An Interactive Appearance-based Document Retrieval System for Historical Newspapers

Hongxing Gao[1], Marçal Rusiñol[1], Dimosthenis Karatzas[1], Apostolos Antonacopoulos[2], Josep Lladós[1]
[1]*Computer Vision Center, Dept. Ciències de la Computació,Edifici O, Universitat Autònoma de Barcelona,*
08193 Bellaterra (Barcelona), Spain.
[2] *Pattern Recognition and Image Analysis (PRImA) Research Lab*
,School of Computing, Science and Engineering, University of Salford, United Kingdom.
{*hongxing, marcal, dimos, josep*}*@cvc.uab.cat, A.Antonacopoulos@primaresearch.org*

Keywords: Document Image Retrieval, Relevance Feedback, Historical Newspapers

Abstract: In this paper we present a retrieval-based application aimed at assisting a user to semi-automatically segment an incoming flow of historical newspaper images by automatically detecting a particular type of pages based on their appearance. A visual descriptor is used to assess page similarity while a relevance feedback process allow refining the results iteratively. The application is tested on a large dataset of digitised historic newspapers.

1 INTRODUCTION

Since the dawn of the digital era, a lot of effort has been devoted to the mass digitization of cultural heritage assets such as books or newspapers. The main motivation that librarians had behind these digitization efforts was to find solutions to preserve and efficiently store these materials. However, it soon became clear that investing just in efficient ways of scanning in order to obtain digital images (the scan-to-archive paradigm) was not enough. The scanning processes covered the preservation and the storage aspects, but the raw image formats were of limited value without ways to provide access to the digital collection contents.

For example, the British Library[1] has digitised so far 5.5 millions of newspaper pages, corresponding to more than 200 different newspaper titles covering a period from the early 1700s until today. Each of these pages have been automatically processed by an OCR engine which creates electronic text in a searchable format. In order to provide a more accurate search, a manual segmentation of each page into article types and some metadata extraction such as date, region of publication, publication title, etc. has been done for each page. Of course this is a tedious and expensive step and many efforts, like the competition presented in (Antonacopoulos et al., 2011), have been devoted to automate this process.

Newspapers are digitised and OCRed in bulks. In order to organize the image database by date, newspaper, location, etc. a user has to manually segment the flow of digitised images into newspaper issues and label them with multiple metadata. Currently, such metadata are manually added, while the automatic extraction of relevant information is an open research issue. In this paper we present an application aimed at assisting the user to semi-autmatically segment the incoming flow of scanned images into newspaper issues by automatically detecting the Front-pages. Moreover, we propose to achieve this taking advantage solely of the visual similarity between front-pages of different issues, avoiding the expensive and error-prone steps of analysis and recognition. We show that this is possible at a large scale.

This paper offers a comparative study of retrieval methodologies in the context of the real-life problem of newspaper page flow classification. Individual pages are globally described by an image descriptor which is fast to calculate. To choose such a simple descriptor is a must in large-scale scenarios. Since the chosen descriptor performs well at retrieving pages by visual similarity, we expect that it is a feasible solution to aim the task of discriminating front pages from non-front pages. We also compare retrieval performance based on different distance metrics to assess visual similarity. Subsequently, we evaluate the usefulness of a relevance feedback step and compare representatives of the main relevance feedback methodologies. We evaluate the different retrieval systems

[1] http://www.britishnewspaperarchive.co.uk/

over a large real-life dataset of historical newspaper images created through the IMPACT EU project, comprising 23004 images of 8 different newspaper titles. We show that it is possible to achieve near perfect retrieval results and close to real-time performance.

The rest of this paper is organized as follows. We detail in Section 2 the visual descriptor used to represent page images. In Section 3 the page retrieval framework is presented. In Section 4 we present the experimental setup and results. Finally, we give some concluding remarks in Section 5.

2 DOCUMENT IMAGE DESCRIPTION

Describing an image with a pyramidal decomposition, proposed by Pierre Hroux (Héroux et al., 1998), provides an effective global way integrated with local information. It expresses the density of newspaper image at different level of scales. The algorithm is performed by a recursive operation of cutting the newspaper into four rectangular regions, the density values of which are used as feature vector. In practice, the number of iterative cuts represents the detail capture ability of the feature vector and defines the length of feature vector. As demonstrated in Figure 1, the first level (D_{01} in feature vector) corresponds to the density over the whole image, the second level gives the density of 4 rectangular cuts: D_{11}, D_{12}, D_{13} and D_{14} . Consequently, 5 level cut overall returns a feature vector with 341 (1+4+16+64+256) elements, and sixth level returns 1365 elements.

The pyramidal decomposition descriptor is scale invariant because the resolution or scale change does not lead to the density alteration. Besides, the feature vector extracted by pyramidal decomposition are relatively tolerant to translation and rotation. The extent of such tolerance depends on the feature vector level: the higher level cut, the finer feature vector extracted and hence the more sensitive to the translation and rotation. In general, scanned document flows exhibit limited skew, hence the pyramidal decomposition descriptor is an adequate choice.

3 RETRIEVAL BY SIMILARITY

In this paper, four commonly used distance metrics are checked to calculate the dissimilarity between the query feature vector and each newspaper image feature vector stored in the dataset. We define the feature vector of query newspaper as $\mathbf{Q} = (q_1, q_2, \cdots, q_n)$ and the newspaper in the dataset as $\mathbf{I} = (i_1, i_2, \cdots, i_n)$. Here, q_j and i_j represents the jth elements of feature vector of the query and dataset image respectively. The different distance metrics between $\mathbf{Q}$ and $\mathbf{I}$ are defined as follows.

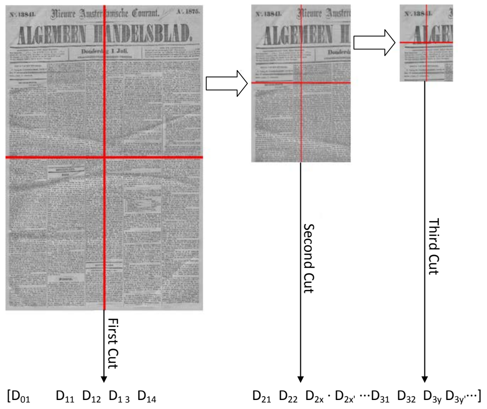
First Cut
Second Cut
Third Cut
[D01 D11 D12 D1 3 D14 D21 D22 D2x · D2x' ···D31 D32 D3y D3y'···]

Figure 1: Demonstration of Pyramidal Decomposition. The extracted feature vector is showed at the bottom.

- **Euclidian distance**

$$d(\mathbf{Q}, \mathbf{I}) = \sqrt{\sum_{j=1}^{N} (q_j - i_j)^2} \tag{1}$$

- **Chi-square distance**, supposing $C = (c_1, c_2, \cdots, c_j, \cdots, c_n)$ as the average feature vector of the images in the dataset

$$d(\mathbf{Q}, \mathbf{I}) = \sqrt{\sum_{j=1}^{N} (q_j - i_j)^2 / c_j} \tag{2}$$

- **Cosine distance**

$$d(\mathbf{Q}, \mathbf{I}) = 1 - \frac{\sum_{j=1}^{N} q_j * i_j}{\sum_{j=1}^{N} q_j^2 * \sum_{j=1}^{N} i_j^2} \tag{3}$$

- **Histogram intersection distance**

$$d(\mathbf{Q}, \mathbf{I}) = 1 - \frac{\sum_{j=1}^{N} min(q_j, i_j)}{min(\sum_{j=1}^{N} q_j, \sum_{j=1}^{N} i_j)} \tag{4}$$

As distance metric computation is an iteratively progress for each image of dataset in on-line way, its computation complexity should be especially considered for image retrieval in large dataset. Among the four introduced distance metrics, Cosine distance metric possesses the lowest computation complexity as it boils down to a simple multiplication of corresponding elements and a summation if the data has been normalized previously while other distance metrics are much more complicated. Besides, for pyramidal decomposition feature vectors of the newspaper images, because the scale of feature vector is 'ignored' in Cosine distance metrics, Cosine distance metrics also hold the invariant to illumination change,which commonly happens for scanned images. So in Section 4.2, Cosine distance showed its advantage on both efficiency (fast to calculate) and effectiveness (ability to assess visual similarity correctly given the feature) over other distance metrics.

4 EXPERIMENTAL RESULTS

4.1 Dataset and Evaluation Measures

The experimental setup is implemented in two steps: we first extracted the feature vectors of all images in the database in an off-line fashion. Subsequently, we evaluated the effect of four alternative distance metrics discussed in section 3 on a subset of 500 newspaper images comprising 2 classes (front-page and non-front-page from a single newspaper title). This experiment aimed to test the effectiveness of the feature vector extraction process and establish the best distance metric for retrieval. We execute the two previously stated relevance feedback strategies using the pyramidal decomposition feature vectors and the Cosine distance metric over the whole newspaper dataset containing 23004 images comprising 16 classes (front-page or non-front-page for 8 different newspaper titles).

In each of the experiments, we use the Mean Average Precision (MAP) to evaluate the performance of the system.

4.2 Evaluation of different distance metrics

In order to evaluate the performance of the different distance metrics to retrieval, we performed an experiment using a subset of the newspaper dataset consisting of 500 images of a single title which were previously classified in two classes (108 front-pages and 392 non-front-pages).

In addition, to study the effect of adding more detail in the pyramidal decomposition feature, we repeated the experiment using different levels of decomposition.

We compute the precision-recall curve and MAP for each image in our small dataset in a leave-one-out fashion. Each image is taken as query and we perform the retrieval versus the remaining 499 images. Consequently we obtain 500 precision-recall curves and 500 MAP values for each distance metric.

Table 1: Average MAP of all the queries.

		Euc.	χ^2	Cos.	HI
All images	L4	0.8247	0.8266	**0.9438**	0.7337
	L5	0.8364	0.8390	**0.9342**	0.7318
	L6	0.8615	0.8652	**0.9326**	0.7384
Only front-page	L4	0.5820	0.5843	**0.8939**	0.4279
	L5	0.6555	0.6585	**0.9255**	0.4580
	L6	0.7469	0.7508	**0.9461**	0.4977
Only non-front-page	L4	0.8916	0.8933	**0.9575**	0.8180
	L5	0.8863	0.8887	**0.9438**	0.8072
	L6	0.8931	0.8967	**0.9289**	0.8048

We also perform the experiment on the front-page subset (only the front-pages are used as queries) and non-front-page subset (only non-front-pages are used as queries) separately in order to evaluate the performance on certain classes of images (see Table1).

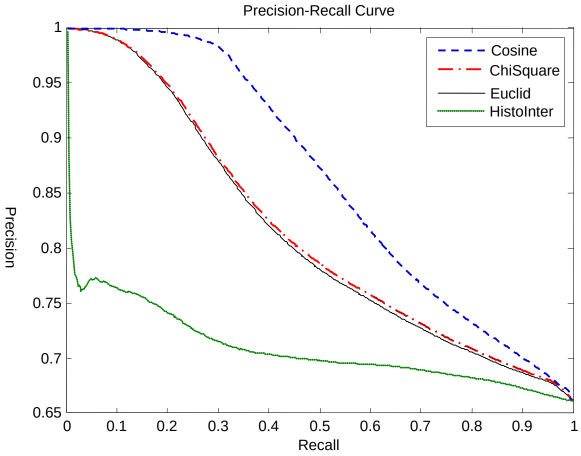

Figure 2: Precision-Recall curve of 500 queries (both front-Pages and non-front-Pages, L5 feature vector).

Despite the different level of feature vector extraction and different type of dataset, the Cosine distance always performs much better than any other distance metric evaluated for the particular feature vector used; Euclidian distance and Chi-Square distance yield similar performance with the latter being marginally better, while Histogram Intersection distance invariably yields the poorest performance during the whole experiment. Consequently, we arrive to the conclusion that the Cosine Distance is much more suitable for the particular application of front-page newspaper image retrieval.

For front-page only images retrieval, increasing the level of pyramid decomposition feature has a clear effect: it produces better results no matter what the distance metrics are used. However, for non-front-pages only retrieval, it does not yield any better performance for higher level feature. This is to be expected because all the front-pages contain a pretty fixed part (the title part) which is better taken advantage of at higher detail levels. The non-front-pages on the other hand are similar only at the level of the general structure (e.g. all will have the same number of columns) which is already captured at low pyramid levels. Consequently, adding more detail by extracting higher level feature does not yield any better performance.

We also calculated the computational cost of extracting the pyramidal decomposition feature over the 500 query images ('Descriptor' in Table 2) as well as the distance calculation between the query image and all dataset images ('Distance' in Table 2). The time shown in Table 2 is the average time consumption over 500 queries. Since feature vector extraction consumes much more time than distance calculation, the computational complexity will rise up sharply for higher feature vector extraction levels as the dimension of feature vector space is increased.

Table 2: Time Consumption of Feature Extraction and Distance Calculation.

Feature Level	Descriptor (s.)	Distance (s.)
L4	0.0277	3.3280e-05
L5	0.0445	9.2561e-05
L6	0.0979	3.5360e-04

4.3 Results on Relevance Feedback

Although the retrieval performance is good on small datasets as demonstrated above, retrieval performance is decreasing when generalizing the previous retrieval method to the whole newspaper containing 23004 images comprising of 16 classes. Consequently, we perform relevance feedback Strategies on the whole newspaper dataset to improve the user's retrieval experience by using the Rocchio (Rocchio, 1971) and Relevance Score (Giacinto, 2007) methods. Within the 23704 images, 2000 images are randomly chosen as queries. For each query, 100 relevant images are expected to be retrieved and the relevance feedback step is executed no more than 9 times over top-20 retrieved result. Figure 3 demonstrates how relevance feedback improves the retrieval result.

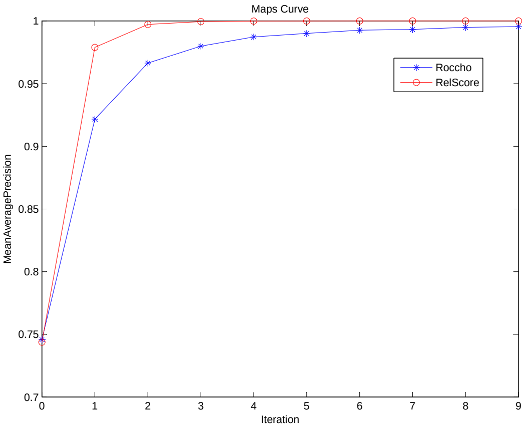

Figure 3: MAP improvement by Relevance Feedback.

5 CONCLUSIONS

In this paper we have presented a retrieval-based application aimed at assisting a user to semi-automatically segment an incoming flow of historical newspaper images by automatically detecting a particular type of pages based on their appearance. An off-the-shelf visual descriptor has been used to assess page similarity while a relevance feedback process allowed refining the results iteratively. The presented application has been tested on a large dataset of digitised historic newspapers.

As future research lines, we are interested in the application of *active learning* and *crowdsourcing* strategies to this particular problem in order that the system automatically selects which are the most relevant images to present to the user to get feedback.

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 China Scholarship Council grant (No.2011674029).

REFERENCES

Antonacopoulos, A., Clausner, C., Papadopoulos, C., and Pletschacher, S. (2011). Historical document layout analysis competition. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1516–1520.

Giacinto, G. (2007). A nearest-neighbor approach to relevance feedback in content based image retrieval. In *Proceedings of the 6th ACM international conference on image ans video retrieval*, pages 456–463.

Héroux, P., Diana, S., Ribert, A., and Trupin, E. (1998). Classification method study for automatic form class identification. *Proceeding of the International Conference on Pattern Recognition*, 1:926–928.

Rocchio, J. (1971). Relevance feedback in information retrieval. In *SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323.