Pattern Recognition Letters 172 (2023) 113–120

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Open set classification of untranscribed handwritten text image documents

Jose Ramón Prieto*, Juan José Flores, Enrique Vidal, Alejandro Hector Toselli

Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València Camino de Vera s/n, València 46022, Spain

ARTICLE INFO

Article history:
Received 1 August 2022
Revised 13 April 2023
Accepted 7 June 2023
Available online 8 June 2023

Edited by: Maria De Marsico

Keywords:
Open set document classification
Handwritten text images
Probabilistic indexing
Neural networks

ABSTRACT

Content-based classification of manuscripts is an important task that is generally carried out by expert archivists. Nevertheless, many historical manuscript collections are so vast that in most cases this task is hardly feasible, even for large, well staffed archives. Nowadays, manuscripts are generally preserved in the form of sets of digital images. Therefore, the technical problem we are interested in is automatic classification of *"image documents"*, each consisting of a set of untranscribed handwritten text images, by the textual contents of the images. The traditional Pattern Recognition classification paradigm does provide the basic tools to deal with this problem. However, in practice, the set of relevant classes of a large documental series is seldom known in advance. Therefore, a classifier trained with a predefined set of classes will systematically fail when new image documents arrive which do not belong to any of the classes assumed in training. Here we adopt the *"Open Set Classification"* framework to extend and consolidate our previous work on image document classification in order to adequately handle new documents from unknown classes. The proposed approaches are based on a relatively novel technology for text image representation known as *"probabilistic indexing"*, which proves very effective to characterise the intrinsic word-level uncertainty exhibited by historical handwritten text images. We assess the performance of this approach on a moderately sized but representative dataset extracted from a huge series of complex notarial manuscripts from the *Spanish Archivo Histórico Provincial de Cádiz*, with good results.

1. Introduction

Huge amounts of digital images of important historical manuscripts are preserved in archives and libraries. Many of these manuscripts are records of daily life affairs. Specifically, we are interested in historical notarial deeds, which make up perhaps the vastest sort of documentary series in archives worldwide. Individual deeds in these series are generally piled up into large bundles or boxes, each typically containing hundreds of deeds and thousands of page images. For series of documents so massive, it is generally difficult or impossible for archives to provide detailed metadata to adequately describe the contents of each bundle, let alone of each individual deed.

Thereupon, bundles, boxes, books, or folders of manuscript images are called "image bundles" or just "*bundles*". A bundle may contain several, often many *"image documents"*, also called "files", "acts" – or *"deeds"* in the case of notarial image documents considered in this work. Image documents are assumed to belong to "types" or *classes*, which are perhaps the most important information needed to describe a manuscript.

So, the task we are interested in is to classify a given untranscribed image document, which may range from a few to a few tens or hundreds of handwritten text images, into a set of classes or types, associated with the topics or (semantic) contents conveyed by the text written in the images. We will refer to this task as *content based image document classification* (CBIDC).

Existing approaches for content-based document classification (DC) assume documents are made up of electronic text, so characters, words and paragraphs are unambiguously given. Therefore the current wisdom to address the proposed CBIDC task would be to first transcribe the images and then apply off-the-self DC techniques. However, manual transcription is not an option and, on the other hand, achieving sufficiently accurate automatic transcripts is generally unfeasible or elusive for large sets of historical manuscripts.[1]

* Corresponding author.
E-mail addresses: joprfon@prhlt.upv.es (J.R. Prieto), juafloar@alumni.upv.es (J.J. Flores), evidal@prhlt.upv.es (E. Vidal), ahector@prhlt.upv.es (A.H. Toselli).

[1] HTR word recognition accuracies as low as 40–60% are reported in [1–3] for historical manuscripts similar to those considered in this work.

https://doi.org/10.1016/j.patrec.2023.06.006

As proposed in previous works [3–5], to overcome these issues we rely on a relatively novel image representation technology called *probabilistic indexing* (PrIx) [6–9]. It has proved very effective in dealing with the intrinsic word-level *uncertainty* generally exhibited by handwritten text and, more so, by historical handwritten text images. PrIx was primarily developed to allow search and retrieval of textual information in large untranscribed manuscript collections [3,7,10,11].[2] However, it has also proved very versatile to properly approach many other tasks (see [8,10,12], e.g.) where text images need to be represented not by "visual features" but by the uncertain textual contents of the different image regions. One of these tasks is CBIDC, considered in this work.

In our proposal, PrIx provides the probability distribution of words which are likely written in the images, from which statistical expectations of *word* and *document frequencies* are estimated. These estimates are then used to compute well-known text features such as *Information Gain* and Tf·Idf, which are in turn considered inputs to a Neural Network classifier.

Note that the CBIDC task here considered is very different from other related tasks, which are often called with similar names. To name a few: "document classification" (DC, mentioned above, which only applies to unambiguous electronic text), "content-based image classification" (applied to single pictures of natural scenes – not text), or "document image classification" (where classes are associated with the visual appearance or page layout of single images). See [4] for a more detailed discussion on these differences, as well as references to previous publications dealing with related problems, but mainly aimed at printed text.

Note also that recent works on document classification, including those based on multimodal approaches and visual transformers [13,14] are far from being applicable to our CBIDC task, where the nature and size of the textual visual objects considered (maybe hundreds of page images) is very different and/or exceedingly large as compared with the single-image objects considered in these works.

On the other hand, it is important to realise that document types do change over the years and, in a realistic scenario, we need to handle image document of classes that had never been seen before. In the traditional classification framework, all these new image documents would be systematically misclassified. Therefore, to properly deal with the proposed task, new image documents which are not of any known class should be detected; that is, the system should refuse or *"reject"* their classification. One key contribution of the present work is to explicitly address this full-fledged CBIDC problem and provide satisfactory solutions within the so-called *"Open Set Classification"* (OSC) framework [15–17].

This work continues research started in [3–5]. After a first, tentative approach to the problem [3], in [4] we distinctively introduced the CBIDC task and explored several ideas to address the underlying basic classification problem. The application considered in [4] was rather artificial and also maybe too ambitious to allow drawing sound conclusions from the empirical results. Then in [5], we selected the most promising methods studied in [4] and applied them to a more focused and realistic CBIDC task. The encouraging results of those studies led to the present work, where we consolidate previous results through wider and more reliable experiments and, as mentioned above, we assume the OSC framework to support the ultimate needs of the practical application of our methods.

OSC has been considered in several recent works, such as [18–23]. While most approaches proposed in these works can hardly be applied to our CBIDC task, we have been able to adapt ideas from [22,23] and compare the resulting methods with the other approaches we propose.

The rest of the paper is organised as follows: in Secs. 2 and 3 we outline the key concepts and details needed to understand PrIx and the approach we propose to embed image documents into a vector space. In Section 4 the proposed techniques for Closed and Open Set classification are presented. Section 5 is devoted to discuss in detail the data set and the empirical settings adopted for the experiments, which are themselves presented in Section 6. Finally Section 7 draws conclusions and suggests further research avenues based on the results of this paper.

2. Probabilistic indexing of handwritten text images

The Probabilistic Indexing (PrIx) framework was proposed to deal with the intrinsic word-level uncertainty generally exhibited by handwritten text in images and, in particular, images of historical manuscripts. In this framework, any element in an image which is likely enough to be interpreted as a word is detected and stored, along with its *relevance probability* (RP) and its location in the image. These text elements are referred to as *"pseudo-word spots"*.

Following [6,9], the RP for an image-region x and a pseudo-word v is denoted as $P(R{=}1 \mid X{=}x, V{=}v)$, but for the sake of conciseness, the random variable names will be omitted and, for $R = 1$, we will simply write R. As discussed in [24], this RP can be approximated as:

$$P(R \mid x, v) = \sum_{b \sqsubseteq x} P(R, b \mid x, v) \approx \max_{b \sqsubseteq x} P(v \mid x, b) \tag{1}$$

where b is a small, word-sized image sub-region or Bounding Box (BB), and with $b \sqsubseteq x$ we denote the set of all BBs contained in x. Note that $P(v \mid x, b)$ is just the posterior probability needed to "recognise" the BB image (x, b). Therefore, assuming the computational complexity entailed by (1) is algorithmically managed [9], any sufficiently accurate isolated word classifier can be used to obtain $P(R \mid x, v)$. Here we use the methods described in [9], as outlined in [4].

This word-level indexing approach has proved to be very robust, and it has been used to very successfully index several large iconic manuscript collections, such as the French Chancery collection [7], the Bentham papers [10], and the Spanish Carabela collection considered in this paper.[3]

3. Feature selection and extraction for CBIDC

Traditional methods to select and extract text features for DC [25] apply only to plain text. For CBIDC, instead, we rely on image PrIx's to *estimate*, rather than *compute* these features.

Since R is a binary random variable, the RP $P(R \mid x, v)$ can be seen as the statistical expectation that v is written in x. As discussed in [4,10], the sum of RPs for all the pseudo-words indexed in an image region x is the expected number of words written in x. Following this estimation principle, all the document and word frequencies needed to select and extract the textual features required for CBIDC can be estimated. This is thoroughly discussed in [4,5]; so only the essential concepts and equations are summarised hereafter.

Let $n(x)$ be the total (or "running") number of words written in an image region x and and $n(X)$ the running words in an image document X which typically encompasses several pages. Let $n(v, X)$ be the frequency of a specific (pseudo-)word v in X. And let $m(v, \mathcal{X})$ be the number of documents in a collection, $\mathcal{X}$, which contain the (pseudo-)word v. The expected values of these counts

[2] See http://prhlt-carabela.prhlt.upv.es/PrIxDemos for a list of public search interfaces based on PrIx

[3] See AHPC in http://prhlt-carabela.prhlt.upv.es/carabela

are [4,10]:

$$E[n(x)] = \sum_{v} P(R \mid x, v),\ E[n(v, X)] = \sum_{x \sqsubseteq X} P(R \mid x, v)$$
$$E[n(X)] = \sum_{x \sqsubseteq X} E[n(x)],\ E[m(v, \mathcal{X})] = \sum_{X \sqsubseteq \mathcal{X}} \max_{x \in X} P(R \mid x, v) \quad (2)$$

The contribution of a word v to the contents of an (image) document X can be characterised by the so-called "term frequency – inverse document frequency", Tf·Idf(v, X) [25]. Let M be the total number of documents in $\mathcal{X}$. Using the above count estimates, Tf·Idf can be computed as follows [4]:

$$\text{Tf·Idf}(v, X) = \frac{E[n(v, X)]}{E[n(X)]} \cdot \log \frac{M}{E[m(v, \mathcal{X})]} \quad (3)$$

One of best known approaches for document representation in DC (and CBIDC alike), is the *bag of words* (BOW) or *vector model* [25]. In this model, a document X is represented as a feature vector, $\vec{X} \in \mathbb{R}^N$, indexed by the N words of an adequate vocabulary V_N where, typically, $\forall v \in V_N$, $X_v = \text{Tf·Idf}(v, X)$.

Clearly, not all the words of an (image) document collection $\mathcal{X}$ are equally informative about the contents or the class of the different $X \in \mathcal{X}$. Therefore *information gain* (IG) is commonly used to rank all the (pseudo-)words in $\mathcal{X}$ in decreasing order of their IG [25]. Then V_N is built up by simply selecting the N (pseudo-)words with higher values of IG. The probabilities required to compute the IG for all v in $\mathcal{X}$ (see [4]) can be estimated using the statistical expectations in Eqs. (2):

$$P(t_v) = \frac{E[m(v, \mathcal{X})]}{M}, \qquad P(c \mid t_v) = \frac{E[m(v, \mathcal{X}_c)]}{E[m(v, \mathcal{X})]}$$
$$P(\bar{t}_v) = 1 - P(t_v), \qquad P(c \mid \bar{t}_v) = \frac{M_c - E[m(v, \mathcal{X}_c)]}{M - E[m(v, \mathcal{X})]} \quad (4)$$

where $\mathcal{X}_c$ is a subset of documents in $\mathcal{X}$ which belong to class c and M_c is the number of documents in X_c.

The notation t_v in Eq. (4) stands for the value of a boolean random variable that is *True* iff, for some random X, the word v appears in X. Therefore, $P(t_v)$ is the probability that $\exists X \in \mathcal{X}$ such that v is written in X, and $P(\bar{t}_v)$ is the probability that *no* document contains v. Similarly, $P(c \mid t_v)$ (resp. $P(c \mid \bar{t}_v)$) is the conditional probability that the class of some document is c if it contains (resp. does not contain) the word v.

4. Image document classification

Let us first consider the most conventional Pattern Recognition (PR) classification paradigm where each image document X in $\mathcal{X}$ is assumed to belong to one of C *known classes*. We will refer to this setting as *"Closed Set Classification"* (CSC).

Using the Tf·Idf vector representation $\vec{X}$ of X, under the minimum-error risk statistical framework, an optimal prediction of the class of X is [26]:

$$c^\star(X) = \arg\max_{c \in \{1,\ldots,C\}} P(c \mid \vec{X}) \quad (5)$$

The posteriors $P(c \mid \vec{X})$ can be computed following several well-known approaches, some of which were discussed and tested in [4,5]. Following the results reported in these papers, only the Multi-Layer Perceptron (MLP) is adopted in the present work. The input to the MLP is $\vec{X}$, the output is a softmax layer with C units, and training is performed by backpropagation using the standard cross-entropy loss. Under these conditions, it is well known that the each output of the MLP, c, approaches $P(c \mid \vec{X})$, $1 \le c \le C$. Thus Eq. (5) directly applies.

Such a CSC classifier is typically evaluated by its probability of error, estimated as the *Error Rate* k_e/K, where k_e is the number of wrong predictions made on a test set of K image documents from the same C classes considered for training [26].

4.1. Open set classification

In the practical application of the methods discussed in this paper, a complete set of classes (i.e., typologies of notarial deeds, such as Will, Debenture, etc.) is seldom known at the training time. Moreover, many of the classes represented in the available ground truth (GT) often contain just one, or maybe a few samples (deeds) which are hardly enough for training or testing. Clearly, these classes should be set aside in the above CSC paradigm. But, in practice, new image documents do arrive which need to be processed anyway and the classical CSC paradigm proves inadequate. Instead, our problem naturally falls under the so called *"Open Set Classification"* framework [15–17,23], where a larger number of (possibly unknown or uncertain) classes, $\tilde{C} > C$, is assumed to exist in $\mathcal{X}$.

Consider first a setup where the system can be trained with samples of all the C *known classes* plus an *additional* "REJECT*class*" which encompasses the remaining $\tilde{C} - C$ unknown classes. Clearly, all the GT classes with too few samples can be properly included in this "class". This is still a fairly traditional PR setting, which amounts to training and classification with $C' = C + 1$ classes [26]. Minimum error-risk classification is also given by Eq. (5), changing C with C', and the traditional *"Error Rate"* can still be reasonably used for OSC evaluation.

A different way to deal with test samples of *unknown classes* is to train the system using only samples of the *Cknown classes*. A threshold t is then needed to establish a class posterior probability below which any test sample should be rejected; i.e., considered to belong to a REJECT class. Formally, let $Q(\vec{X}) \stackrel{\text{def}}{=} \max_{1 \le c \le C} P(c \mid \vec{X})$. Then:

$$c^\star(X) = \begin{cases} \arg\max_{c \in \{1,\ldots,C\}} P(c \mid \vec{X}) & \text{if } Q(\vec{X}) \ge t \\ \text{REJECT} & \text{otherwise} \end{cases} \quad (6)$$

Following this scheme, several approaches can be used for OSC with REJECT and training with only the C known classes. In addition to directly using a MLP, trained with Tf·Idf input vectors from the C known classes as discussed at the beginning of this section, we have adapted the ideas of [23] and [22] to our OSC CBIDC task.

In the model proposed in [23], called "one versus rest" (1-vs-rest), the output layer of a neural network is configured as a vector of C *sigmoid* activation functions. That is, each output c corresponds to a Bernouilli distribution, $P(b_c \mid \vec{X})$, $1 \le c \le C$, where b_c is the value of a binary random variable, which is 1 if the class of X is c and 0 otherwise. Here, we have applied this idea to our MLP architectures by simply changing the SoftMax output layer (which corresponds to the categorical distribution $P(c \mid \vec{X})$), with a 1-vs-rest layer and using the corresponding C–variate binary cross-entropy loss for training, as in [23]. This model will be referred to as "binary-outputs MLP" (bMLP).

On the other hand, in [22] a Convolutional Prototype Network (CPN) is proposed as a general approach for OSC (and CSC alike). In that work, an input convolutional stack is devoted to feature extraction from the input objects which generally consist of simple (and single) images. In our CBIDC task an input consists of multiple (from a few to hundreds) complex handwritten text images – an input that a conventional convolutional stack would hardly be able to handle. But CBIDC feature extraction is already largely and satisfactorily solved by representing these sets of images with Tf·Idf vectors computed from image PrIx's, as discussed in Secs. 2 and 3. Therefore we kept the Tf·Idf input and MLP layers of our main approach and adopted from [22] only the prototype and output layers, along with the corresponding training rules. Such an architecture is called MLP-PN. As in the other approaches, two types of loss were used for MLP-PN: A classical discriminative loss, called Distance-based cross-entropy (DCE) [22] and a "One Vesus All" loss (OVA) [22], similar to the 1-vs-rest binary cross-entropy proposed in [23] which, as discussed above, we refer to as "binary-outputs

Fig. 1. Example of page images from JMDB_4949 and JMBD_4950.

MLP". We will hereafter refer to the resulting models as pMLP and bpMLP, respectively.

If a single, fixed threshold t can be assumed or somehow estimated,[4] both bMLP and pbMLP can straightforwardly implement OSC with REJECT, exactly as in Eq. (6), by assuming that $P(c \mid \vec{X}), 1 \leq c \leq C$, are the output probabilities yield by bMLP or pbMLP. Also, the OSC Error Rate can be straightforwardly measured for in the same way for MLP, bMLP and pbMLP.

Letting the user adjust the reject threshold is a convenient, practical option to help tailoring a trained system to the rejection needs of each specific batch of test data. To assess rejection performance in this scenario, a ROC curve [25] can be plotted to characterise the system for all the possible thresholds. The area under this curve, called AUROC, is a commonly accepted scalar measure that adequately assesses the system's overall performance for all reject thresholds. A ROC curve assumes binary decisions. In our case, the task is to decide whether a test deed is or is not from one of the C known classes.

5. Dataset and experimental settings

In this section, we provide details of the dataset and the empirical framework adopted for the experiments presented in Section 6. To allow reproducibility, we make publicly available all the required data and code.[5]

5.1. A Handwritten Notarial document dataset

The dataset considered in this work is a small part of a huge series of historical notarial documents held by the Spanish Archivo Histórico Provincial de Cádiz (AHPC). It consists of 16 849 manuscript bundles or "protocol books", containing in total more than 4.2 million deeds or files and 25 million pages.

50 of these bundles were included in the collection compiled in the Carabela project [3], where the corresponding PrIx's were also produced.[6] In the present work we selected two of these books, JMBD_4949 and JMBD_4950, dated 1723–1724. Fig. 1 shows examples of page images of these books.

Note that *no* typical GT annotations (such as text lines or transcripts) are available for these manuscripts. As explained below, only coarse-grained GT annotations aimed at bundle segmentation and deed classification were produced.

The bundles were manually divided into sequential segments or sections, each corresponding to a single deed, which was then annotated with a class label. A first section of about 50 pages, which form a kind of table of contents, was also identified in each book, but these sections were not used in the present work. It is worth noting that each deed may contain from two to dozens of pages, and separating these deeds is not straightforward. In future works, we plan to develop methods to also perform this task automatically but, for the present work, we take the manual segmentation as given.

Table 1
Number of documents and page images for JMBD_4949 and JMBD_4950: per class, per document & class, and totals.

Class	Deeds	Pages				
ID's		Avg	Min	Max	St-dev	Total
PA	240	3.3	2	24	3.5	803
LP	72	4.8	2	30	5.4	345
DB	44	4.8	2	32	5.6	212
LE	32	4.8	2	16	2.6	152
TE	29	8.6	4	48	9.4	248
SA	21	22.9	4	122	29.8	480
RI	17	4.0	4	4	0.0	68
CS	12	11.5	2	26	9.0	138
DP	10	3.8	2	8	1.9	38
ST	9	2.4	2	4	0.8	22
CN	6	5.3	2	14	3.9	32
TF	6	5.3	4	8	1.9	32
Reject	57	9.2	2	70	12.2	526
Total	555	5.6	2	122	9.2	3096

The experts found 95 deeds in JMBD_4949 and 260 in JMBD_4950, a total of 555 deeds, belonging to about 41 different types or classes. However, the classes of some deeds were not clear and, for many of the clearly identified classes, only very few deeds were available. To allow the classification results to be sufficiently reliable, only those classes having at least *one* deed in each book and *six* deeds in total were taken into account. This way, 498 deeds were retained from 12 classes considered sufficiently represented and all the other, belonging to 29 unclear or poorly represented classes, were collectively deemed to belong to a special "class" called REJECT (RJ).

The twelve well-represented classes are: Power of Attorney (*PA*), Letter of Payment (*LP*), Debenture (*DB*), Lease (*LE*), Testament (*TE*), Sale (*SA*), Risk (*RI*), Census (*CS*), Deposit (*DP*), Statement (*ST*), Cession (*CN*) and Treaty of Fact (*TF*). See details of this dataset are in Table 1.

The *Closed Set* machine learning task consists in training a model to classify a deed known to belong to one of the $C = 12$ proper classes into one of these same classes. The corresponding *Open Set* task is to also let the system reject samples (deeds) from the remaining 29 classes. That is, $\tilde{C} = 12 + 29 = 41$ and the proportion of known classes is 29.3%.

5.2. Empirical settings

PrIx's typically contain huge amounts of different pseudo-word hypotheses. However, many of these hypotheses have low relevance probability and most of the low-probability pseudo-words are not real words. Therefore, as a first step, entries with less than three characters, as well as those with too low RP ($P(R \mid x, v) < 0.1$), were pruned out. This reduced the original set of 809 787 different pseudo-words to a vocabulary V of 55 927 pseudo-words for the two bundles considered.

Then, as discussed in Section 3, the pseudo-words in V were sorted by decreasing IG and the first N entries were selected to define a BOW vocabulary V_N. Exponentially increasing values of N from 16 up to 16 384 were considered in the experiments.

Finally, a Tf·Idf N-dimensional vector was calculated for each deed, $X \in \mathcal{X}$. For experimental simplicity, Tf·Idf(v, X) was esti-

[4] In some of our experiments we have adopted the heuristic method proposed in [23], which can be used to estimate C different thresholds, one per class. However, since we have not observed any improvement by using multiple thresholds, in this work we stick with the simpler single-threshold setting.

[5] https://github.com/JoseRPrietoF/docClassPrIx

[6] The images of this collection and a search interface based on PrIx are available at http://prhlt-carabela.prhlt.upv.es/carabela

mated just once for each $v \in V$, using a normalising factor $E[n(X)]$ (see Eq. (3)) computed for all $v \in V$, rather than just $v \in V_N$ for every N considered in the experiments.[7]

Tf·Idf deed vectors were further normalised by subtracting the mean and dividing by the standard deviation, resulting in zero-mean and unit-variance input vectors.

Three MLP configurations with different numbers of layers were considered. In all the cases, every layer except the last one is followed by batch normalisation and ReLU activation functions [27]. The basic configuration was a plain C-class or C'-class perceptron where the input is totally connected to each of the $C = 12$ or $C' = 13$ neurons of the output layer (hence no hidden layers are used). For the sake of simplifying the terminology, here we consider such a model as a "0-hidden-layers MLP" and refer to it as MLP-0. The next configuration, MLP-1, was a proper MLP with one hidden 128-neurons layer. This layer was expected to do some kind of word clustering, hopefully improving the classification ability of the output layer. Finally, an MLP-2 and a bMLP-2, with two hidden 128-neurons layers, were tested. Deeper models were tried too, but they did not yield significant improvements.

The parameters of each MLP, bMLP, pMLP and bpMLP were initialised following [28] and trained according to the standard (MLP/pMLP) or binary (bMLP/bpMLP) cross-entropy loss for a minimum of 20 and a maximum of 500 epochs, applying early stopping with a patience factor of 50 epochs. For MLP-0, the RMSprop optimiser was used with a learning rate of 0.1, while for all the other models the optimiser was SGD [29], with a learning rate of 0.01. Specifically for pMLP and bpMLP, following [22],[8] a single prototype per class has been adopted and, after trying several sizes, the best results are presented for a prototype size of 128.

As discussed in Section 4.1, to use models trained only with the 12 known classes for OSC, a threshold t is required which has to be determined or somehow estimated. Two simple heuristic methods were considered.

The first one is that proposed in [23], which we compute as: $t = 1 - \sqrt{\sum_X (1 - P_{\hat{c}}(X))^2 / K}$, where the sum spans the $K = 498$ samples of known classes, $P_{\hat{c}}(X) = P(\hat{c}(X) \mid X)$ for MLP or $P_{\hat{c}}(X) = P(b_{\hat{c}(X)} \mid X)$ for bMLP, and $\hat{c}(X)$ is the correct class of X according to the GT.

The second, rather crude heuristic comes from the observation that the exact value of t is not critical, provided it is around the average values of the max class posteriors of the test samples (see Section 6.2). So we can just set the threshold to this average. While this estimate is based on *test* sample posteriors, it is totally fair, since the class labels are not used at all.

As suggested by Table 1 (Section 5.1), we consider all the deeds available in the bundles JMBD_4949 and JMBD_4950 as a single dataset. This arrangement is different from the one we adopted in [5], where each bundle was considered a (smaller) dataset by itself. Even though the number of samples is now much larger (498 in 12 classes for CSC and 555 in 41 classes for OSC), again they are not enough to establish a fixed training/test *partition*. Instead, as in [5], we adopt the *leaving one out* (LOO) protocol, which entails certain issues in some of the experimental procedures.

First, to simplify the computation of IG and Tf·Idf, the calculations were performed only once for all the classes and samples, because we have observed that leaving or not a single sample out hardly changes the results of these calculations significantly. Second, for computing the first of the above explained threshold estimates, the posterior probabilities of *all* the 498 samples $X \in \mathcal{X}$ of known classes have been used. While this simplification breaks to some extent the test-set independence principle, it should be noted that the values of these estimates are not critical, as will be discussed in Section 6.2.

[7] $E[n(X)]$ is the expected number of running words in V, which can be larger than the same estimate if only the words in V_N are considered (in the summation of the first equation of Eq. (2)). For every N, this normalising factor is thus the same for all the components of the Tf·Idf vectors, and it has not been observed to significantly affect the classification results.

[8] We adapted the code provided by the authors of [30], available from: https://github.com/YangHM/Convolutional-Prototype-Learning.

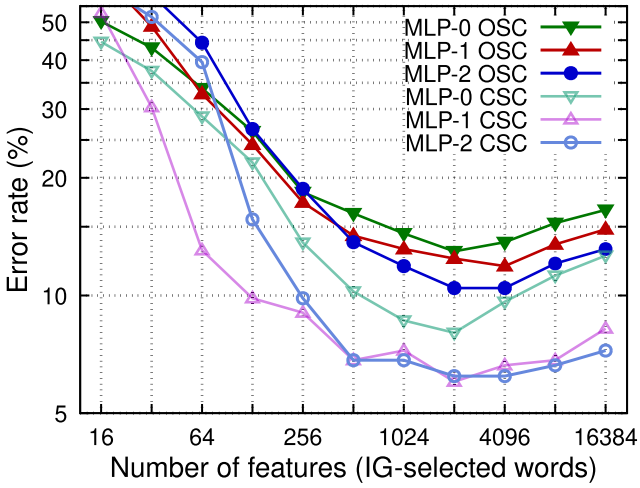

Fig. 2. Leaving-one-out classification error rate on JMBD_4949 and JMBD_4950 with three threshold-less MLP models, both for Closed and Open Set Classification. OSC: training and testing with 12 known classes; OSC: training and testing with 12 known plus REJECT (13 "classes"). All the results are based on PrIx document and word frequency estimates. 95% confidence intervals (not shown for clarity) are all smaller than ±4.4% and smaller than ±3.0% for all the error rates below 15%.

6. Experiments and results

Results using the methods presented in Section 4 and the dataset and empirical settings discussed in Section 5 are reported below. First we focus on CSC and also on OSC methods which rely on training with samples (deeds) of classes considered unknown to avoid the need of a reject threshold. The second subsection is devoted to OSC with models trained only with samples of known classes – which thereby requires a threshold to reject test samples deemed *not* to belong to any known class.

6.1. Threshold-less closed and open set classification

Fig. 2 shows two sets of results all obtained according to Eq. (5) (Section 4.1). First, traditional CSC results achieved using three MLP models trained and tested only with samples of the 12 known classes. Then OSC with the same models but now trained with samples of the 13 classes: 12 known proper classes plus a special REJECT "class", which includes samples from 29 additional classes. In both cases, results are shown for increasing dimension (number of IG-selected words) of the Tf·Idf image document embeddings.

CSC results are obviously better than those of their OSC counterparts. Under the traditional CSC framework, these results suggest that, using MLP-1 and 512 words (or more) for Tf·Idf representation, more than 93% of our image documents (deeds) could be automatically tagged with the correct classes.

MLP-2 yields the best OSC and CSC results, with input image documents embedded into a 2048-dimensional Tf·Idf vector space. The first column of Table 2 summarises these results.

Table 2 also reports comparable results using a bMLP-2 classifier (c.f., Section 4.1). Even though the bMLP-2 output layer and training loss do not aim to maximise class discrimination, this model achieves almost the same results as MLP-2. The classifica-

Table 2
Classification error rate of threshold-less methods. CSC: training and testing with 12 known classes; OSC: training and testing with 12 known plus a REJECT "class". Results are shown for n=2048 words and both PRIx image representations and plain text HTR image transcripts.

Images represented as:	PRIx				HTR
Classifier	MLP-2	bMLP-2	pMLP-2	pbMLP-2	MLP-2
CSC ($C=12$)	**6.2**	**6.2**	7.0	11.7	8.0
OSC ($C'=13$)	**10.5**	11.0	10.8	18.2	12.3

tion accuracy achieved is remarkable, given the complexity of the task: classify sets of images of untranscribed manuscripts (with as many as 122 images per set, see Table 1) into 12 (or 12+1) different classes, which only differ from each other in nuances characterised by subtle combinations of words.

Results for the prototype network models (MLP-PN) pMLP-2 and pbMLP-2 discussed in Section 4.1 are also included in this table. For pure CSC (12 classes), the results of pMLP-2 are comparable with those of MLP-2 and bpMLP-2, but the accuracy of the pbMLP-2 model, trained is a similar way as bMLP-2, is clearly lower. Results for these models trained with an additional REJECT class (OSC, $C'=13$) follow a similar tendency as all the other models, even though pbMLP-2 does not reach comparable accuracy.

For completeness, Table 2 also reports results obtained with exactly the same MLP-2 classifier, but using state-of-the-art *HTR image transcripts* [1,2], rather than PRIx, to represent the images. In this case, documents and word frequencies needed for IG and Tf·Idf were naively *computed* (using Eq. (1–4) of [4]) from the noisy plain-text HTR output. As expected, these results fall short of those obtained with the proposed approach, where document and word frequencies are *estimated* (rather than computed) using *PRIx image representations*.

Table 3 shows the confusion matrix and the error rate per class for MLP-2 OSC. It is worth noting that the REJECT "class" is involved in 38 out of the 58 total errors.

6.2. Threshold-based open set classification and rejection

Here models are trained only on samples of the 12 known classes, but the test set includes samples both from these 12 classes and also from the other 29 classes considered unknown. So the task entails both classification and rejection. OSC Error Rates are reported in Table 4. As in the previous subsection, these error rates include three types of errors: a) conventional known-class misclassification, b) rejecting samples from known classes and c) failing to reject samples from unknown classes. Given that no REJECT class is trained, OSC must follow Eq. (6) (Section 4.1), which requires a threshold t. Table 4 reports results for two fixed thresholds and for another two thresholds, estimated as discussed in Section 5.2. An "oracle threshold" is also included which was just determined as the one for which the test-set Error Rate was lowest.

Table 4
OSC classification + rejection bMLP-2 error rate for different thresholds (t), using PRIx and n=2048 words with the bMLP-2 model. It was trained with $C=12$ classes and tested with samples of all $\tilde{C}=41$ classes (12 known, plus 29 REJECT "classes"). 95% confidence intervals are within ±3.2%, or ±2.2% for the lowest error rate.

Threshold estimate	bMLP-2	(t)
Fixed 0.0	15.9	(0.00)
Fixed 0.5	16.4	(0.50)
$1-\sigma$ [23]	**6.5**	(0.75)
Avg. max class posterior	7.2	(0.94)
Best on test ("oracle")	6.5	(0.75)

Table 5
Rejection performance for bMLP-2 OSC with PRIx and n=2048 words. Training with $C=12$ classes, testing with samples of all $\tilde{C}=41$ classes. AUROC values (%) and rejection error rate (%) for various thresholds t.

Model	Threshold (t)							AUROC
	0.00	0.50	0.75	0.76	0.94	0.97	0.98	
bMLP-2	10.3	11.2	**2.2**	**2.2**	3.2	3.6	4.1	**98.3**

The four models whose results appear in Table 2 were tested in this threshold-based, full OSC scenario. The oracle-threshold OSC error rates achieved were as follows: MLP-2: 13.0%, bMLP-2: 6.5%, pMLP-2: 16.57% and pbMLP-2: 18.37%. Given the great superiority of bMLP-2, detailed results are shown in Table 4 only for this model.

Results with the two estimated thresholds are similar and close to the oracle. In fact, estimates are not critical for bMLP-2 because similar error rates are observed for any threshold in the range $[0.70, 0.97]$.

Overall we can conclude that bMLP-2 provides excellent accuracy in full, threshold-based OSC, very close to the best result achieved in basic CSC, but now including also the duty of rejecting samples from unknown classes.

Table 5 shows the AUROC result (see Section 4.1), which assess *rejection* performance taking into account all the possible thresh-

Table 3
Confusion matrix for PRIx MLP-2 OSC with $n=2048$ 2048.

JMBD_4949 & JMBD_4950															
	PA	LP	DB	LE	TE	SA	RI	CS	DP	ST	CN	TF	RJ	Total	Err (%)
PA	229	0	0	0	1	0	0	2	0	0	1	0	7	240	4.6
LP	2	66	2	0	0	1	0	0	0	0	0	0	1	72	8.3
DB	3	1	37	0	0	0	0	0	0	0	0	0	3	44	15.9
LE	1	1	0	29	0	0	0	0	0	0	0	0	1	32	9.4
TE	1	0	0	0	27	0	0	0	0	0	0	0	1	29	6.9
SA	0	0	0	0	0	19	0	0	0	0	0	1	1	21	9.5
RI	0	0	0	0	0	0	17	0	0	0	0	0	0	17	0.0
CS	0	0	0	0	0	0	0	8	0	0	0	0	4	12	33.3
DP	0	0	0	0	0	0	0	0	10	0	0	0	0	10	0.0
ST	0	0	0	0	2	0	0	0	0	5	0	0	2	9	44.4
CN	0	0	0	0	0	1	0	0	0	0	5	0	0	6	16.7
TF	0	0	0	0	0	0	0	0	0	0	0	6	0	6	0.0
REJECT	3	6	3	0	0	2	0	3	0	1	0	0	39	57	31.6
Total	239	74	42	29	30	23	17	13	10	6	6	7	59	555	(10.5)

olds. The table also shows the Error Rates of the corresponding binary classification task (REJECT – *not*-REJECT) for some thresholds. The rejection performance achieved by bMLP-2 is close to perfect, which explains the OSC superiority of bMLP-2 discussed above.

7. Conclusions

This work shows how to perform accurate content-based classification of untranscribed image documents (CBIDC). This task is challenging because each image document typically encompasses many images of handwritten text which are hard to read, even by humans. Our approach is cost-effective, because it does not need image transcripts. The only ground truth needed for model training is the class label of each training document and, once trained, the models provide accurate automatic CBIDC for new, also untranscribed multi-page image documents.

Our methods overcome the need of explicit transcripts by relying on *probabilistic indexing* (PrIx), a technology which provides robust representations of text images in terms of *textual* rather than *visual features*. We show that, using PrIx representations, our classification models consistently provide better results than using a popular, naive approach, where images are represented by their noisy automatic HTR transcripts.

Extending our previous works, here we report consolidated results using a sufficiently large set of image documents which belong to a rich set of classes. Our present study includes both the traditional classification viewpoint (CSC) and the "Open Set" (OSC) framework which is much more realistic and close to practical requirements.

Various OSC methods have been proposed or adopted and studied, all based on PrIx image representation and image documents embedding into a Tf·Idf vector space. Some methods follow the classical paradigm of training a CSC model with an additional class which collects samples of what would be "unseen classes". Other, more interesting approaches only need training with samples of known classes and use a rejection threshold on the class posterior probabilities of known classes. Our results clearly show that, among these later methods, the model referred to as bMLP greatly outperforms all the others, achieving a *combined classification and rejection accuracy* close to 94%.

According to the experts who annotated the GT data used in our experiments, this accuracy is close to the limit of human-labeling uncertainty. So we believe that no further efforts are deserved to improve the technology (the OSC methods in particular), until larger and more challenging data sets can be compiled and annotated with the required GT — a task that will certainly be expensive.

To deal with increasingly challenging types of image documents, we believe that the internal structure of the documents will need to be modeled. So, in future works we plan to explore other classification models, such as recurrent neural networks, that can account for the sequential regularities exhibited by textual contents in successive page images of formal documents.

So far, all our studies on CBIDC have assumed the image documents are given. However, in real applications, these documents are typically embedded into large bundles, without explicit separation of the specific page images encompassed by each document. Therefore, in future research works we also plan to develop new methods that allow not only to classify image documents, but also automatically segment large document bundles into the individual image documents they contain.

Finally, we know very well that our practical CBIDC OSC task is in essence *incremental* [16,31,32]. Therefore, we will certainly develop and/or adopt existing incremental learning techniques to provide final practical solutions to the CBIDC needs of archives and libraries which hold and manage large historical manuscript collections.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Work partially supported by : Universitat Politȩcnica de Valȩncia under grant FPI-I/SP20190010, Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121), by grant PID2020116813RBI00a of MCIN/AEI/ 10.13039/501100011033 and by a María Zambrano grant of the Spanish Ministerio de Universidades and the European Union NextGenerationEU/PRTR.

References

[1] J.A. Sánchez, V. Romero, A.H. Toselli, M. Villegas, E. Vidal, A set of benchmarks for handwritten text recognition on historical documents, Pattern Recognit. 94 (2019) 122–134.
[2] V. Romero, A.H. Toselli, E. Vidal, J.A. Sánchez, C. Alonso, L. Marqués, Modern vs diplomatic transcripts for historical handwritten text recognition, in: Int. Conf. on Image Analysis and Processing (PatReCH workshop), volume LCNS 11808, Springer, 2019, pp. 103–114.
[3] E. Vidal, V. Romero, A.H. Toselli, J.-A. Sánchez, V. Bosch, L. Quirós, J.M. Benedí, J.R. Prieto, M. Pastor, F. Casacuberta, et al., The carabela project and manuscript collection: large-scale probabilistic indexing and content-based classification, in: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2020, pp. 85–90.
[4] J.R. Prieto, V. Bosch, E. Vidal, C. Alonso, M.C. Orcero, L. Marquez, Textual-content-based classification of bundles of untranscribed manuscript images, in: 2020 25th Int. Conf. on Pattern Recognition (ICPR), IEEE, 2021, pp. 3162–3169.
[5] J.J. Flores, J.R. Prieto, D. Garrido, C. Alonso, E. Vidal, Classification of untranscribed handwritten notarial documents by textual contents, in: IbPRIA-22, Springer, 2022, pp. 14–26.
[6] A.H. Toselli, E. Vidal, V. Romero, V. Frinken, HMM Word graph based keyword spotting in handwritten document images, Inf. Sci. (Ny) 370–371 (2016) 497–518.
[7] T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A.H. Toselli, E. Vidal, Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project, in: 14th ICDAR, volume 01, 2017, pp. 311–316.
[8] E. Lang, J. Puigcerver, A.H. Toselli, E. Vidal, Probabilistic indexing and search for information extraction on handwritten german parish records, in: 16th ICFHR, 2018, pp. 44–49.
[9] J. Puigcerver, A Probabilistic Formulation of Keyword Spotting, Univ. Politȩcnica de Valȩncia, 2018 Ph.D. thesis.
[10] A.H. Toselli, V. Romero, J.A. Sánchez, E. Vidal, Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, in: Int. Conf. on Document Analysis and Recogn. (ICDAR), IEEE, 2019, pp. 108–113.
[11] A.H. Toselli, E. Vidal, J. Puigcerver, E. Noya-García, Probabilistic multi-word spotting in handwritten text images, Pattern Anal. Appl. 22 (1) (2019) 23–32.
[12] J.R. Prieto, E. Vidal, J.A. Sánchez, C. Alonso, D. Garrido, Extracting descriptive words from untranscribed handwritten images, in: Iberian Conf. on Patt. Recogn. and Image Analysis, Springer, 2022, pp. 540–551.
[13] S. Sevim, S.I. Omurca, E. Ekinci, Document image classification with vision transformers, in: M.N. Seyman (Ed.), Electrical and Computer Engineering, Springer International Publishing, Cham, 2022, pp. 68–81.
[14] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, L. Zhou, LayoutLMv2: Multi-modal pre-training for visually-rich document understanding, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2579–2591.
[15] W.J. Scheirer, L.P. Jain, T.E. Boult, Probability models for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2317–2324.
[16] C. Geng, S.-J. Huang, S. Chen, Recent advances in open set recognition: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3614–3631.
[17] A. Mahdavi, M. Carvalho, A survey on open set recognition, in: 2021 IEEE Fourth Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE), 2021, pp. 37–44.
[18] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, T. Naemura, Classification-reconstruction learning for open-set recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 4016–4025, doi:10.1109/CVPR.2019.00414.

[19] H. Huang, Y. Wang, Q. Hu, M.-M. Cheng, Class-specific semantic reconstruction for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. PP (2022), doi:10.1109/tpami.2022.3200384.
[20] L. Chambers, M.M. Gaber, DeepstreamOS: fast open-set classification for convolutional neural networks, Pattern Recognit. Lett. 154 (2022) 75–82.
[21] Y. Shu, Y. Shi, Y. Wang, T. Huang, Y. Tian, P-ODN: prototype based open deep network for open set recognition, 2019,
[22] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, C.-L. Liu, Convolutional prototype network for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022) 2358–2370.
[23] L. Shu, H. Xu, B. Liu, DOC: deep open classification of text documents, in: Proceedings of the 2017 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2911–2916.
[24] E. Vidal, A.H. Toselli, J. Puigcerver, A probabilistic framework for lexicon-based keyword spotting in handwritten text images, arXiv preprint arXiv:2104.04556 (2021).
[25] C.D. Manning, P. Raghavan, H. Schtze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[26] R.O. Duda, P.E. Hart, et al., Pattern Classification and Scene Analysis, volume 3, Wiley New York, 1973.
[27] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd Int. Conf. on Int. Conf. on Machine Learning - Volume 37, in: ICML'15, JMLR.org, 2015, p. 448456.
[28] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, J. Machine Learn. Research 9 (2010) 249–256.
[29] S. Ruder, An overview of gradient descent optimization algorithms 14 (2017) 2–3.
[30] H.-M. Yang, X.-Y. Zhang, F. Yin, C.-L. Liu, Robust classification with convolutional prototype learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3474–3482.
[31] J. Leo, J. Kalita, Incremental deep neural network learning using classification confidence thresholding, IEEE Trans. Neural Netw. Learn. Syst. (2021).
[32] E. Lopez-Lopez, X.M. Pardo, C.V. Regueiro, Incremental learning from low-labelled stream data in open-set video face recognition, Pattern Recognit. 131 (2022) 108885.