Document Image Retrieval by Layout Analysis

G. Pirlo[1] , M. Chimienti[2], M. Dassisti[3], D. Impedovo[4], A. 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*

(corresponding author: giuseppe.pirlo@uniba.it)

A new layout-based document image retrieval system is presented in this paper. The system is specifically designed for commercial form retrieval and uses mathematical morphology to extract structural components from the document image. Document layout description is performed by the Radon Transform whereas Dynamic Time Warping is used for matching. The experimental results have been carried out on both real and simulated data sets. They demonstrate the effectiveness of the proposed approach and their robustness with respect to different classes of commercial forms and shifted/rotated document images.

Index Terms *— Document management, Document Image Retrieval, Mathematic Morphology, Radon Transform, Dynamic Time Warping.*

1. Introduction

The increasing number of documents available in databases and digital libraries makes document retrieval a critical task of current document management systems. Traditional document retrieval systems – based on set-theoretic, algebraic and probabilistic models - require a document to be present in text form and the querying method is based on a specific textual content in the document [Doermann, 1998; Manning et al., 2009]. Whatever the model used, text-based document retrieval systems 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 error-prone process, specifically in the

case of multi-lingual/multi-font documents and poor-quality document images [Marukawa et al., 1997; Taghva et al., 1996; Lorpesti, 1996], as discussed in comprehensive surveys on this topic [Doermann, 1998; Mitra and Chaudhuri, 2000].

Along with the spreading of multimedia documents, it is useful to search a document on the basis of its structure and not only on the basis of its textual content. In such cases, methods adopted for document retrieval use feature vectors in which each feature is extracted from a specific region of the document image. For instance, some researchers 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 [Tzacheva et al., 2002]. In another approach, a hierarchical zoning strategy was proposed to overcome the problem of optimal grid selection, in order to face with the treatment of set of documents of different characteristics [Duygulu and Atalay, 2002]. A system that extracts text lines and describes the layout by means of relationships between pairs of these lines was also discussed in the literature [Huang et al., 2005], whereas some researchers used Brick Wall Coding Features (BWC) features to represent bounding boxes of the words [Erol et al., 2008]. 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. Several approaches can also be combined to identify a document, such as barcode, micro optical patterns, encoding hidden information, paper fingerprint, character recognition, local features and RFID. Owing to utilize SIFT. Unfortunately, the retrieval process is time consuming and requires special equipment [Liu and Liao, 2011].

In this paper a Layout-based Document Image Retrieval (LDR) system is presented, that is specifically devoted to commercial form processing, such as invoices, waybills, receipts, etc., in which layout is strongly characterized by a grid-structure. In fact, in these 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. In the first step the system uses a technique based on mathematical morphology for removing textual components from the document image and for extracting the grid-based structure in the document layout. Subsequently Radon Transform is used to obtain the feature vector characterizing the specific grid structure of the document. Dynamic Time Warping (DTW) is finally adopted to perform document matching.

The paper is organized as follow. The architecture of the system is presented in Section 2. Section 3 describes the preprocessing phase, which uses operators of mathematical morphology. The feature extraction phase is presented in Section 4. In Section 5 the matching process based on DTW is discussed while the decision combination process is illustrated in Section 6. The experimental results are reported and discussed in Section 7. Section 8

presents the conclusion of the work and highlights some directions for further research.

2. The Radon Transform for Layout-based Document Image Retrieval

The LDR system presented in this paper is based on three main phases: Acquisition and Preprocessing; Feature Extraction; Matching. 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. More precisely, the input document is acquired as a standard 256 gray-level – 100dpi PDF file. Figures 1 shows an input document concerning a real invoice.

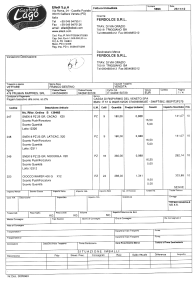


Figure 1. Input document image I=I(x,y).

Successively, after noise removal, document is resampled to 100 dpi and grid-based structure is extracted by mathematical morphology [Serra, 1982]. More precisely, let I=I(x,y) be the document image ($1 \le x \le X$, $1 \le y \le Y$) and let be

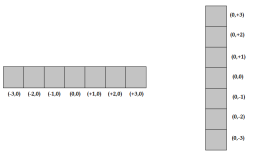
- B_{hor} the horizontal structure element defined as (see Figure 2a):

$B=\{(-s,0), \ldots, (-1,0), (0,0), (1,0), \ldots , (s,0)\}$;

- B_{ver} the horizontal structure element defined as (see Figure 2b):

$B=\{(0,-s), \ldots, (0,-1), (0,0), (0,1), \ldots , (0,s)\}$;

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



(a) (b)

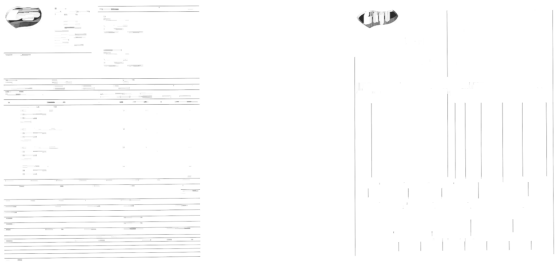
Figure 2. Structure elements (s=3): (a) B_{hor} , (b) B_{ver}.

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 (see Figure 3):

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

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

being "•" the closure operator, while "⊕" and "Θ" indicate respectively Minkowski sum and difference.



(a) (b)

Figure 3. 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:

$$I^* = I_{hor} \text{ XOR } I_{ver} \quad (2)$$

Figure 4 shows an example of document image after preprocessing.

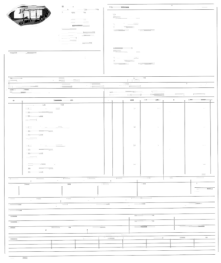


Figure 4. The preprocessed image I^*.

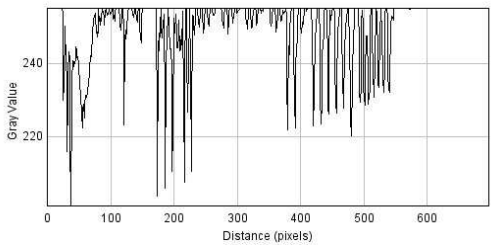
In the feature extraction step, in order to extract grid-based layout document images, the Radon Transform was considered. It is worth noting that the Radon Transform was extensively used in image analysis and has a number of important applications, like those related to MRI and computed tomography [Cormack, 1983; Deans, 1983]. The complete description of the Radon Transform is beyond the scope of this paper (see further details in [Jafari-Khouzani and Soltanian-Zadeh, 2005; Seo et al., 2004]). 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 (ρ – xcos ϑ - ysin ϑ) =0, for each ϑ and ρ. More precisely the Radon Transform of a function $I^*(x,y)$ in an Euclidean space is defined by [Hjouj and Kammler, 2008]:

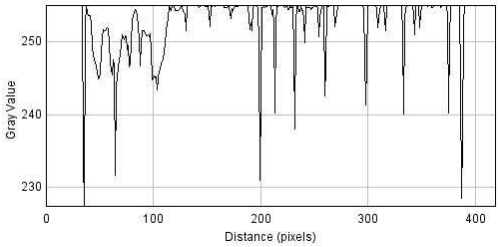
$$S_{\vartheta,\rho} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I^*(x,y)\cdot\delta(\rho - x\cos\vartheta - y\sin\vartheta)dxdy \quad (3)$$

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

Therefore, computing 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 5 presents the results of the Radon Transform applied to the preprocessed image I^* for the parameter values related to horizontal (ϑ=0 , ρ=0) and vertical (ϑ=π/2 , ρ=0) projections.



(a) Horizontal Projection



(b) Vertical Projection

Figure 5. Feature extraction by Radon Transform.

Dynamic Time Warping (DTW) is used for matching the feature vectors extracted by the radon transform from two document images. More precisely, 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 [Salvador and Chan, 2004; Lemire, 2009]:

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

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

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

The DTW detects the warping function $W^*(S^r,S^t)= c^*_1,c^*_2,\ldots,c^*_{K^*}$ which satisfies the condition of [Salvador and Chan, 2004]:

- Monotonicity (i.e. $i_{k-1} \leq i_k$, $j_{k-1} \leq j_k$ for k=2,...K) (6a)
- Continuity (i.e. $i_k - i_{k-1} \leq 1$, $j_k - j_{k-1} \leq 1$ 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 [Salvador and Chan, 2004; Lemire, 2009]:

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

.

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

The matching procedure provides two distance-based ranked lists of documents, obtained respectively from horizontal and vertical projections. The decision making process obtains the final decision combining the two ranked lists using the Borda-count strategy [Kittler et al., 1998; Xu et al., 1992]. According to this strategy, let $D=\{D_1, D_2,\ldots, D_k,\ldots, D_K\}$ be 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 [Ho et al., 1994]:

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

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 [Ho et al., 1994].

3. Experimental Results

The experimental test was carried out on a dataset of 33 commercial forms belonging to 16 different categories. Figure 6 shows some examples of commercial forms in the dataset. In this case they belong to the category n. 3.

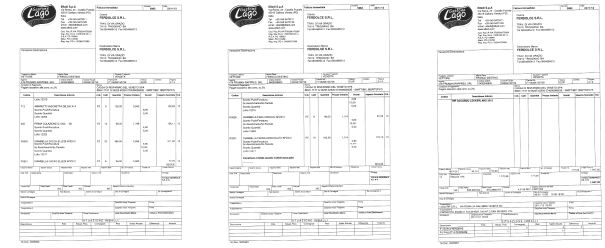


Figure 6. Examples of commercial forms of the same category.

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. Table 1 reports the number of forms for each category.

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

In the testing phase the leave-one-out method was considered to verify 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 follows [Huang et al., 2005]:

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

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.

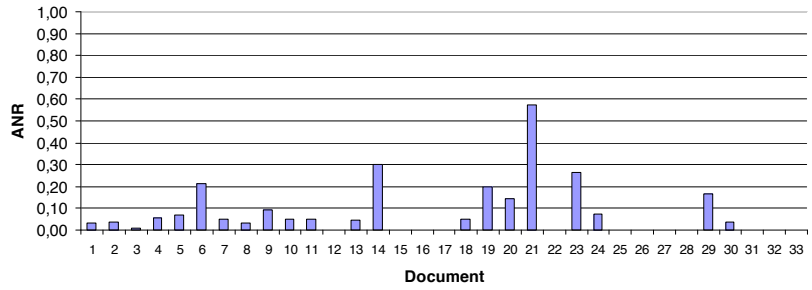


Figure 7. Experimental Results.

In order to estimate the robustness of the new approach two additional tests have been carried out using shifted and rotated document images as input.

When shifted documents are fed into the system the experimental results are shown in Figure 8. In this case a shift of 5 pixel is considered in the four main directions and the average result is computed. Also in this case the value of ANR is equal to 0.08, on average.

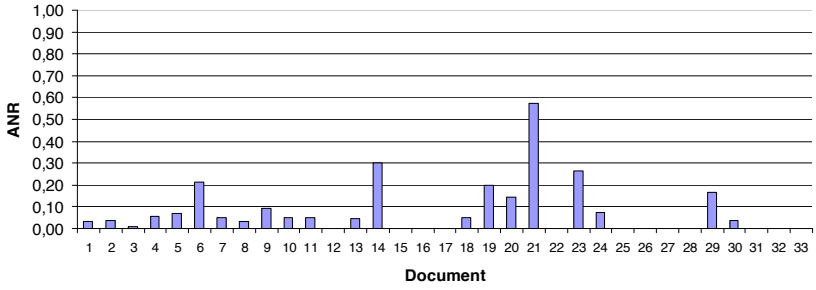


Figure 8. Experimental Results: shifted documents

Conversely, Figure 9 shows the results when rotated document images are fed into the system. In this case a rotation of 2° clockwise and anticlockwise is considered and the average result is computed. In this case the value of ANR is equal to 0.14 on average.

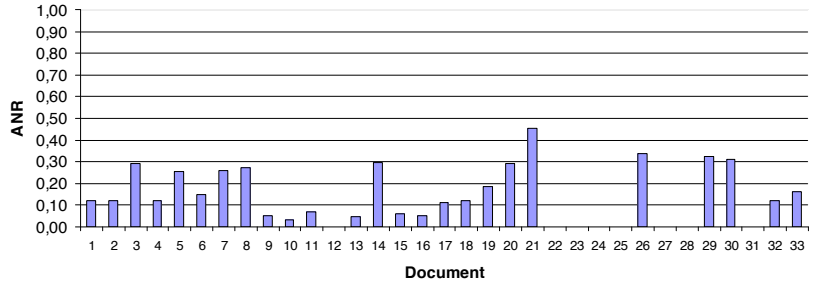


Figure 9. Experimental Results: rotated documents

Figure 10 shows the results when document images are shifted and rotated before to be fed into the system. In this case a shift of 5 pixels and a rotation of 2° clockwise and anticlockwise are considered. In t his case the value of ANR is equal to 0.15 on average.

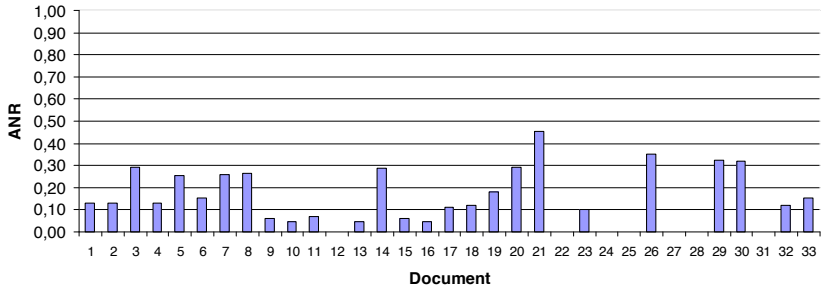


Figure 10. Experimental Results: shifted and rotated documents

4. Experimental Results

This paper presented a new system for layout-based document image retrieval. The system was specifically designed for retrieval of commercial forms as invoices, waybills and receipts, to optimize document management and sustainability. It used a morphologic filtering technique and the Radon Transform to obtain multiple document image descriptions. Document matching was then performed by Dynamic Time Warping whereas the Borda-count decision combination strategy was used to combine multiple decisions.

The experimental results, carried out on a dataset of real commercial documents, demonstrate the effectiveness of the proposed solutions and the robustness also with respect to shifted and rotated input document images.

5. References

Lopresti, D., Robust Retrieval of noisy text, *Proc. of ADL'96,* 1996, 76–85.

Cormack, A. M., Computed tomography: Some history and recent developments, *Proc. Symposia in Applied Mathematics*, 27, 1983, 35–42.

Deans, S. R., *The Radon Transform and Some of Its Applications*, New York: Wiley, 1983.

Doermann, D. , The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Understanding,* 70, 3, 1998, 287–298.

Duygulu, P., Atalay, V., A Hierarchical Representation of Form Documents for Identification and Retrieval, *International Journal on Document Analysis and Recognition,* 5, 1, 2002, 17–27.

Erol, B., Ant´unez, E., Hull, J. J., Hotpaper: multimedia interaction with paper using mobile phones, *Proceeding of the 16th ACM international conference on Multimedia*, 2008, 399–408.

Hjouj, F., Kammler, D. W., Identification of Reflected, Scaled, Translated, and Rotated Objects From Their Radon Projections, IEEE Trans. Image Processing, 17, 3, 2008, 301-310.

Ho, T.K., Hull, J.J., Srihari, S.N., Decision combination in multiple classifier systems, IEEE Trans. Pattern Anal. Mach. Intell., 16, 1, 1994, 66–75.

Huang, M., Dementhon, D., Doermann, D., Golebiowski, L., Document ranking by layout relevance, *Proc. 8th ICDAR*, 2005, 362–366.

Jafari-Khouzani, K., Soltanian-Zadeh, H., Radon Transform orientation estimation for rotation invariant texture analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27, 6, 2005, 1004–1008.

Kittler, J., Hatef, M., Duin, R.P.W., Matias, J., On combining classifiers, IEEE Trans. on Pattern Analysis Machine Intelligence, 20, 3, 1998, 226-239.

Lemire, D., Faster Retrieval with a Two-Pass Dynamic-Time-Warping Lower Bound, Pattern Recognition, 42, 9, 2009, 2169-2180.

Liu, Q., Liao, C., PaperUI, Proceeding of the 4th International Workshop on Camera-Based Document Analysis and Recognition, 2011, 3–10.

Manning, C. D. , Raghavan, P. , Schütze, H., An Introduction to Information Retrieval, Cambridge Press, 2009.

Marukawa, K., Hu, T., Fujisawa, H., Shima, Y., Document retrieval tolerating character recognition errors - Evaluation and application, *Pattern Recognition,* 30, 8, 1997, 1361-1371.

Mitra, M., Chaudhuri, B., Information retrieval from documents: A Survey, *Information Retrieval,* 2, 2/3, 2000, 141–163.

Salvador, S., Chan, P., Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, *Proc. KDD Workshop on Mining Temporal and Sequential Data*, 2004, 70-80.

Seo, S., Haitsma, J., Kalker, T., Yoo, C. D., A robust image fingerprinting system using the Radon transforms, *Signal Process.: Image Commun.*, 19, 4, 2004, 325–339.

Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press, 1982.

Taghva, K., Borsack, J., Condit, A., "Evaluation of model-based retrieval effectiveness with OCR text", *ACM TOIS,* 14, 1, 1996, 64–93.

Tzacheva, A., El-Sonbaty, Y., El-Kwae, A., Document Image Matching Using a Maximal Grid Approach, *Proceedings of the SPIE Document Recognition and Retrieval IX,* 2002, 121-128.

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, 1992, 418-435.