the well principled conditional random field framework, we proposed to define and combine complementary stroke interactions and we evaluated their contribution to the stroke classification performance. We not only considered spatial and temporal contexts but also included additional types of interactions (intersecting strokes, horizontally aligned strokes, strokes forming continuous paths), that were all shown to be beneficial for the classification performance. The best combination of contextual sources performs significantly higher than previous results on the same database. We consider that these improvements about the modelling of contextual knowledge will be useful for designing better systems towards higher-level document understanding tasks, such as segmentation and structural analysis.

Conflict of interest statement

None declared.

Acknowledgements

This work is supported by the Chinese Academy of Sciences under the Fellowships for Young International Scientists program (No. 2012Y1GB0001), and by National Natural Science Fundation of China under the Research Fund for International Young Scientists (No. 61250110082).

References

[1] R. Plamondon, S. Srihari, E. Polytech, Q. Montreal, Online and off-line handwriting recognition: a comprehensive survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 63–84.
[2] A. Delaye, C.-L. Liu, Text/non-text classification in online handwritten documents with conditional random fields, in: C.-L. Liu, C. Zhang, L. Wang (Eds.), Proceedings of the Chinese Conference on Pattern Recognition 2012, Communications in Computer and Information Science, vol. 0321, Springer, Heidelberg, 2012, pp. 514–521.
[3] C.-L. Liu, S. Jaeger, M. Nakagawa, Online recognition of chinese characters: the state-of-the-art, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 198–213.
[4] A. Delaye, E. Anquetil, Fuzzy relative positioning templates for symbol recognition, in: Proceedings of the 11th International Conference on Document Analysis and Recognition, 2011, pp. 1220–1224.
[5] A. Delaye, E. Anquetil, Hbf49 feature set: a first unified baseline for online symbol recognition, Pattern Recognition 46 (2013) 117–130.
[6] D. Willems, R. Niels, M. van Gerven, L. Vuurpijl, Iconic and multi-stroke gesture recognition, Pattern Recognition 42 (2009) 3303–3312.
[7] H. Mouchère, C. Viard-Gaudin, D. Kim, J. Kim, U. Garain, Crohme2011: competition on recognition of online handwritten mathematical expressions, in: Proceedings of the 11th International Conference on Document Analysis and Recognition, 2011, pp. 1497–1500.
[8] S. Macé, E. Anquetil, Eager interpretation of on-line hand-drawn structured documents: the DALI methodology, Pattern Recognition 42 (2009) 3202–3214.
[9] G. Feng, C. Viard-Gaudin, Z. Sun, On-line hand-drawn electric circuit diagram recognition using 2d dynamic programming, Pattern Recognition 42 (2009) 3215–3223.
[10] T. Ouyang, R. Davis, Learning from neighboring strokes: combining appearance and context for multi-domain sketch recognition, in: Proceedings of the 23rd Annual Conference Neural Information Processing Systems, 2009, Curran Associates, Inc, 2009, pp. 1401–1409.
[11] Z. Lin, J. He, Z. Zhong, H.-Y. Shum, Table detection in online ink notes, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1341–1346.
[12] X. Zhang, M. Lyu, G. Dai, Extraction and segmentation of tables from Chinese ink documents based on a matrix model, Pattern Recognition 40 (2007) 1855–1867.
[13] A. Jain, A. Namboodiri, J. Subrahmonia, Structure in on-line documents, in: Proceedings of the 6th International Conference on Document Analysis and Recognition, 2004, pp. 844–848.
[14] E. Indermühle, Analysis of Digital Ink in Electronic Documents, Ph.D. Thesis, University of Bern, 2012.
[15] K. Mochida, M. Nakagawa, Separating drawings, formula and text from free handwriting, in: Proceedings of the 11th Conference of the International Graphonomics Society, 2003, pp. 216–219.
[16] X.-D. Zhou, D. Wang, C.-L. Liu, A robust approach to text line grouping in online handwritten japanese documents, Pattern Recognition 42 (2009) 2077–2088.
[17] M. Shilman, Z. Wei, S. Raghupathy, P. Simard, D. Jones, Discerning structure from freeform handwritten notes, in: Proceedings of the 7th International Conference on Document Analysis and Recognition, 2003, pp. 60–65.
[18] M. Ye, P. Viola, Learning to parse hierarchical lists and outlines using conditional random fields, in: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004, pp. 154–159.
[19] X.-D. Zhou, C.-L. Liu, Text/non-text ink stroke classification in Japanese handwriting based on markov random fields, in: Proceedings of the 9th International Conference on Document Analysis and Recognition, 2007, vol. 1, IEEE, pp. 377–381.
[20] C. Bishop, M. Svensen, G. Hinton, Distinguishing text from graphics in on-line handwritten ink, in: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004, IEEE, pp. 142–147.
[21] E. Indermühle, M. Liwicki, H. Bunke, IAMonDo-database: an online handwritten document database with non-uniform contents, in: Document Analysis Systems, 2010, pp. 97–104.
[22] A. Awal, G. Feng, H. Mouchère, C. Viard-Gaudin, et al., First experiments on a new online handwritten flowchart database, in: Proceedings of the SPIE-IS&T Electronic Imaging, Document Recognition and Retrieval XVIII, vol. 1, 2011.
[23] A. Bhat, T. Hammond, Using entropy to distinguish shape versus text in hand-drawn diagrams, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., 2009, pp. 1395–1400.
[24] J. Rodriguez, G. Sanchez, J. Llados, Categorization of digital ink elements using spectral features, in: W. Liu, J. Llados, J.-M. Ogier (Eds.), Graphics Recognition. Recent Advances and New Opportunities, Lecture Notes in Computer Science, vol. 5046, Springer, Berlin/Heidelberg, 2008, pp. 181–190.
[25] D. Willems, S. Rossignol, L. Vuurpijl, Features for mode detection in natural online pen input, in: Proceedings of the 12th Biennial Conference of the International Graphonomics Society, 2005, pp. 113–117.
[26] D. Willems, S. Rossignol, L. Vuurpijl, Mode detection in on-line pen drawing and handwriting recognition, in: Proceedings of the 8th International Conference on Document Analysis and Recognition, 2005, IEEE, pp. 31–35.
[27] E. Peterson, T. Stahovich, E. Doi, C. Alvarado, Grouping strokes into shapes in hand-drawn diagrams, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010.
[28] M. Weber, M. Liwicki, Y. Schelske, C. Schoelzel, F. Strauß, A. Dengel, MCS for online mode detection: evaluation on pen-enabled multi-touch interfaces, in: Proceedings of the 11th International Conference on Document Analysis and Recognition, 2011, IEEE, pp. 957–961.
[29] S. Otte, D. Krechel, M. Liwicki, A. Dengel, Local feature based online mode detection with recurrent neural networks, in: Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition, 2012, pp. 531–535.
[30] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the International Conference on Machine Learning, 2001, pp. 282–289.
[31] E. Indermuhle, V. Frinken, H. Bunke, Mode detection in online handwritten documents using BLSTM neural networks, in: Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition, 2012, pp. 302–307.
[32] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 855–868.
[33] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, C. Rother, A comparative study of energy minimization methods for Markov random fields, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision ECCV 2006, Lecture Notes in Computer Science, vol. 3952, Springer, Berlin/Heidelberg, 2006, pp. 16–29.
[34] D. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical Programming 45 (1989) 503–528.
[35] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: D.S.A.J. Smola, P. Bartlett, D. Schuurmanns (Eds.), Advances in Large Margin Classifiers, vol. 10, Cambridge, MA, 1999, pp. 61–74.
[36] P. Cowans, M. Szummer, A graphical model for simultaneous partitioning and labeling, in: 10th International Workshop on Artificial Intelligence and Statistics, 2005.
[37] C.A. Sutton, A. McCallum, An introduction to conditional random fields, Foundations and Trends in Machine Learning 4 (2012) 267–373.
[38] E. Saund, Finding perceptually closed paths in sketches and drawings, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 475–491.

Adrien Delaye received his MS and PhD degrees from Institut National des Sciences Appliquees (INSA) of Rennes, France, in 2008 and 2011, respectively. He is a post-doctoral researcher at the Institute of Automation, Chinese Academy of Sciences, in Beijing, China. His research concerns pattern recognition, spatial reasoning, structured learning, and is mainly applied to handwritten document processing and pen-based interaction.

Cheng-Lin Liu received the BS degree in Electronic Engineering from Wuhan University, Wuhan, China, the ME degree in Electronic Engineering from Beijing Polytechnic University (current Beijing University of Technology), Beijing, China, the PhD degree in Pattern Recognition and Intelligent Systems from the Institute of Automation of Chinese Academy of Sciences, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a Research Staff Member and later a Senior Researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. From 2005, he has been a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China, and is now the Deputy Director of the laboratory. His research interests include pattern recognition, image processing, neural networks, machine learning, and the applications to character recognition and document analysis. He has published over 160 technical papers at prestigious international journals and conferences. He won the IAPR/ICDAR Young Investigator Award of 2005. He serves on the editorial board of Pattern Recognition Journal, Image and Vision and Computing, and International Journal on Document Analysis and Recognition. He is a Fellow of the IAPR and a Senior Member of the IEEE.