Noname manuscript No.
(will be inserted by the editor)

Multimodal Page Classification in Administrative Document Image Streams

Marçal Rusiñol · Volkmar Frinken · Dimosthenis Karatzas · Andrew D. Bagdanov · Josep Lladós

Received: date / Accepted: date

Abstract In this paper we present a page classification application in a banking workflow. The proposed architecture represents administrative document images by merging visual and textual descriptions. The visual description is based on a hierarchical representation of the pixel intensity distribution. The textual description uses latent semantic analysis to represent document content as a mixture of topics. Several off-the-shelf classifiers and different strategies for combining visual and textual cues have been evaluated. A final step uses an n-gram model of the page stream allowing a finer-grained classification of pages. The proposed method has been tested in a real large-scale environment and we report results on a dataset of 70,000 pages.

Keywords Digital mail room · multimodal page classification · visual and textual document description

1 Introduction

Big corporations and public institutions such as banks, insurance companies, city halls, or national health services along with their citizens and clients create massive amounts of documents –faxes, letters, forms, invoices, etc.– that more often than not have to be dealt with in a close to real-time manner. These are vital communications with clients, providers and other stakeholders that flow into, through, and out of the organization. Processing paper-based correspondence is a labor-intensive task. Letters are opened, read, sorted, routed and delivered. Depending on the contents, the contained documents are then forwarded to the appropriate recipient for the required action. The needs of the market have been the leading force behind a huge amount of research and development across the document life-cycle from digitization to image analysis and from indexing and classification to knowledge management, re-purposing and routing. The collective application of the above processes for the management of document flows at large scales is known as the Digital Mail Room.

Document Analysis research provides solutions for automating the screening process and determining the document type (whether invoice, contract, letter, etc.), and for extracting the relevant information from each document with minimal human intervention. This information is stored in appropriate databases for future querying and feeding outbound communications.

In this paper we present a page classification application tested in a banking workflow. When asking for a mortgage or a line of credit, the bank asks clients to deliver paperwork in bulk in order to study the viability of the financial transaction. Such paperwork contains tax forms, invoices, contracts, etc. Before analyzing the risk of this operation, the bank digitizes all this material and categorizes each page for forwarding to a specific analyst. This paper presents our proposed page classification system aimed at mitigating the load of manual effort devoted to page categorization.

M. Rusiñol · D. Karatzas · A.D. Bagdanov · J. Lladós
Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
E-mail: {marcal, dimos, bagdanov, josep}@cvc.uab.es

V. Frinken
Department of Advanced Information Technology
Kyushu University, Japan
E-mail: vfrinken@ait.kyushu-u.ac.jp

1.1 Related Work

Document image classification is a mature research topic and many different approaches have been proposed in the literature. The interested reader is referred to the survey papers on document image representation, retrieval and classification from Doermann [12] and Chen and Blostein [8]. Besides the supervised machine learning techniques used for classifying the incoming images (k-NN, decision trees, SVM, neural networks, etc.), the different methods can be categorized according to the used document representation.

First there are methods that define document classes in terms of visual similarity. The proposed descriptors normally use statistics computed over low-level features in order to encode how documents "look". For instance, Héroux et al. proposed in [24] a document descriptor that encodes in a hierarchical fashion pixel densities within a grid partition. Sidiropoulos et al. [38] proposed a similar descriptor that encodes the average of gray intensity over an adaptive grid. In [19], Gordo et al. proposed a document description based on multi-scale run-length histograms. Such simple descriptors are helpful when dealing with problems where documents from the same class are visually similar although their contents might change (e.g. forms).

More elaborate methods encode document similarity in terms of their structure. Structural features are obtained from either a logical or physical layout analysis of document pages. Physical layout analysis decomposes document images into blocks and document similarity can be expressed in terms of the spatial relationships among these blocks. For example, Bagdanov and Worring construct an attributed relational graph [3] in order to model the layout structure within a document genre. Cesarini et al. use $X-Y$ trees in [6] to both physically segment and describe the document types. In [16] Gaceb et al. use a hierarchical graph coloring strategy to simultaneously perform segmentation and physical description. As an example of the family of methods that describe documents in terms of the structure of logical elements, we cite the work by Dengel and Dubiel presented in [10]. In this case the document description encodes how logical elements are located and which are the spatial relationships among them. Structural descriptors are much more robust for assessing visual similarity among document classes than image-based methods. However, they have the drawback that computing a mapping between two layout structures is computationally expensive.

Finally, document classes can be defined in terms of content similarity [1]. After a complete transcription of the documents, document similarity can be expressed by means of textual content. In general, text documents are represented as a set of words together with their associated frequencies in each document, which is known as the bag-of-words model. Because of its simplicity for classification purposes, most text classification methods use the bag-of-words representation, combined with a wide variety of different classifiers [39,36]. Refinements of the bag-of-words model for both feature selection (e.g. [40]) and feature transformation (e.g. latent semantic analysis [9], probabilistic latent semantic analysis [25], and latent Dirichlet allocation [5]) have been proposed in order to model document contents in terms of a mixture of topics. Textual content descriptors are of course suitable when dealing with document classes that "talk" about the same topic although the visual appearance of documents within a class may differ. In the specific case of administrative documents such as invoices, some ad-hoc document representations in terms of discriminative keywords have been proposed in the literature [27,22]. Such domain-specific representations offer high classification performances but are usually not generalizable to broader document collections.

In our application, we deal with document classes where visual similarity is strong evidence (such as forms and invoices from the same provider) and classes that exhibit strong textual content similarity and no predefined standard look-and-feel (such as audit reports or contracts). This fact motivates the use of an architecture that combines multiple modalities. However, very few attempts on fusing different information modalities for document image classification can be found in the literature. For instance the work by Erol and Hull presented in [14] achieved a semantic classification of administrative documents by merging features from different domains, namely, textual, color, handwriting or layout features. More recently, the paper presented by Augereau et al. in [2], inspired byu our previous work [32], proved the success of the combination of visual and textual features for administrative document classification.

On the other hand, there are very few works in the literature dealing with multipage documents [18,32] or with the treatment of image streams [21]. We strongly believe that in digital mailroom scenarios in which documents are often digitized in bulk, the use of the context of which pages come before or after the others is strong evidence to exploit when categorizing individual pages.

1.2 Contributions

In this paper we present a page classification application in a banking workflow. The proposed strategy represents administrative document images by merging vi-

sual and textual descriptions. Several off-the-shelf classifiers and combination strategies have been tested.

The main contributions of the paper are twofold. First, we describe a multimodal representation of administrative document images. Very few attempts at combining different views of document images have been studied in the literature. We show how the combination of both modalities clearly outperforms the exclusive use of either visual or textual information. In addition, we present an exhaustive analysis of the performance of different state-of-the-art classifiers and combination strategies. The second contribution is an n-gram model of the sequential distribution of different pages as they appear in the processed stream, yielding an extra improvement on the final page classification. The proposed method has been tested in a real large-scale environment and we report results of experiments on 70,000 pages.

The remainder of this paper is organized as follows. In Section 2 we present an overview of the proposed architecture. Section 3 is devoted to the multimodal document image description strategy. In Section 4 we briefly describe the classifiers and the combination schemes we used. The addition of sequential information by means of an n-gram model is detailed in Section 5. Experimental results are presented in Section 6. Finally, conclusions and further research lines are discussed in Section 7.

2 System Overview

Our proposed system architecture is illustrated in Fig. 1. Given a flow of incoming documents, both visual and textual descriptors and specific classifiers are computed in parallel. The visual modality encodes the appearance of the document image in terms of pixel densities and the classifier outputs the probabilities of belonging to each document class. Regarding the textual modality, a commercial OCR engine transcribes document images. After some pre-processing steps, a bag-of-words representation of the document is projected onto a topic space by means of Latent Semantic Analysis (LSA). Finally, the textual classifier outputs class probabilities as well. Later, a combination step weights the influence of visual and textual cues in order to output the individual page classification. An n-gram model of the page stream is finally used. A rejection threshold is set on the confidence value of the classification in order to refuse categorizations with weak evidence.

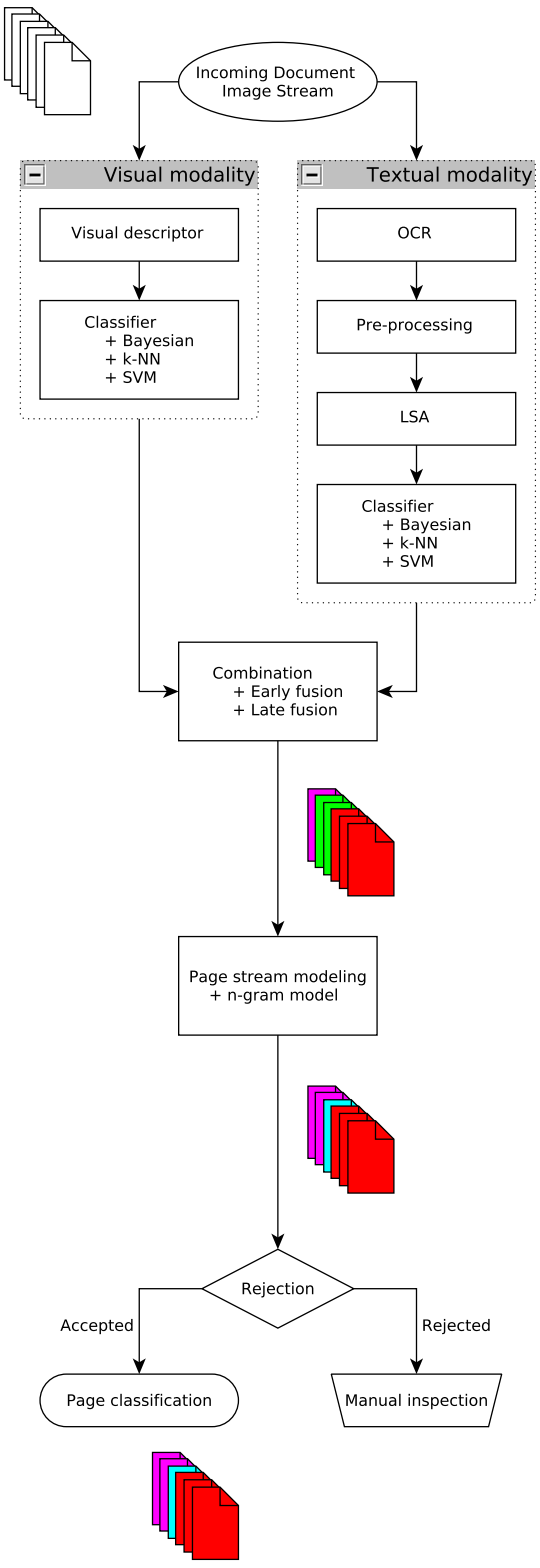


Fig. 1: Overview of the proposed system architecture.

3 Multimodal Description of Document Images

Our proposed description method combines textual information with a global visual document image descrip-

tion. We first detail the visual description of document images and then the textual content descriptor.

3.1 Visual Description

Within the document analysis and retrieval literature, many descriptors encoding the visual appearance of document images have been proposed. In this paper we use a simple description of documents that encodes pixel densities at different scales. In order to remove small details and noise from the incoming images, a Gaussian smoothing operator is used to blur the images before computing the visual descriptor.

We use the multiscale descriptor presented by Héroux et al. in [24] which encodes pixel densities at different locations and at different scales. This descriptor, although extremely simple and efficient to compute, has proven to yield very competitive results [24,17] when compared to more elaborate structural descriptors. Besides being discriminative it also tolerates slight skew deformations. Each document image is recursively split into rectangular regions to form a pyramid. In each region the pixel density is computed and stored in the corresponding position of the feature vector. We can see an example of the first levels of the pyramid in Figure 2. In our experimental setup, we use four scale levels, yielding an 85-dimensional visual descriptor $\mathbf{f}_v$. The visual feature vectors are finally normalized by their L_2-norm.

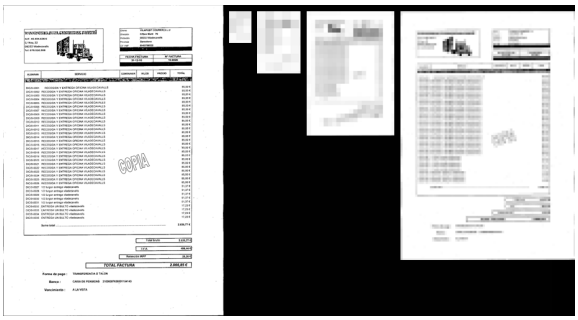


Fig. 2: Example of the multiscale pixel density descriptor.

3.2 Textual Description

In order to use textual information as another source to perform classification, we have used latent semantic analysis. The document images are OCRed with the commercial OCR from ABBYY[1]. Given the ASCII representations of each document image, a text preprocessing step is applied. Then the use of LSA overcomes some of the problems of directly using a bag-of-words model since it adds semantic coherence to the obtained results. We begin by detailing the preprocessing steps.

3.2.1 Text Preprocessing

Before extracting the textual descriptor for each document image, we apply several off-the-shelf preprocessing methods that help increase the robustness of the obtained textual description.

The first preprocessing step is to reduce inflected and derived words to their root in order to treat them equally. This process is known as stemming. We have used the Spanish version of the Porter stemming algorithm implemented in the Snowball [29] system. Then, stopword filtering is applied to eliminate very common words that do not convey any semantic information. In our experimental setup we end with a dictionary containing nearly 600,000 terms.

Finally, we represent each document image by its bag-of-words vector $\mathbf{f}_t$. Each $\mathbf{f}_t$ is then weighted by applying the tf-idf model [33]. This normalization emphasizes the terms that are frequent in a particular document and infrequent in the complete document corpus. The tf-idf weighting scheme assigns to each term t a weight in the document d given by

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t, \tag{1}$$

where $\text{tf}_{t,d}$ is the term frequency, i.e. the number of occurrences of term t in document d, and idf_t is the inverse document frequency computed as

$$\text{idf}_t = \log\left(\frac{N}{\text{df}_t}\right), \tag{2}$$

where N is the total number of documents and the document frequency df_t corresponds to the number of documents in the collection that contain the term t.

3.2.2 Latent Semantic Analysis

The classic bag-of-words model has some shortcomings. Often, the words appearing in the document to be classified are not the same as those from the documents in the corpus. Users in different contexts often use different words to describe the same information. This phenomenon is known as *synonymy*. The problem of *synonymy* often hinders the accuracy of classifiers. In addition, the dimensionality of bag-of-words representations might explode when dealing with large datasets and it is often advised to somehow reduce the size of the representation. In order to overcome this problem,

[1] ABBYY Finereader Engine 9

Deerwester et al. introduced in [9] the latent semantic analysis technique. The motivation of using LSA is that, given a text classification framework, it is able to classify documents that are conceptually similar in meaning in a given class, even if they do not share a significant set of words among them, while also drastically reducing the dimensionality of the feature vector.

The LSA model assumes that there exists some underlying semantic structure in the descriptor space. This semantic structure is defined by assigning to each document a set of topics, which can be estimated in an unsupervised way using standard statistical techniques. The goal is to obtain a transformed space where documents having similar topics but with different terms will lie close.

From our corpus, we select a subset of documents that will be used as the training set to build the LSA space. We represent the training set with a term-by-document matrix $\mathbf{A} \in \mathbb{R}^{M \times Q}$, where M is the number of different terms and Q is the number of documents in the training set. The transformed space is obtained by decomposing the training term-by-document matrix into three matrices by a truncated Singular Value Decomposition (SVD). In order compute the truncated SVD aiming to reduce the descriptor space to T topics we proceed as follows:

$$\mathbf{A} \simeq \hat{\mathbf{A}} = \mathbf{U}_T \mathbf{S}_T \left(\mathbf{V}_T\right)^\top , \tag{3}$$

where $\mathbf{U}_T \in \mathbb{R}^{M \times T}$, $\mathbf{S}_T \in \mathbb{R}^{T \times T}$ and $\mathbf{V}_T \in \mathbb{R}^{Q \times T}$. In our experimental setup, we use a value of $T = 300$ topics, which offers a good tradeoff between the descriptor's dimensionality and the achieved discriminative power.

When encoding the whole corpus, each feature vector $\mathbf{f}_t$ is projected to the topic space in order to obtain the topic descriptor $\hat{\mathbf{f}}_t$ by

$$\hat{\mathbf{f}}_t = \mathbf{f}_t^\top \mathbf{U}_T \left(\mathbf{S}_T\right)^{-1} . \tag{4}$$

Finally, each topic descriptor $\hat{\mathbf{f}}_t$ from the corpus is normalized using the L_2-norm. In this work we used the LSA implementation technique proposed by [31]. This technique introduces a streamed distributed algorithm for incremental SVD updates which has the advantage that it does not need a single-pass matrix decomposition algorithm that operates in constant memory with regard to the collection size. It presents and important advantage when dealing with large data collections.

4 Multimodal Classification

For the sake of completeness, in our experiments we have tested a number of different off-the-shelf classifiers and different strategies for combining visual and textual cues. We first briefly enumerate the classifiers used in our study and then the combination approaches.

4.1 Supervised Classifiers

In order to provide a thorough analysis of classifier performance, we have chosen to test classifiers from three different families. We evaluate a distance-based classifier, a number of Bayesian classifiers, and finally some kernel classifiers.

To represent distance-based classifiers, we used a simple k-nearest neighbor (k-NN) classifier. The cosine distance is used between the input and the training samples.

The second family of tested classifiers are probabilistic ones [13] based on Bayes' theorem. Within this family we have used the naïve Bayes classifier (NB) and both linear and quadratic Bayes normal classifiers (LDC and QDC respectively).

Finally, we have tested kernel classifiers. We have chosen the support vector machines classifier with three different kernels [28]: the radial-basis function kernel, the χ^2 kernel and the histogram intersection kernel (SVM-RBF, SVM-χ^2 and SVM-HI, respectively).

For the k-NN and the Bayes classifiers we have used the implementations in the PRTools [23] package. For the SVM classifiers we have used the LibSVM [7] library.

4.2 Combination of Visual and Textual Classifications

For each visual and textual page description, we obtain feature vectors $\mathbf{f}_v$ and $\hat{\mathbf{f}}_t$ describing the pages. Separate visual and textual classifiers are trained to test which is the most pertinent information. In addition, in order to combine both information cues, in our experiments we have tested an early fusion and four different late fusion strategies:

- **Early fusion** encodes the documents in a *single* histogram descriptor. Both features vectors $\mathbf{f}_v$ and $\hat{\mathbf{f}}_t$ are concatenated into a single feature vector $\mathbf{f}_{vt}$ which is L_2-normalized again. Here, a new classifier handling combined feature vectors has to be trained.
- **Late fusion** strategies perform individual page classifications for each of the information cues and then combine the classifier outputs into a single probability vector. Each visual and textual classifier outputs an *a posteriori* probability vector P_v and P_t, respectively, and the late fusion strategies combine them into a single probability vector P_{vt} identifying the most probable class both in terms of visual and

textual evidences. We have tested four different late fusion approaches:

- **SUM** combines the probability vectors by adding them with a previous power normalization step.

$$P_{vt} = P_v^i + P_t^j, \text{ with } i, j \in [0, 1].$$

- **PROD** multiplies the probability vectors after also applying a power normalization factor to each of them.

$$P_{vt} = P_v^i \times P_t^j, \text{ with } i, j \in [0, 1].$$

- **MAX** computes the fusion by taking the maximum of the probability vectors after a power normalization step.

$$P_{vt} = \max\left(P_v^i, P_t^j\right), \text{ with } i, j \in [0, 1].$$

- **LOG** uses logistic regression to achieve a single combined probability vector.

$$P_{vt} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 P_v + \beta_2 P_t)}}, \text{ with } \beta_0, \beta_1, \beta_2 \in [0, 1].$$

5 Modeling the page stream with *n*-grams

In a document stream, individual documents are not independent of each other and contextual information can be used to increase classification accuracy. We have used an n-gram model similar to the ones proposed in [35,26]. In order to include contextual information, the conditional occurrence probabilities are estimated on the training set. Assume we have a document stream $\hat{d} = d_1, \ldots, d_D$, where each document of a certain type $c(d_i) \in \{1, \ldots, C\}$ is to be classified. Then, the n-gram probability of a document is defined as the probability of the document d_i being of a certain type a conditioned on the types of the $n-1$ previous documents

$$p_s\left(c(d_i) = a \,|\, c(d_{i-n+1}), \ldots, c(d_{i-1})\right) \ .$$

With this is mind, the goal is to find the sequence of document classes $\hat{c} = c_1, \ldots, c_D$ that maximizes at the same time the class probabilities according to the individual classifiers

$$\prod_{i=1}^{D} p_s(c(d_i) = c_i)$$

as well as the probability according to the n-grams

$$\prod_{i=1}^{D} p_s\left(c(d_i) = c_i \,|\, c(d_{i-n+1}), \ldots, c(d_{i-1})\right) \ .$$

This can be solved with a token passing algorithm, similar to those used for speech or handwriting recognition with hidden Markov models [41]. In such a token passing algorithm, each token represents a certain classification hypothesis of the documents from the beginning up to a certain point. A token ϑ is defined by three values: $\vartheta.p$ is the classification probability of the followed hypothesis, $\vartheta.h$ is the history, which is a link to a token at a previous time step, and $\vartheta.c$ is the class of the current document.

The algorithm is initialized by different classification hypotheses for the first document which are stored in separate tokens in a list L_1. Then, the algorithm iterates over all documents in the sequence. For a time step t, a token ϑ in list L_{t-1} is used to generate C new tokens in list L_t. Such a new token ϑ' stands for the hypothesis that the documents $d_1 \ldots d_{t-1}$ are classified according to token ϑ and the new document d_t is classified as $c(d_t) = j$. The values of the new token are therefore

$$\begin{aligned} \vartheta'.c &= j \\ \vartheta'.p &= \vartheta.p \cdot p_s(c(d_t) = j) \cdot p_s\left(c(d_t) = j \,|\, \cdots\right) \\ \vartheta'.h &= \vartheta \end{aligned} \tag{5}$$

To keep the runtime from becoming exponential, only the best N tokens are kept in each list before the next list is created. The final class sequence can be retrieved in backward order, starting from the token with the highest probability in the last list L_D and following the links back to the beginning. An illustration of the algorithm is given in Fig. 3.

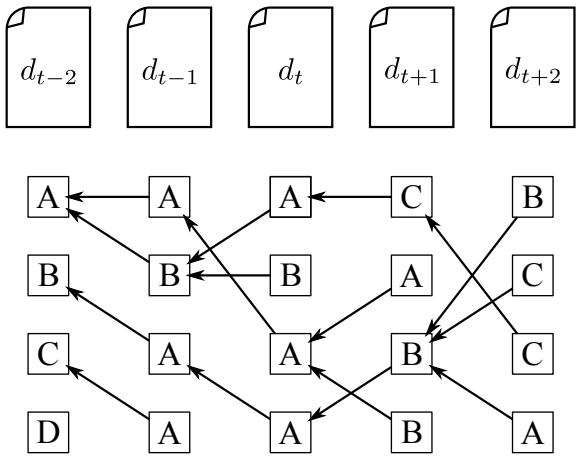


Fig. 3: Illustration of the token passing algorithm. For each document the ordered list with recognition hypotheses is given. The final classification is $BAABB$

6 Experimental Results

We first validate the visual and textual descriptors using public document image datasets, then we introduce our in-house dataset, the evaluation measures and the experimental framework used to assess the performance of the proposed system and then analyze the obtained results.

6.1 Descriptor Evaluation using Public Datasets

In order to validate the proposed visual and textual features, we have run a simple classification experiment over the NIST Tax Forms Dataset (SPDB2) [11] and the MARG [15] medical papers dataset. The NIST dataset contains 5590 tax form images spread over 20 different categories, whereas the MARG dataset consists of 1553 first pages of medical papers categorized into 9 different layout classes. Here visual and textual features were used alone and we used a 1-NN classifier in a one-versus-all setup. The obtained results are given in Table 1.

We observe that the proposed features are really performant on the NIST dataset; both visual and textual representations perfectly classified the documents in a one-versus-all setup. On the MARG dataset, which is ground-truthed at the layout level (i.e. two documents are considered from the same class if they share the same layout structure), obviously a textual content representation fails to achieve good classification accuracies whereas the visual representation is more suitable in those scenarios (although it does not reach the classification accuracies of pure layout descriptors).

6.2 In-house Dataset

Our dataset consists of nearly 70,000 real document images sampled from a banking workflow. This corresponds to a portion of certain types of documents received during two months. This dataset contains 13 different document classes which have been manually labeled.

Note that the class labels are defined in terms of *documents*. Since the incoming documents are multipage, this requires that different *pages* from the same document share the same label although they might look different and contain different textual content. Examples of such documents are given in Figure 4.

In order to train the classifiers and validate the different parameters involved in our architecture, we have split this collection into two different sets, each corresponding to a month of documents. The training set consists of 38,313 images and the test set has 31,424 images.

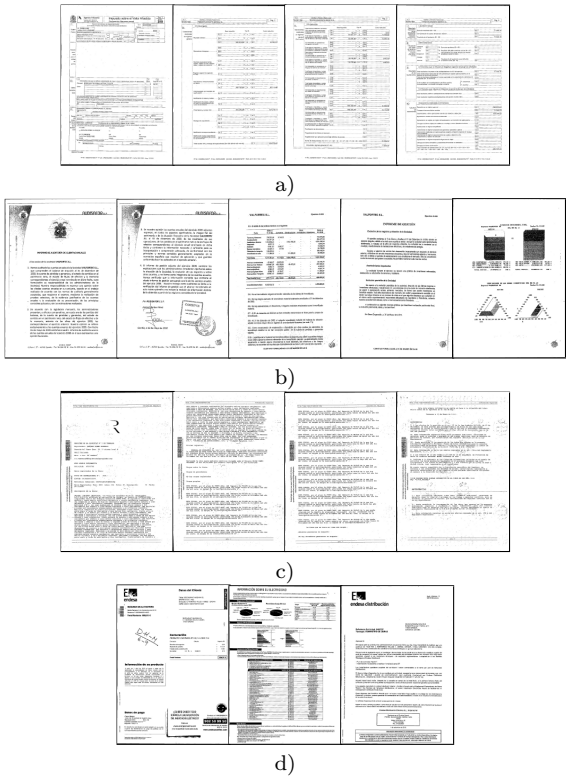


Fig. 4: Example of four classes from our mailroom stream. a) Tax forms, b) balance sheets, c) property registry records and d) Proofs of investment (usually invoices).

Over the training set we have used a 10-fold cross-validation strategy to train the classifiers and validate the different parameter values such as the k in the k-NN classifier, c and γ for the SVMs, the power normalization factors i and j for the late fusion strategies, and so on.

In order to measure the performance of the system, we report the classification accuracy rates for all the classifiers, information cues and combination strategies. In addition, in order to evaluate the rejection ability of the system, we report the accuracy-coverage plot which is an indicator of the accuracy evolution as we keep on rejecting classifications.

6.3 Results

In Table 2 we present the classification accuracies obtained by the different page representations, the different classifiers and combination strategies. If we first look at the independent visual and textual classification, we see that in our use case the textual features

Table 1: Classification rates for the NIST and MARG datasets.

	NIST		MARG	
	Visual	Textual	Visual	Textual
Our proposal	100	100	83.64	29.23
[34]	99.82		-	
[37]	99.70		-	
[20]	100		94.78	
[4]	-		92.60	
[30]	-		97.50	

Table 2: Classification rates for different classifiers and different combination strategies. The best performance without combination, with the early fusion an late fusion are highlighted.

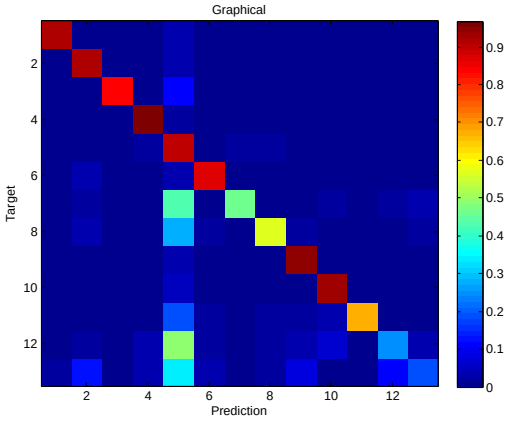
		LDC	QDC	NB	10-NN	SVM-RBF	SVM-χ^2	SVM-HI
No Combination	Visual	72.90	77.98	63.03	82.77	**85.54**	83.25	81.36
	Textual	86.79	82.81	81.05	92.11	**94.78**	94.57	94.67
Early Fusion		64.77	75.07	65.33	89.10	93.78	92.64	**94.30**
Late Fusion	SUM	90.34	91.95	84.06	-	95.48	95.12	95.18
	PROD	90.41	91.05	84.51	-	**95.49**	95.13	95.19
	MAX	89.70	90.85	81.43	-	95.30	94.86	94.85
	LOG	89.74	90.93	81.71	-	95.41	94.99	95.04

clearly outperform the visual representation no matter which classifier we use. The best results are obtained with the SVM classifier using a radial basis kernel for both the visual and textual representations. These results indicate that, in our scenario, textual content is more discriminative than visual page appearance. Regarding the classifiers, the Bayesian ones perform much worse than the kernel-based ones, although the k-NN classifier performs respectably.

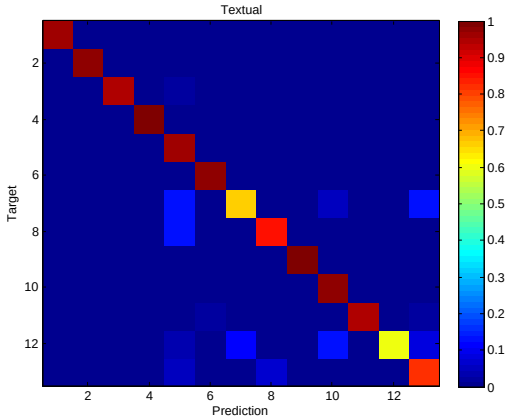
Concerning the early fusion combination, we observe that the best classifier is the SVM using the histogram intersection kernel. Again, Bayesian classifiers do not compare to the performance achieved with the kernel-based ones. However, it is worth noting that the early fusion strategy is not a good option in our scenario. No matter the classifier, better performance is obtained when using the textual description alone than when trying to combine the visual and textual page descriptors with early fusion.

Late fusion experiments for the k-NN classifier were not carried out since it does not output an *a posteriori* probability vector. For the rest of the classifiers, we see that no matter which late fusion strategy we use, we obtain better performance than the use of the single modalities alone. This means that, even if the textual representation outperformed the visual one when used alone, the addition of visual information helps to disambiguate some cases where the textual information is not pertinent. We have conducted a paired-sample t-test at the 0.01 significance level and the gain in performance when combining modalities against the performance of the textual description alone was statistically significant for all the classifiers. Except for the QDC classifier, the best performance is obtained when using the PROD late fusion combination strategy. Here again the best classifier is the SVM with the radial basis function kernel, although the difference with the other kernel-based classifiers and fusion strategies is not significant. There is however an important improvement when compared with the Bayesian classifiers.

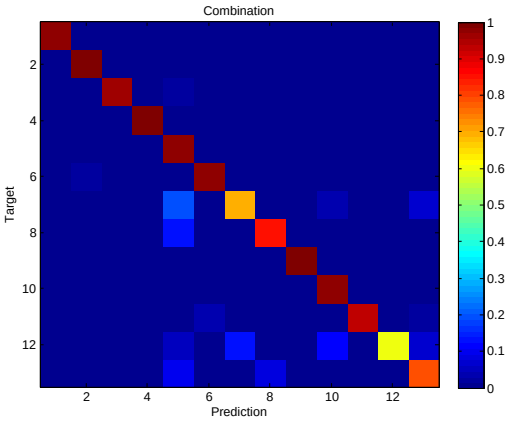
In Fig. 5 we report the confusion matrices for the 13 classes when using the visual, the textual modalities alone and the late fusion PROD combination with the SVM-RBF classifier. Observe how the visual classifier misclassifies many more elements than the textual or the combined approaches. We also see that the main confusions are between just a few document classes (e.g. classes 7, 8, 11, 12 or 13), and that the use of textual information attenuates this. Such failure cases correspond to classes that are visually heterogeneous, such as invoices, receipts or audit reports. The slight improvement with respect to the textual representation when combining textual and visual information is inappreciable in the confusion matrices, but is present in eight of the thirteen classes. In fact, these eight classes are the ones in which the visual descriptor provides acceptable results. We conclude that for classes that can be either visually or textually represented, the use of combined features provides better accuracies. Whereas for classes in which one of the modalities is not really



a)



b)



c)

Fig. 5: Confusion matrices obtained for the a) visual, b) textual and c) multimodal experiments when using an SVM with the RBF kernel and the PROD late fusion strategy.

suitable (visually representing an invoice class or textually representing a class sharing a certain layout) the combination of textual and visual modalities hinders the final performance against the sole use of the most suitable modality.

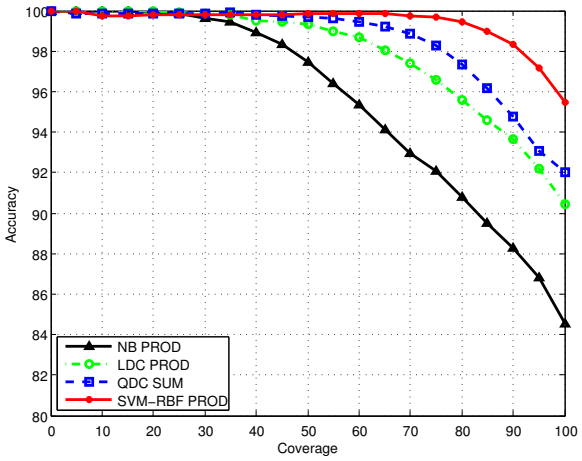


Fig. 6: Coverage-Accuracy plot when using the rejection strategy.

We report in Fig. 6 the coverage versus accuracy plots for different classifiers when using a rejection threshold over the final probability issued by the combination of classifiers. Depending on the application scenario, the threshold can be set to a low value in order to provide low rejection rates at the risk of accepting falsely classified pages. Or it can be increased in a conservative fashion, and in such a scenario just highly confident classifications are accepted and thus rejecting a significant number of positive samples. The coverage accuracy plot shows the accuracy drop when the system is asked to provide an answer for more and more elements from the collection (i.e. as the rejection threshold is set from conservative to more tolerant values). Again, the SVM classifier yields the best performance. For instance, in order to reach a 98% classification accuracy, the system with the SVM classifier rejects approximately 10% of the incoming pages whereas the naïve Bayes classifier rejects more than 50%. We also appreciate the significant drop in accuracy for the same document coverage. Finally, for the most rigid and critical scenarios, we see that the proposed method is able to reach a 100% recognition rate by just rejecting 35% of the pages and forwarding them to manual inspection.

In Table 3 the classification rates when using document n-grams after classification with the SVM-RBF classifier are given. The results show that the inclusion of n-gram probabilities increases recognition accuracy both when using the textual and visual features alone

Table 3: Classification rates using different n-gram document probabilities together with the SVM-RBF classifier with different features. Again, the best performance for each feature are highlighted.

used n-gram	Visual Features	Textual Features	Combined Late Fusion
2-grams	88.53	96.38	96.83
3-grams	**88.66**	96.47	96.90
4-grams	**88.66**	**96.51**	96.93
5-grams	88.57	96.44	**96.84**
6-grams	88.50	96.39	96.82
7-grams	84.26	93.48	94.33

and in the combined scenario. In particular, the best results are obtained when using the late fusion strategy in combination with 5-grams, where the accuracy is boosted from 95.49% to 96.84%.

7 Conclusions

In this paper we have presented a page classification application tested in a banking workflow. The proposed architecture represents administrative document images by merging two different modalities, namely visual and textual descriptions. The visual description relies on the pixel intensity distribution whereas the textual description uses latent semantic analysis to represent document content as a mixture of topics. The performance of Bayesian, distance-based and kernel-based classifiers has been analyzed. Early and late fusion strategies aimed at combining the visual and textual cues have also been compared. A final step modeling the page stream by means of an n-gram model has been used to refine the individual page classification. The proposed method has been tested in a real large-scale environment and we report results on experiments with 70,000 pages.

Acknowledgements This work has been partially supported by the Spanish Ministry of Education and Science under projects TIN2011-24631, TIN2012-37475-C02-02, RYC-2009-05031 and RYC-2012-11776; by the People Programme (Marie Curie Actions) of the Seventh Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ; and by the CREST project from Japan Society for the Promotion of Science (JSPS).

References

1. Aggarwal, C., Zhai, C.: Mining Text Data, chap. A Survey of Text Classification Algorithms. Springer (2012)
2. Augereau, O., Journet, N., Vialard, A., Domenger, J.: Improving classification of an industrial document image database by combining visual and textual features. In: Proceedings of the Eleventh IAPR International Workshop on Document Analysis Systems (2014)
3. Bagdanov, A.: Fine-grained document genre classification using first order random graphs. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 79–83 (2001)
4. van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: Proceedings of the International Conference on Document Image Analysis for Libraries (2006)
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
6. Cesarini, F., Lastri, M., Marinai, S., Soda, G.: Encoding of modified X-Y trees for document classification. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 1131–1136 (2001)
7. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3), 27:1–27:27 (2011)
8. Chen, N., Blostein, D.: A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal on Document Analysis and Recognition **10**(1), 1–16 (2006)
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6), 391–407 (1990)
10. Dengel, A., Dubiel, F.: Computer understanding of document structure. International Journal of Imaging Systems and Technology **7**(4), 271–278 (1996)
11. Dimmick, D., Garris, M., Wilson, C.L.: Structured forms database. Tech. rep., National Institutte of Standards and Technology (1991)
12. Doermann, D.: The indexing and retrieval of document images: A survey. Computer Vision Image Understanding **70**(3), 287–298 (1998)
13. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley-Interscience (2000)
14. Erol, B., Hull, J.: Semantic classification of business images. In: Electronic Imaging, pp. 60,730G–60,730G (2006)
15. Ford, G., Thoma, G.: Ground truth data for document image analysis. In: Proceedings of the Symposium on Document Image Understanding and Technology, pp. 199–205 (2003)
16. Gaceb, D., Eglin, V., Lebourgeois, F.: Classification of business documents for real-time application. Journal of Real-time Image Processing (2011). DOI 10.1007/s11554-011-0227-4
17. Gordo, A., Gibert, J., Valveny, E., Rusiñol, M.: A kernel-based approach to document retrieval. In: Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems, pp. 377–384 (2010)
18. Gordo, A., Perronnin, F.: A bag-of-pages approach to unordered multi-page document classification. In: International Conference on Pattern Recognition, pp. 1920–1923 (2010)
19. Gordo, A., Perronnin, F., Valveny, E.: Document classification using multiple views. In: Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems, pp. 33–37 (2012)
20. Gordo, A., Perronnin, F., Valveny, E.: Large-scale document image retrieval and classification with runlength histograms and binary embeddings. Pattern Recognition **46**(7), 1898–1905 (2013)
21. Gordo, A., Rusiñol, M., Karatzas, D., Bagdanov, A.: Document classification and page stream segmentation for digital mailroom applications. In: International Conference on Document Analyis and Recognition (2013)

22. Hamza, H., Belaïd, Y., Belaïd, A., Chaudhuri, B.: An end-to-end administrative document analysis system. In: Proceedings of the Fourteenth International Conference on Pattern Recognition, pp. 175–182 (2008)
23. van der Heijden, F., Duin, R., de Ridder, D., Tax, D.: Classification, parameter estimation and state estimation - an engineering approach using Matlab. John Wiley & Sons (2004)
24. Héroux, P., Diana, S., Ribert, A., Trupin, E.: Classification method study for automatic form class identification. In: Proceedings of the Fourteenth International Conference on Pattern Recognition, pp. 926–928 (1998)
25. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the Twentysecond Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
26. Meilender, T., Belaïd, A.: Segmentation of continuous document flow by a modified backward-forward algorithm. In: Proceedings of the Document Recognition and Retrieval (2009)
27. Misue, K., Sakakibara, Y.: Building of a document classification tree by recursive optimization of keyword selection function. US Patent US5463773 A (1995)
28. Odone, F., Barla, A., Verri, A.: Building kernels from binary strings for image matching. IEEE Transactions on Image Processing **14**(2), 169–180 (2005)
29. Porter, M.: Snowball: A language for stemming algorithms (2001)
30. Rangoni, Y., Belaïd, A., Vajda, S.: Labelling logical structures of document images using a dynamic perceptive neural network. International Journal on Document Analysis and Recognition **15**(1), 45–55 (2012)
31. Řehůřek, R.: Subspace tracking for latent semantic analysis. In: Proceedings of the 33rd European Conference on Information Retrieval Research, pp. 289–300 (2011)
32. Rusiñol, M., Karatzas, D., Bagdanov, A.D., Llados, J.: Multipage document retrieval by textual and visual representations. In: International Conference on Pattern Recognition, pp. 521–524 (2012)
33. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988). DOI 10.1016/0306-4573(88)90021-0
34. Sarkar, P.: Image classification: classifying distributions of visual features. In: Proceedings of the International Conference on Pattern Recognition (2006)
35. Schmidtler, M., Amtrup, J.: Automatic document separation: A combination of probabilistic classification and finite-state sequence modeling. In: Natural Language Processing and Text Mining, pp. 123–144 (2006)
36. Sebsatiani, F.: Machine learning in automated text categorization. Journal ACM Computing Surveys **34**(1), 1–47 (2002)
37. Shin, C., Doermann, D., Rosenfeld, A.: Classification of document pages using structure-based features. International Journal on Document Analysis and Recognition **3**(4), 232–247 (2001)
38. Sidiropoulos, P., Vrochidis, S., Kompatsiaris, I.: Content-based binary image retrieval using the adaptive hierarchical density histogram. Pattern Recognition **44**(4), 739–750 (2011)
39. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49 (1999)
40. Yang, Y., Pederson, J.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412–420 (1997)
41. Young, S., Russell, N., Thornton, J.: Token passing: A simple conceptual model for connected speech recognition systems. Tech. Rep. CUED/F-INFENG/TR38, Cambridge University (1998)