Multipage Document Retrieval by Textual and Visual Representations

Marçal Rusiñol[1], Dimosthenis Karatzas[1], Andrew D. Bagdanov[2] and Josep Lladós[1]

[1]*Computer Vision Center, Dept. Ciències de la Computació*
Universitat Autònoma de Barcelona, Bellaterra, Spain
{marcal,dimos,josep}@cvc.uab.cat
[2]*Media Integration and Communication Center*
Università degli Studi di Firenze, Florence, Italy
bagdanov@dsi.unifi.it

Abstract

In this paper we present a multipage administrative document image retrieval system based on textual and visual representations of document pages. Individual pages are represented by textual or visual information using a bag-of-words framework. Different fusion strategies are evaluated which allow the system to perform multipage document retrieval on the basis of a single page retrieval system. Results are reported on a large dataset of document images sampled from a banking workflow.

1 Introduction

Every day public organizations, social security services and large companies handle large volumes of various administrative documents such as identity cards, forms, mails, etc. The Document Image Analysis and Recognition (DIAR) community has look into solutions for automating the screening process and for extracting relevant information from each document with minimal human intervention. One of the key issues in digital mailroom applications is document image retrieval. Although many works dealing with single-page document representations can be found (e.g. [3, 1]), the literature dealing with variable-length multipage documents, representing a most realistic scenario, is scarce.

In [4], Frasconi et al. proposed a hidden Markov model-based system aimed at categorizing documents by looking at sequences of pages. The contextual information provided by the analysis of page sequences help to significatnlty improve single page classification accuracy. Since the main aim of their work was to categorize the textual contents of multipage documents, they propose the use of a bag-of-words representation of the documents. On the other hand, Gordo and Perronnin proposed in [5] a bag-of-pages approach that treats multipage documents as unordered sets of pages. After a learning stage, each page is assigned to a prototype in a page vocabulary, leading to a histogram representation of multipage documents. In this case, the authors were more interested in the visual appearance of the different pages more than the contents themselves. The pages are encoded using visual descriptor based on multiscale run-length histograms.

In this paper we focus on a retrieval scenario rather than a classification problem. Given a dataset of multipage documents, the user feeds the system with a multipage document query and expects to retrieve similar documents, where "similarity" can be defined over different modalities. We start by proposing an individual page retrieval framework that encodes pages by either visual or textual information. We then test several page-fusion strategies aiming to perform document (multipage) retrieval. Then we examine the possibility to combine the retrieval results obtained at the document level through the two different representation modalities (visual and textual). Results are reported on a large dataset of multipage document images sampled from a real banking workflow.

The rest of this paper is organized as follows. We detail in Section 2 both the visual and textual descriptions

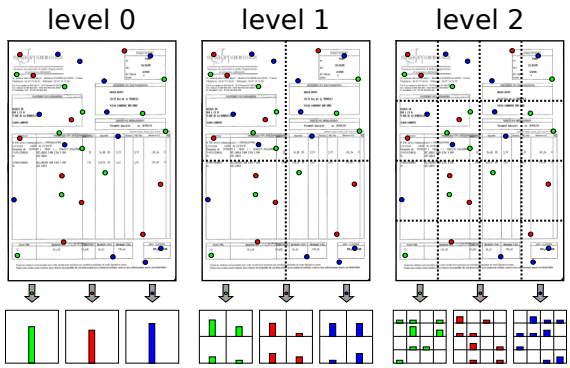

Figure 1. An illustrative example of Spatial Pyramid Matching (SPM).

of page images. In Section 3 the multipage document retrieval framework is presented and in Section 4 we present the experimental setup and results. Finally, we give some concluding remarks in Section 5.

2 Individual Page Description and Retrieval

We propose two different modalities to describe document page images. A visual description captures the overall appearance of the page, which is usually preferred when documents from the same class share the same overall structure although their contents might differ. The textual descriptor describes pages in terms of their contents and is more suitable when documents from the same class share the same topic though they might look very different. Both modalities are represented using a bag-of-(textual or visual)-words framework.

2.1 Visual Description

As visual description of document pages we use the Spatial Pyramid Matching (SPM) method proposed by Lazebnik et al. in [6]. First we densely extract SIFT [7] features from the document images. In our experimental setup, SIFT descriptors are computed over 40×40 pixel patches in a regular grid of 20 pixels. Features having an overall weak gradient magnitude are discarded. We then construct a visual codebook through the application of the k-means clustering to quantize the SIFT features into visual words. The vocabulary sizes in our experiments range from $K = 64$ to $K = 512$. Finally, in order to add some coarse spatial information to the final visual descriptor, the document image is divided at different levels of resolution. For each level we count how many features quantized in the same type fall in each spatial bin. The final visual descriptor is formed by concatenating the weighted histograms of all visual words at all resolutions. In our configuration we use three resolution levels, the first one encodes the whole image, the second and third ones divide the image in 4 and 16 equally sized bins respectively. We can see an illustrative example of our SPM configuration in Fig. 1.

2.2 Textual Description

In order to use textual information as another cue, we use Latent Semantic Analysis (LSA) on the textual content of documents. The administrative document images are first OCRed with the commercial OCR system from ABBYY. Given the ASCII representations of each document image, a text preprocessing step is applied. We first stem the words with the Porter stemming algorithm implemented in the Snowball [8] system. Then, a stopword filtering step is applied. We then represent each document image as a bag-of-words vector. Each bag-of-words vector is weighted by applying the tf-idf model. Finally we use LSA, presented by Deerwester et al. [2], in order to reduce the dimensionality of the feature vector, to be more tolerant to possible OCR errors and to add semantic coherence to the descriptor. The LSA model assumes that there exists some underlying semantic structure in the descriptor space. This semantic structure is defined by modeling each document as a mixture of topics which can be estimated in an unsupervised way by a truncated singular value decomposition. Each bag-or-words feature vector is projected to the topic space in order to obtain the final textual descriptor. In our experimental evaluation we tested several numbers of topics values ranging from $T = 64$ to $T = 512$.

3 Multipage Document Retrieval

Whether we use the visual or textual page description, we obtain a histogram encoding of pages. These histograms are $L2$-normalized and retrieval of similar pages given a query is computed using the cosine distance between the query descriptor and the pages stored in the dataset. In order to adapt these page representations and retrieval scheme to work with multipage documents, we have tested three different fusion strategies.

- **Early fusion** encodes the multipage documents in a *single* histogram descriptor. Since both the visual and textual descriptors are based on a bag-of-words framework, the multipage document descriptor counts the occurrences of a given (visual)

word among all the pages of the document. Note that we still use the SPM configuration for the visual modality and the LSA technique for the textual one.

- **Late fusion** strategies perform individual page retrievals for each page of the query documents and then combine the resulting lists. We have tested two different late fusion approaches.
  - **CombMAX** combines the resulting lists in terms of the scores of each page. We associate each multipage document in the dataset with the maximum score that any of its pages received.
  - **Borda count** combines the resulting lists just in terms of the page ranking independently of their scores. The topmost page on each ranked list gets n votes, where n is the dataset size. Each subsequent rank gets one vote less than the previous. The final ranked list is obtained by adding all the votes per document and sorting again.

We have also tested combining the two different modalities (visual and textual) at the document (multipage) level. However, here, as it is difficult to normalize the textual and visual descriptors, neither early fusion nor combMAX can be used. In order to merge both information cues into a single resulting ranked list we perform multipage retrieval for both modalities independently at the document (multipage) level, and as a final step we combine the two resulting lists into a single one using the Borda count method.

4 Dataset and Experimental Results

Our dataset consists of more than 7200 multipage document images sampled from a real banking workflow, resulting in a total of nearly 70,000 pages. We can see an example of those documents in Fig. 3. The dataset is organized into 13 semantic classes such as tax forms, land registry deeds, balance sheets, etc. We use a leave-one-out cross validation in all retrieval experiments so that each mutlipage document is used once as a query against the rest of the dataset. The results are reported in terms of the mean average precision (mAP) and precision and recall plots.

We can see in Table 1 the mAP values for both the visual and textual descriptors as a function of the fusion method and several values of the visual vocabulary size K and the number of topics T used in LSA. In general, the late fusion method based on scores performs better

Table 1. mAP for visual and textual descriptions and all fusion strategies.

Modality	K / T	Multipage combination		
		Early fusion	CombMAX	Borda
Visual	64	43.13	47.2	42.66
	128	43.45	47.73	42.94
	256	43.02	47.08	42.68
	512	42.92	46.16	42.22
Textual	64	66.97	73.57	74.24
	128	65.13	74	72.89
	256	63.31	73.63	70.84
	512	61.17	72.52	67.69

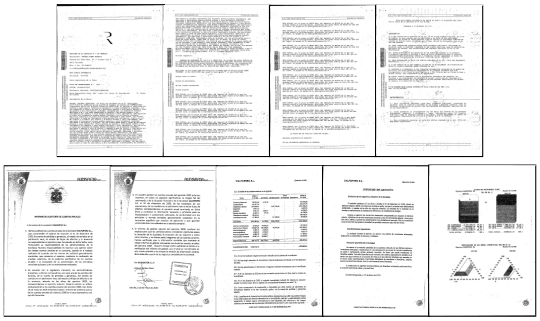

Figure 3. Multipage documents.

than the early fusion strategy. Although there is a significant improvement in using combMAX against Borda count in the visual scenario, this difference is not appreciable when using textual descriptors where the best performance is achieved with Borda count.

Regarding descriptor size, we can observe that usually the best performance is achieved when using rather small vocabularies and few topics.

Finally, it is worth noting the significant improvement in performance when describing documents by textual features rather than with visual ones. This is mainly due to the fact that in our banking dataset the document classes are more semantic than physical. That is, many documents from the same class share the same topic even though they are completely dissimilar in terms of layout. This is also evident when looking at the precision and recall plots in Fig. 2.

The fact that textual features perform better than visual ones in our scenario also affects the multimodal experiments combing the visual and textual modalities. We can see in Table 2 the results when combining textual and visual information with the best K and T configuration for each combination strategy. Although the

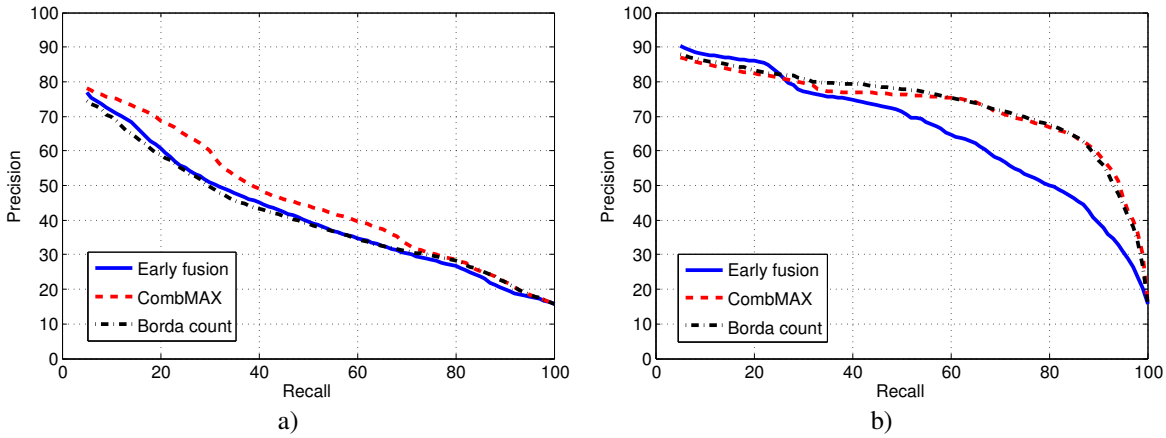

Figure 2. Precision and Recall plots. a) Visual description and b) Textual description.

Table 2. mAP when combining textual and visual results with Borda count.

Combination	K	T	mAP
Early fusion	128	64	56.03
CombMAX	128	128	52.6
Borda	128	64	58.38

obtained results outperform the use of only visual information they do not reach the performance of the textual modality alone. This is again a clear indicator that the visual description of those documents introduces noise in the whole system.

5 Conclusions

In this paper we have presented a multipage document image retrieval system powered by textual and visual flavors of the bag-of-words framework. We have studied the effect of three different fusion strategies for performing multipage document retrieval based on single-page similarities. Results were reported by using a large dataset of multipage document images from a real banking workflow.

In our specific scenario, textual representation of documents achieved the best results when the retrieval of multipage documents is performed using late fusion. However, this behavior might be biased by the nature of the documents in our use case scenario. Since the document classes are defined at a semantic level, it is not really fair to perform a retrieval in terms of visual similarity of document images. In addition, we observed that the excessive increase of the vocabulary length did not seem to enhance the results.

Acknowledgements

This work has been supported by the Spanish projects RYC-2009-05031, TIN2011-24631, TIN2009-14633-C03-03, Consolider Ingenio 2010: MIPRCV (CSD200700018) and the grant 2009-SGR-1434 of the Generalitat de Catalunya.

References

[1] N. Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *Int. J. Doc. Anal. Recog.*, 10(1):1–16, 2006.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.*, 41(6):391–407, 1990.

[3] D. Doermann. The indexing and retrieval of document images: A survey. *Comput. Vis. Image Und.*, 70(3):287–298, 1998.

[4] P. Frasconi, G. Soda, and A. Vullo. Hidden Markov models for text categorization in multi-page documents. *J. Intell. Inf. Syst.*, 18(2–3):195–217, 2002.

[5] A. Gordo and F. Perronnin. A bag-of-pages approach to unordered multi-page document classification. In *Proc. of the Int. Conf. Pattern Recog.*, pages 1920–1923, 2010.

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of the Conf. Comput. Vis. Pattern Recog.*, pages 2169–2178, 2006.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[8] M. Porter. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/introduction.html, 2001.