Layout-Based Document-Retrieval System by Radon Transform Using Dynamic Time Warping

Giuseppe Pirlo[1,*], Michela Chimienti[2], Michele Dassisti[3], Donato Impedovo[4], and Angelo Galiano[4]

[1] Dipartimento di Informatica, Università degli Studi di Bari "A. Moro", via Orabona 4, 70125-Bari, Italy
[2] Laboratorio Kad3, C.da Baione, 70043 Monopoli (BA), Italy
[3] Dip. Meccanica, Management e Matematica, Politecnico di Bari, viale Japigia 182, 70126 - Bari, Italy
[4] Dyrecta Lab, Via V. Simplicio 45, 70014 Conversano (BA), Italy
giuseppe.pirlo@uniba.it

Abstract. In the context of sustainability of document management technologies, this paper presents a new system for layout-based document retrieval specifically designed for commercial form retrieval. The system first uses a technique based on mathematical morphology to extract grid-based structural components from the document image. Successively, Radon Transform is used for document layout description. A document matching technique based on dynamic time warping is finally adopted. The experimental results carried out on real and simulated data set, demonstrate the effectiveness of the approach with respect to different classes of commercial forms.

Keywords: Document management, Document Image Retrieval, Sustainability, Mathematic Morphology, Radon Transform, Dynamic Time Warping.

1 Introduction

Information Retrieval (IR) is a critical task of document management systems as the number of documents available in databases and digital libraries exponentially grows. Quite often useless reprinting becomes a necessary activity in case of document loss or unavailability. This is also due to standard systems for document retrieval that use text data. They require a document to be present in text form and the querying method is based on a specific textual content in the document. Several advanced techniques have been proposed, based on set-theoretic, algebraic and probabilistic models [1, 2, 3]. Whatever the model used, one of the main drawback of text-based document retrieval systems is that they require a document in text form, since the search for similar documents is based on comparing the textual contents. As a consequence, a preliminary stage of image to text conversion by an Optical Character Recognizer (OCR) is required when a document is in image form. OCR is a time-consuming

* Corresponding author.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 61–70, 2013.

error-prone process, specifically in the case of multi-lingual/multi-font documents and poor-quality document images [4, 5, 6]. The interested reader can refer also to two comprehensive surveys on this topic [7, 8].

In many cases, document search does not depend on the textual content while it is useful to search a document on the basis of its structure. In such cases, methods adopted for document retrieval is a feature vector, in which each feature is extracted from a specific region of the document image. For instance, Tzacheva et al. [9] used a static zoning strategy for document image decomposition to extact a fixed-size feature vector from the document image. In this approach, a regular grid is superimposed to the document image in order to extract regional characteristics. Duygulu et al. [10] proposed a hierarchical zoning strategy to overcome the problem of optimal grid selection, in order to face with the treatment of set of documents of different characteristics. Huang et al. [11] presented a system that extracts text lines and describes the layout by means of relationships between pairs of these lines. Erol et al. [12] used Brick Wall Coding Features (BWC) features, that are local features which represents bounding boxes of the words. Although the features are scale invariant and robust to slight perspective distortion, the accuracy of their system is very low. In addition the method does not work correctly when documents are written in languages such as Japanese and Chinese, in which words are not separated. The system of Liu and Liao [13] combines several approaches to identify a document, as for instance barcode, micro optical patterns, encoding hidden information, paper fingerprint, character recognition, local features etc. . Unfortunately, the retrieval process is time consuming and requires special equipment.

In this paper we deal with commercial forms, such as invoices, waybills, receipts, etc., where layout is strongly characterized by a grid-structure. In this particular cases, traditional document-image approaches are not effective since they are not able to describe documents on the basis of the grid-based structure. A layout-based Document Retrieval (LDR) system is proposed in this paper to handle automatically the documents. In a first step it uses a technique based on mathematical morphology for extracting the grid-based structure in the document layout, removing textual components. Subsequently Radon Transform is used to obtain the feature vector characterizing the specific grid structure of the document. A Dynamic Time Warping (DTW) - based technique finally performs document matching.

The paper is organized as follow. Section 2 presents overall structure of the system. In Section 3 the preprocessing phase is described, based on mathematical morphology operators. Feature extraction is presented in Section 4. Section 5 discusses the matching process based on DTW. The decision combination process is reported in Section 6. Section 7 presents the experimental results while Section 8 presents the conclusion of this work and highlights some directions for further research.

2 System Architecture

Figure 1 shows the structure of the LDR system presented in this paper. After document image acquisition, the document is preprocessed and transformed by Radon

Transform. The features extracted are then stored in the reference database in the enrollment stage. In the running stage, an unknown document is first scanned and preprocessed, successively the features are extracted compared to the those stored into the database. The matching module performs matching by Dynamic Time Warping (DTW) and outputs the ranked list of similar documents.

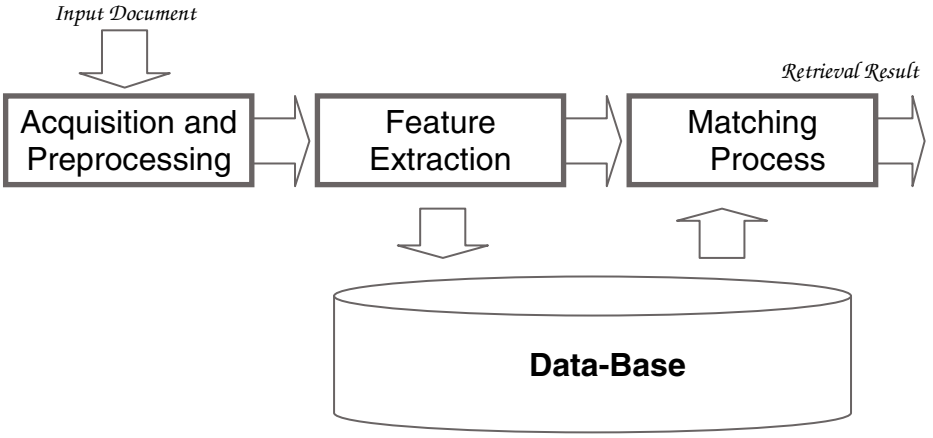


Fig. 1. The LDR system

Each processing step is performed by a well-defined software module. The following Sections will describe each module in detail.

3 Data Acquisition and Preprocessing

The data acquisition and pre-processing module is a key part of the system. It controls the acquisition of the input document as a standard 256 gray-level – 100dpi PDF file. Figures 2 shows an input document concerning a real invoice.

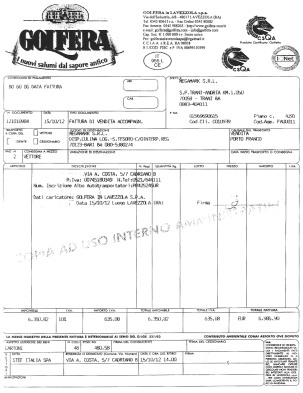
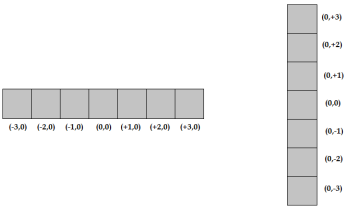


Fig. 2. Input document image I=I(x,y)

Successively, after noise removal, document is resampled to 100 dpi and grid-based structure is extracted by mathematical morphology [14, 15]. More precisely, let I=I(x,y) be the document image ($1 \le x \le X$, $1 \le y \le Y$) and let be

- B_{hor} the horizontal structure element defined as (see Figure 3a):
  $B=\{(-s,0), \ldots, (-1,0), (0,0), (1,0), \ldots , (s,0)\}$;
- B_{ver} the horizontal structure element defined as (see Figure 3b):
  $B=\{(0,-s), \ldots, (0,-1), (0,0), (0,1), \ldots , (0,s)\}$;

being s a small positive integer which determine the size of the structure element.



(a) (b)

Fig. 3. Structure elements (s=3): (a) B_{hor} , (b) B_{ver}

In the preprocessing phase from the image I(x,y) two filtered images $I_{hor}=I_{hor}(x,y)$ and $I_{ver}=I_{ver}(x,y)$, which contains respectively horizontal and vertical segments, are obtained by a closure operator as follows:

$$I_{hor}= I \bullet B_{hor} = (I \oplus B_{hor}) \Theta B_{hor} \quad (1a)$$

$$I_{ver}= I \bullet B_{ver} = (I \oplus B_{ver}) \Theta B_{ver} \quad (1b)$$

being "$\bullet$" the closure operator, while "$\oplus$" and "Θ" indicate respectively Minkowski sum and difference.

(a) (b)

Fig. 4. Example of filtered images: (a) I_{hor} , (b) I_{ver}

Finally, $I_{hor}(x,y)$ and $I_{ver}(x,y)$ are combined to reconstruct the preprocessed image I^* according to XOR operator:

$$I^* = I_{hor} \ \text{XOR} \ I_{ver} \tag{2}$$

Figure 5 shows an example of document image after preprocessing.

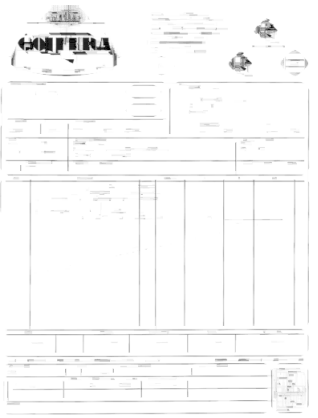


Fig. 5. The preprocessed image I^*

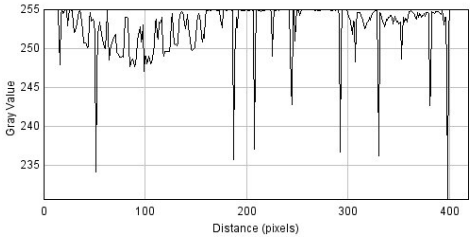
4 Feature Extraction

The feature extraction module, that has been specifically designed for grid-based layout document images, used the Radon Transform then has been extensively used in image analysis and has a number of important applications, like those related to MRI and computed tomography [16, 17]. The complete description of the Radon Transform is beyond the scope of this paper (see further details in [18, 19]). For the aim of this paper we only remind that the Radon Transform computes projection sum of the image intensity along a oriented at line ($\rho - x\cos\vartheta - y\sin\vartheta = 0$), for each ϑ and ρ. More precisely the Radon Transform of a function $I^*(x,y)$ in an Euclidean space is defined by [20]:

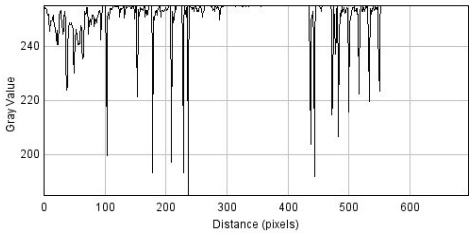
$$S_{\vartheta,\rho} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I^*(x,y)\cdot\delta(\rho - x\cos\vartheta - y\sin\vartheta)dxdy \tag{3}$$

where the d (r) is Dirac function, which is infinite for argument zero and zero for all other arguments (it integrates to one).

Therefore, reckoning of the Radon Transform of a two dimensional image intensity function $I^*(x,y)$ results in its projections across the image at arbitrary orientations ϑ and offsets ρ. Figure 6 presents the results of the Radon Transform applied to the preprocessed image I^* for the parameter values related to horizontal ($\vartheta=0$, $\rho=0$) and vertical ($\vartheta=\pi/2$, $\rho=0$) projections.



(a) Horizontal Projection



(b) Vertical Projection

Fig. 6. Feature extraction by Radon Transform

5 Matching

The matching procedure is performed applying the Dinamic Time Warping (DTW) on the feature vectors extracted by the radon transform. In particular, let be F^r, S^t the feature vectors of M elements extracted from the document images I^{*r} and I^{*t}, a warping function between S^r and S^t is any sequence of couples of indexes identifying points of S^r and S^t to be joined [21, 22]:

$$W(S^r,S^t)=c_1,c_2,\ldots,c_K, \tag{4}$$

where $c_k=(i_k,j_k)$ $(k,i_k,j_k$ integers, $1\le k\le K$, $1\le i_k\le M$, $1\le j_k\le M)$. Now, if we consider a distance measure $d(c_k)=d(z^r(i_k), z^t(i_k))$ between elements of S^r and S^t, we can associate to $W(S^r,S^t)$ the dissimilarity measure

$$D_{W(S^r,S^t)} = \sum_{k=1}^{K} d(c_k) \, . \tag{5}$$

The DTW detects the warping function $W^*(S^r,S^t)= c^*_1,c^*_2,\ldots,c^*_{K^*}$ which satisfies the condition of [21]:

- Monotonicity (i.e. $i_{k-1}\le i_k$, $j_{k-1}\le j_k$ for k=2,...K) (6a)
- Continuity (i.e. $i_k - i_{k-1} \le 1$, $j_k - j_{k-1} \le 1$ for k=2,...K) (6b)
- Boundary (i.e. $i_1 = 1$, $j_1 = 1$ and $i_K = M$, $j_K = M$) (6c)

and which provides the distance value between S^r and S^t defined as [21, 22]:

$$D_{W^*(S^r,S^t)} = \min_{W(S^r,S^t)} D_{W(S^r,S^t)} . \quad (7)$$

The value in eq. (7) represents the similarity between the document images I^{*r} and I^{*t}. Hence, according to this strategy, give a document image as input the matching module will outputs the ranked list of the k top similar document images retrieved from the database.

6 The Decision-Making Process

The matching procedure provides two distance-based ranked lists of documents, obtained respectively from horizontal and vertical projection. The decision making process combines the two ranked lists to let take the final decision. In this paper the Borda-count strategy was considered [23, 24]. According to this strategy, let be $D=\{D_1, D_2,\ldots, D_k,\ldots, D_K\}$ the set of K documents enrolled into the system for reference and D" the unknown input document. Furthermore, let be:

- $L^h : < D^h_1, D^h_2, \ldots, D^h_k, \ldots, D^h_K >$ the ranked list of documents obtained from the match of the horizontal projection ($D^h_k \in D$, for k=1,2,,,,,K and $D^h_{k1} \neq D^h_{k2}$ for $k_1 \neq k_2$);
- $L^v : < D^v_1, D^v_2, \ldots, D^v_k, \ldots, D^v_K >$ the ranked list of documents obtained from the match of the vertical projection ($D^v_k \in D$, for k=1,2,,,,,K and $D^v_{k1} \neq D^v_{k2}$ for $k_1 \neq k_2$);

The Borda-count approach assigns to each reference document D_k a confidence score $S(D_k)$ defined as [25]:

$$S(D_k) = S^h(D_k) + S^v(D_k) \quad (8)$$

being $S^h(D_k)=K-i$, if $D_{k=} D^h_i$; $S^v(D_k)=K-j$, if $D_{k=} D^v_j$.
Hence, the final list of ranked documents is

$$L^* : < D^*_1, D^*_2, \ldots, D^*_k, \ldots, D^*_K > \quad (9)$$

so that D^*_{k1} precedes D^*_{k2} in L^* if and only is $S(D_{k1}) \geq S(D_{k2})$, and – of course - $_1$ is the top candidate document [25].

7 Experimental Results

For the experimental test, a set of 33 documents have been considered. They are authentic commercial forms belonging to 16 different categories, as Table 1 reports.

Table 1. Dataset

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Number of Documents	9	4	3	3	2	2	1	1	1	1	1	1	1	1	1	1

Documents were scanned (100dpi , 256 gray-level) and preprocessed. Finally they were stored into a database along with the values of the Radon Transform concerning the horizontal ($S_{0,0}$) and vertical ($S_{0,\pi/2}$) projection. In the testing phase each document has been considered for verifying the effectiveness of the system. In order to estimate the quality of the ranked list provided by the system for a given query, the Average Normalized Rank (ANR) was adopted, defined as in [26]:

$$ANR = \frac{1}{N \cdot N_w} \cdot \sum_{i=1}^{N_w} \left(R_i - \frac{N_w + 1}{2} \right) \tag{10}$$

being

- N the number of documents in the set,
- N_w the number of relevant documents (for the given query) in the set,
- R_i is the rank of each relevant document in the set.

It is worst noting that ANR ranges in [0,1]:

- ANR=0 means that relevant documents are at the top of the ranked list (right position);
- ANR=1 means that relevant documents are at the bottom the ranked list (wrong position).

Figure 7 shows the experimental results. They demonstrate that the proposed approach is very robust with respect to different categories of documents. On average the value of ANR is equal to 0.08. Furthermore, 26 cases out of 33 the ANR is less than 10%, whereas only in one case it is greater than 0.5. Of course, further experiments are in progress and will be shown at the conference.

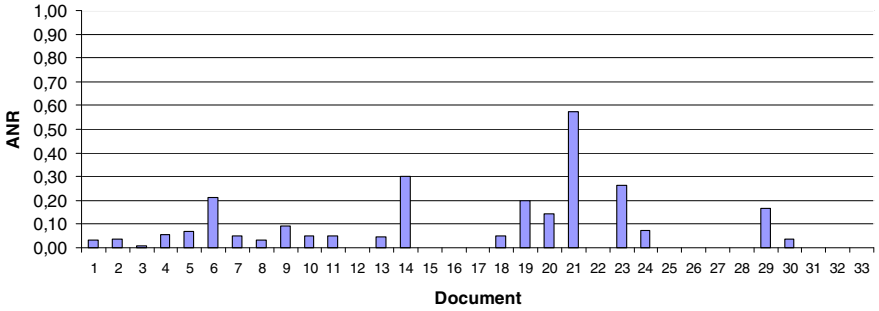


Fig. 7. Experimental Results: Average Normalized Rank (ANR) vs document type

8 Conclusion

A new system for layout-based document image retrieval was presented in this paper.

The system was specifically designed for retrieval of commercial forms as invoices, waybills and receipts, to optimize document management and sustainability.

It uses a morphologic filtering technique and the Radon Transform to obtain multiple document image descriptions. Document matching is then performed on each description by Dynamic Time Warping and a Borda-count decision combination strategy is finally used to combine multiple decisions.

The experimental results, carried out on a dataset of real commercial documents, demonstrate the effectiveness of the proposed solutions. Of course, further experiments are in progress to evaluate system robustness with respect to the size of the dataset as well as to document quality (i.e. faxed/photocopied documents) and image alterations in the acquisition process (i.e. document shift, rotation, etc.). The results will be shown at the conference.

References

1. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Cambridge Press (2009)
2. Doermann, D.: The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding 70(3), 287–298 (1998)
3. Ko, Y.: A study of term weighting schemes using class information for text classification. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1029–1030. ACM, NY (2012)
4. Marukawa, K., Hu, T., Fujisawa, H., Shima, Y.: Document retrieval tolerating character recognition errors - Evaluation and application. Pattern Recognition 30(8), 1361–1371 (1997)
5. Taghva, K., Borsack, J., Condit, A.: Evaluation of model-based retrieval effectiveness with OCR text. ACM TOIS 14(1), 64–93 (1996)
6. Lopresti, D.: Robust Retrieval of noisy text. In: Proceedings of the Third Forum on Research and Technology Advances in, pp.76–85 (1996)
7. Doermann, D.: The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding 70(3), 287–298 (1998)
8. Mitra, M., Chaudhuri, B.: Information retrieval from documents: A Survey. Information Retrieval 2(2/3), 141–163 (2000)
9. Tzacheva, A., El-Sonbaty, Y., El-Kwae, A.: Document Image Matching Using a Maximal Grid Approach. In: Proc. SPIE Document Recognition and Retrieval IX, pp. 121–128 (2002)
10. Duygulu, P., Atalay, V.: A Hierarchical Representation of Form Documents for Identification and Retrieval. International Journal on Document Analysis and Recognition 5(1), 17–27 (2002)
11. Huang, M., Dementhon, D., Doermann, D., Golebiowski, L.: Document ranking by layout relevance. In: Proc. Eighth International Conference on, vol. 1, pp. 362–366 (2005)
12. Erol, B., Antúnez, E., Hull, J.J.: Hotpaper: multimedia interaction with paper using mobile phones. In: Proceeding of the 16th ACM International Conference on Multimedia, pp. 399–408 (2008)
13. Liu, Q., Liao, C.: PaperUI. In: Iwamura, M., Shafait, F. (eds.) CBDAR 2011. LNCS, vol. 7139, pp. 83–100. Springer, Heidelberg (2012)
14. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press (1982)

15. Pirlo, G.: Removing Underlines from Handwritten Text: An experimental investigation. In: Downton, C., et al. (eds.) Handwriting Recognition, pp. 497–502. World Scientific Publishing Co. Pte. Ltd., Singapore (1997) (in Progress)
16. Cormack, A.M.: Computed tomography: Some history and recent developments. In: Proc. Symposia in Applied Mathematics, vol. 27, pp. 35–42 (1983)
17. Deans, S.R.: The Radon Transform and Some of Its Applications. Wiley, NY (1983)
18. Jafari-Khouzani, K., Soltanian-Zadeh, H.: Radon Transform orientation estimation for rotation invariant texture analysis. IEEE Trans. Pattern Anal. Mach. Intell. 27(6), 1004–1008 (2005)
19. Seo, S., et al.: A robust image fingerprinting system using the Radon transforms. Signal Process. Image Commun. 19(4), 325–339 (2004)
20. Hjouj, F., Kammler, D.W.: Identification of Reflected, Scaled, Translated, and Rotated Objects From Their Radon Projections. IEEE Trans. Image Processing 17(3), 301–310 (2008)
21. Salvador, S., Chan, P.: Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. In: Proc. KDD Workshop on Mining Temporal and Sequential Data, pp. 70–80 (2004)
22. Lemire, D.: Faster Retrieval with a Two-Pass Dynamic-Time-Warping Lower Bound. Pattern Recognition 42(9), 2169–2180 (2009)
23. Kittler, J., Hatef, M., Duin, R.P.W., Matias, J.: On combining classifiers. IEEE Trans. on Pattern Analysis Machine Intelligence 20(3), 226–239 (1998)
24. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Transaction on Systems, Man and Cybernetics 22(3), 418–435 (1992)
25. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. IEEE Trans. Pattern Anal. Mach. Intell. 16(1), 66–75 (1994)
26. Huang, M., Dementhon, D., Doermann, D., Golebiowski, L.: Document ranking by layout relevance. In: Proc. 8th ICDAR, pp. 362–366 (2005)