Multi-domain intelligent system for document image retrieval

Donato Barbuzzi*, Alessandro Massaro, Angelo Galiano and Leonardo Pellicani

Dyrecta Lab srl,
Via V. Simplicio, 45, 70014,
Conversano (BA), Italy
Email: donato.barbuzzi@dyrecta.com
Email: alessandro.massaro@dyrecta.com
Email: maurizio.galiano@dyrecta.com
Email: leonardo.pellicani@dyrecta.com
*Corresponding author

Giuseppe Pirlo

Bari University,
Via E. Orabona, 4 – 70125,
Bari, Italy
Email: giuseppe.pirlo@uniba.it

Matteo Saggese

Kibematsrl,
Via del Pescaccio, 30 – 00166,
Rome, Italy
Email: m.saggese@kibemat.it

Abstract: This paper presents an experimental analysis on document image retrieval using a multi-domain intelligent system. More specifically, on the same document image, the combination of three different domains: layout, logo and signature are discussed. This new method analyses every single decision provided by multi-domain system so that, in the training phase, a new sample classified with a dissimilar confidence to the previous trained samples is used to update the system. DTW, Euclidean distance and cosine similarity have been used, respectively for the analysis of layout, logo and signature. Finally, the weighted combination of individual decisions was considered. The experimental results, carried out on 30 rotated forms belonging to 13 different companies, demonstrate the superiority of the proposed approach with respect to single-domain retrieval systems, based on the ANR performance index. The ANR parameter is able to evaluate the multi-domain system.

Keywords: document management system; document image retrieval; multi-expert intelligent system; feedback-based strategy; instance selection.

Reference to this paper should be made as follows: Barbuzzi, D., Massaro, A., Galiano, A., Pellicani, L., Pirlo, G. and Saggese, M. (2019) 'Multi-domain intelligent system for document image retrieval', *Int. J. Adaptive and Innovative Systems*, Vol. 2, No. 4, pp.282–297.

Biographical notes: Donato Barbuzzi received the Computer Science degree cum laude and PhD from University of Bari Aldo Moro, in 2011 and 2015, respectively. He worked from September to December 2011 as a collaborator in the Interfaculty Center RetePuglia. His current research interest is in the field of multi-expert systems for pattern recognition. Currently, he is a Systems Analyst in the Research Institute at Dyrecta Lab s.r.l.

Alessandro Massaro received his Laurea degree in Electronic Engineering and the PhD in Telecommunication Engineering from the University of Ancona, Italy, in 2001 and 2004, respectively. From 2004 to 2006, he worked as post-doc in University of Ancona. In 2006, he spent one year in research and development at medical and industrial optics industry (endoscope design and optical systems). He worked for two years with CNR-INFM, University of Salento, as PI. He was the Team Leader in Robotics Lab. and in Smart Materials Platforms of the Center for Bio-Molecular Nanotechnology of IIT, Arnesano, Lecce, Italy. His research interests are in photonic band gap circuits, in CAD tools of integrated optics, MEMS technology and systems, and smart material implementation. He worked with the CNRIMIP/Nanotech of Bari, Italy, in diamond and nanodiamond technology for aerospace applications. Currently, he is responsible for the Research and Development Laboratory of Dyrecta Lab s.r.l.

Angelo Galiano received his MSc in Education in 2009 from Unipegaso University (Naples, Italy). He has worked since 1996 in the field of ICT for many companies (including Enel spa) covering roles of gradually increasing importance. He founded, on 2001, Dyrecta an ICT SME, successively grown in Dyrecta Lab when, on 2011, the company obtained the recognition of certified private research laboratory by MIUR (Italian Ministry of Research). He is the CEO of Dyrecta Lab. He is the Scientific Coordinator as well as member of the scientific committee of many (public and private) industrial research projects. His research interests are in the field of HCI, computer vision and augmented reality.

Leonardo Pellicani is graduating in Computer Science at University of Bari Aldo Moro. He attended different courses in the Faculty of Statistics. Successively, he worked both as teacher for different training courses and as consultant in the field of informatics. Moreover, he worked as an administrator in the field of tourism. Currently, he is both Sales Manager and Senior Researcher in Dyrecta Lab s.r.l. His main responsibilities are related to the projects and problem solving.

Giuseppe Pirlo received his Computer Science degree cum laude in 1986. Since 1991, he has been an Assistant Professor at the University of Bari, where he is currently an Associate Professor. His interests cover the areas of document processing and pattern recognition, biometry, computer arithmetic and e-learning. He has developed several scientific projects and published over 200 papers. He is an Associate Editor of *IEEE Trans. on Human Machine Systems* and Reviewer for *IEEE T-PAMI*, *IEEE T-SMC*, *IEEE T-EC*, *PR*, *IJDAR*, *IPL*. He was general co-chair of ICFHR 2012 and EAHSP 2013. He was a Guest Editor of Handwriting Recognition and other PR Applications of the *PR Journal* and Handwriting Biometrics of the *Biometrics Journal*. He is an IEEE and IAPR member.

Matteo Saggese studied Economics at University of Rome Tor Vergata. He worked for many years in technologies applied to process reengineering, information systems, telecommunications, and security. He was a manager in private companies in Italy, developing hi-tech and innovative projects for services companies, for people and health, control management, remote systems control and their security. He was a teacher in many training events. At this time, he is working on system and process management for an Italian private company in Rome for research and development of innovative solutions and services. At the same time, he works on industrial project and company development for another small Italian company based in Rome too.

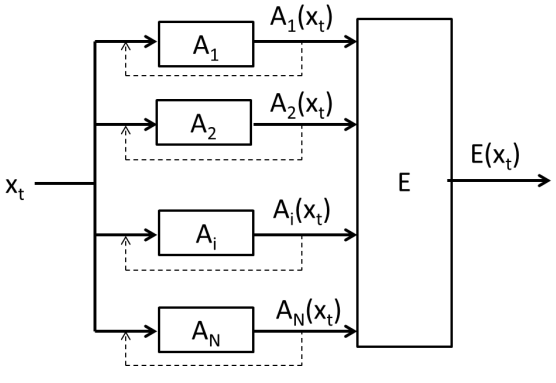
1 Introduction

Currently, information retrieval (IR) systems are mainly based on textual queries, with obvious limitations related both to flexibility of search options and to time for information discovery. Conversely, multi-expert systems were used for two reasons:

1 to capture the variability of the numerous patterns

2 to select the most profitable samples when new data become available (Pirlo et al., 2009, 2013; Impedovo and Barbuzzi, 2014; Barbuzzi et al., 2013).

Figure 1 Feedback in a multi-expert parallel system



The main contribution of this paper is to propose innovative strategies for document retrieval, respect to the state-of-the-art, based on structural, graphical and handwriting analysis according to a multi-expert scenario (see Figure 1). Therefore, queries will be performed respect to the retrieving of layout, logo and/or signature of a document image. Moreover, in the training phase, a single expert will be iteratively retrained by enlarging the original training set with new document images, respect to a specific retraining rule. The feedback-based strategy (or retraining rule) reinforces the classifier's knowledge base, when the classified image is dissimilar respect to its original training set (Barbuzzi et al., 2014).

As already discussed in literature, the achievement of better performance depends on the original training set, but also on the combination strategy of the multi-expert system,

on the data distribution and similarity between samples in the training set, feedback set and test set (Impedovo and Barbuzzi, 2014; Barbuzzi et al., 2015).

The proposed approach has the fundamental objective of minimising the average normalised rank (ANR), using few training data to avoid overfitting. The experimental test has been carried out on 30 rotated forms belonging to 13 different companies.

The paper is organised as follows. Section 2 presents a brief overview on the state-of-the-art in the field of IR. Section 3 describes in detail the feedback-based strategy. Section 4 shows the operating conditions. Section 5 analyses the experimental results and, finally, Section 6 summarises the work.

2 Information retrieval – overview

The main function, in IR, is based on search criteria (named query) within a document collection. In traditional systems, the query is textual and it contains the structured information in order to find the relevant documents. Moreover, this information allows to find the ranking documents according to the specific score.

Generally, the 'document retrieval' models are analysed as 'bag of words' style. In this case the models ignore both position of terms within the document and document structure (Ko, 2012).

It is well-known that a document retrieval system is characterised by:

- document collection D
- term dictionary V, where $t \in V$ is a term
- document $d \in D$, where d = $\{t_1, \ldots, t_n\}$ is a set of terms
- length $l(d) = n$, the number of terms included in d
- frequency tf_{ij} of a term t_i within the document d_j
- weight w_{ij} of a term t_i within the document d_j
- set Q of possible queries, where $q \in Q$ is a query defined as a set of terms $q = \{t_1, \ldots, t_k\}$
- function of document-query matching *rel*: $Q \times D \rightarrow R^+$ that associate at each document its relevance, respect to the query.

where for each term $t_i \in d_j$ a particular weight w_{ij} is associated. This weight corresponds to the relevance degree of t_i in d_j, for example it is based on empirical observation of the term within the document. So, each document can be seen as:

$$d_j = \left(\left(t_1, w_{1j} \right), \left(t_2, w_{2j} \right), \ldots, \left(t_{l(dj)}, w_{l(dj)} \right) \right) \tag{1}$$

Therefore, the function of document-query matching is used to estimate the similarity between document and query information.

$$rel\left(d_j, q \right) = f\left(w_{1j}, w_{2j}, \ldots, w_{l(dj)}, q \right) \tag{2}$$

In the bag of words approach, a document is considered as a container of terms and ($rel(d_j, q)$ is used to estimate the similarity between the terms belonging to the document and the information contained in the query (Hiemstra and Baeza-Yates, 2009).

A traditional matching model uses Boolean algebra as document-query matching. In this model, each term is or present or absent in a document and no weight is associated to it. The query's terms are logically combined, using the Boolean algebra operators, with the document terms in order to obtain an exact matching between the expressions. Unfortunately, this matching model does not allow a ranking between detected documents (Turtle, 1994).

In the vector space model, each document is seen as a weighed vector. The weight w_{ij} associated to the term t_i is calculated on the term frequency (TF) and Term Frequency – Inverse Document Frequency (TF-IDF) schemas. More specifically, in the TF schema, the weight associated to the term is calculated with respect to its frequency tf_{ij} in the document. Conversely, in the TF-IDF schema, the weight associated to the term is calculated with respect to its normalised frequency tf_{ij} in the document. So, if a term is present in many documents in the collection, its contribution is minimal $w_{ij} = tf_{ij} \times idf_i$ (Salton et al., 1975). Finally, in order to obtain the relevance degree, both the cosine distance and the Okapi measure are used as document-query matching (Jones et al., 2000). The ranking on the detected documents is based on the Bayesian rules (Wei and Croft, 2007).

Today the document search engines are oriented on non-text-based system, i.e., the query is not directly associated to the presence of terms and its features are extracted by a specific region of the document image. Tzacheva et al. (2001) suggest to transform a scanned form into a frameset composed of a number of cells. The maximal grid is the grid that encompasses all the horizontal and vertical lines in the form and can be easily generated from the cell coordinates. The number of cells from the original frameset, included in each of the cells created by the maximal grid, is then calculated. Those numbers are added for each row and column generating an array representation for the frameset. Duygulu and Atalay (2002) propose a hierarchical zoning strategy to overcome the problem of optimal grid selection, in order to identify and retrieve similar documents respect to the edit distances between the generated trees. Huang et al. (2005) present a system that extracts text lines and describe the layout by means of relationship between pairs of these lines. 'Mobile Retriever' (Liu and Doermann, 2008) aims to seamlessly link physical and digital documents by allowing users to snap a picture of the text of a document and to retrieve its electronic version from a database. Erol et al. (2008) use the brick wall coding (BWC) features that are local features which represent 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. Both 'Mobile Retriever' system and 'HotPaper' method do not work correctly when documents are written in Japanese and Chinese languages, in which words are not separated. The system of Liu and Liao (2011) combines several approach to identify a document, for example barcode, micro optical patterns, encoding hidden information, paper fingerprint, character recognition, local features and so on.

Unfortunately, the retrieval process is time consuming and it requires special equipment. So, this paper proposes to choose a single domain (layout, logo or signature) or their combination, in order to improve the retrieval efficiency.

3 Feedback-based strategy

3.1 Instance selection

Let:

- $P = \{x_k \mid k = 1, 2, \ldots, \mid P \mid\}$ be the documents set. In particular, P is used to train the system and its data are unlabelled.
- $P' = \{x_k \mid k = 1, 2, \ldots, \mid P' \mid\}$ be the rotated documents set. In particular, P' is used to test the system.
- A_i be the i^{th} classifier, $i = 1, 2, \ldots, N$.
- $F_i(x_k) = (F_{i,1}(x_k), F_{i,2}(x_k), \ldots, F_{i,r}(x_k), \ldots, F_{i,R}(x_k))$ be the numeral feature vector used by A_i to represent the specific pattern $x_k \in P$ (with R numeral features).
- KB_i be the knowledge base of A_i after the processing of new instance selected.
- E be the multi expert system which combines the individual classifier decisions in order to obtain the final classification result.

Initially, in the first stage ($s = 1$), the classifier's knowledge base A_i is empty. Therefore, it is initially defined as:

$$KB_i = \{\varnothing\} \tag{3}$$

Successively, the set P of unknown samples is provided to the multi-expert system both for classification and for learning.

So,

$$KB_i = \left\{ F_i(x_k) \middle| k = 1, 2, \ldots, \left| KB_i \right| \right\} \tag{4}$$

Therefore, KB_i is the numeral feature vector set of the i^{th} classifier for the k patterns that belongs to the knowledge base (KB_i).

P' is considered as set of real cases for testing in order to avoid biased or too optimistic results. When considering new data (samples of P), in order to inspect and take advantage of the behaviour of the single classifier in a multi-expert scenario, the following simple strategy is proposed and evaluated in this work:

Being:

- $s(A_i(x_k))$ the (z-score) normalised distance between the new sample x_t and the entire knowledge base (KB_i) of the classifier A_i
- $s(A_i(x_t))$ the (z-score) normalised distance between the new sample x_t and itself;
- τ the acceptance threshold.

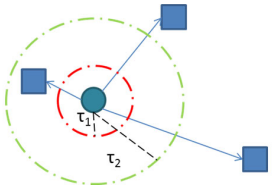
The following rule chooses the element for retraining:

$$\left\{ x_t \middle| \forall x_k \in P \wedge \left(s\left(A_i(x_k)\right) - s\left(A_i(x_t)\right) > \tau \right) \right\}$$

Figure 2 shows the retraining rule applied to the new sample (represented by the circle in the figure) respect to the thresholds τ_1 and τ_2. In the first case (τ_1), the new sample is

added to the expert's knowledge base (represented by the squares in the figure) because it disagrees with each other sample of the training set. While, in the second case (τ_2) the new sample is not added to the expert's knowledge base.

Figure 2 Retraining rule respect to the threshold τ (see online version for colours)



$A_i(x_k)$	mean	standard deviation	$A_i(x_t)$
1.50	5.00	3.63	0.00
10.00			
3.50			

$s(A_i(x_k))$	$s(A_i(x_t))$
-0.96	-1.38
1.38	
-0.41	

τ_1	0.40
τ_2	3.00

More specifically (see Figure 2), using τ_1 as acceptance threshold, it is possible to observe that no training data is included in new sample's neighbourhood, represented by red dashed line. So, the expert accepts the input data because it is dissimilar to the other data in expert's knowledge base. Instead, using τ_2 as acceptance threshold, the expert does not accept the new data because it is similar to the training sample within its neighbourhood (see green dashed line).

In equation (5), hereafter, A_i is updated with those patters where the normalised output of A_i (in terms of distance) disagrees with each other output of its knowledge base respect to a threshold τ.

The system is able to select only samples to be used for the updating process, in order to improve the multi-expert's performance.

3.2 Algorithm

In order to describe this feedback-based strategy, it is detailed in Algorithm 1. A sample in P is used to enlarge the knowledge base of A_i, if and only if its normalised distance by each other stored image in the knowledge base is different, respect to a threshold τ.

Algorithm 1: Feedback-based retraining process

1. Given:
   - $P = \{x_1, \ldots, x_M, \ldots, x_T\}$: document set
   - KB_i: the knowledge base of the expert A_i, $i = 1, 2, \ldots, N$
2. For each selected sample by the user, $h = M + 1, \ldots, T$
3. For each expert:
4. Determine for the sample $x_h \in P$ an output score (normalised distance) of the expert on its knowledge base: $s(A_i(x_h, KB_i)$
5. Apply the rule (5):

```
IF (5) is true THEN
    Add xh in KBi
end IF
End for
End for.
```

Obviously, many new samples will not give any feedback to a specific expert. This aspect depends on: the acceptance threshold, the classification algorithm and the training set at the previous step. So, the final goal of this paper is to find a good compromise to investigate on the retraining rule in order to obtain successful results.

4 Operating conditions

In this paper 30 real commercial forms, belonging to 13 different companies, have been used (Figure 3 shows some examples of these forms in the dataset).

Figure 3 Examples of commercial forms

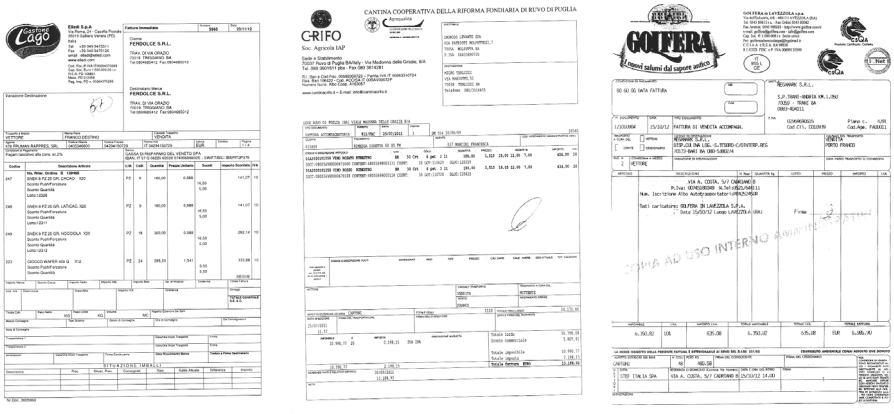
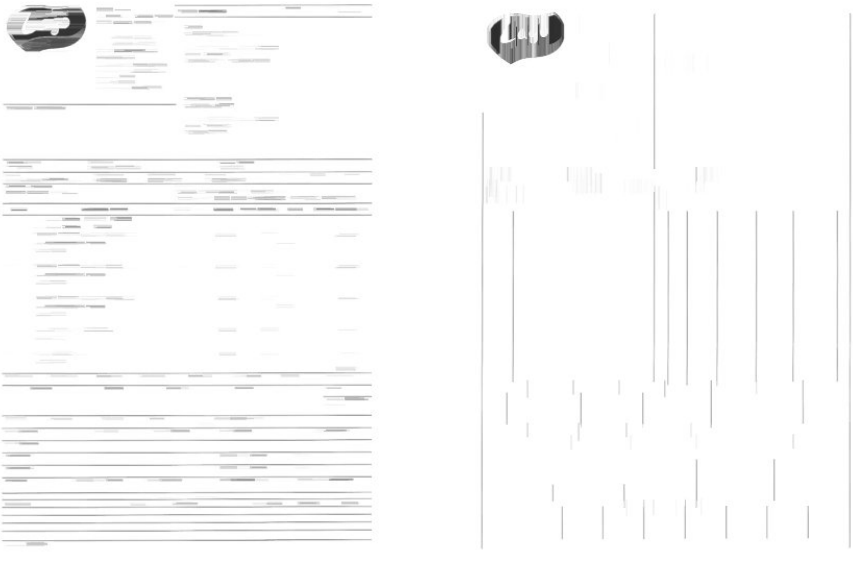


Figure 4 Examples of filtered image



(a) (b)

Documents have been acquired at a resolution of 100 dpi and a depth of 8 bit (256 grey-level). Moreover, each domain (layout, logo and/or signature) has been pre-processed respect to innovative and original techniques.

For layout analysis, two grid-based structures have been extracted by mathematical morphology. More specifically, a *closing* operator with structuring elements (SEs) of *horizontal* and *vertical lines* have been applied, respectively (see Figure 4).

In the feature extraction step, in order to extract the numerical feature vector by each grid-based structure, the radon transform has been considered. Figure 5 shows the results of the radon transform applied to the image in Figure 4(a). While Figure 6 illustrates the results of the radon transform applied to the image in Figure 4(b).

Figure 5 Horizontal projection of Figure 4(a) ($\vartheta = 0$ and $\rho = 0$)

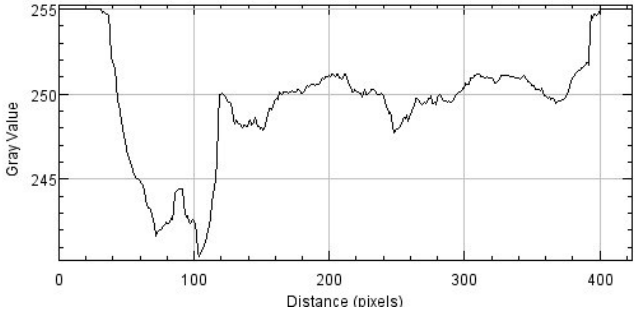
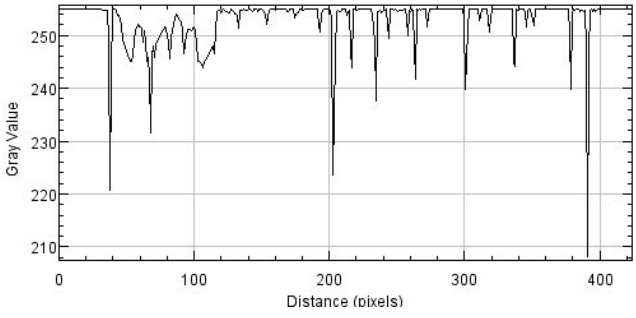


Figure 6 Vertical projection of Figure 4(b) ($\vartheta = \pi/2$ and $\rho = 0$)



Dynamic time warping (DTW), to match the feature vectors extracted by the radon transform from two document images, has been used. Further details, related to the layout-based analysis, have been described in the work (Pirlo et al., 2014) that represents our starting point for this research in the intelligent multi-domain system.

For logo analysis, in order to pre-process the image extracted manually, *Haar wavelet filter* and *adaptive median filter* have been used. More specifically, the Haar wavelet filter is efficacy for Gaussian noise removing, but it has some difficult related to the noise removing introduced by the sensor. So, a solution of this problem could be represented by an adaptive median filter. This filter type detects abnormal pixels from their context, compared to a multiple of the standard deviation of the neighbours. So, these abnormal pixels have been replaced with median value of their neighbours. Figure 7(b) shows the results of Haar adaptive median filter applied to the image in Figure 7(a). In particular, it

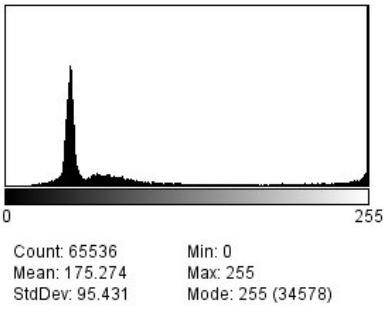
is evident from Figure 7 that clusters of noising pixels are filtered to provide a clearest image.

Figure 7 Examples of filtered image (see online version for colours)

(a) (b)

Concerning the feature extraction, the histogram of the pixel intensity values has been used. Figure 8 shows the grey-level histogram extracted from the image in Figure 7(b).

Figure 8 Grey-level histogram

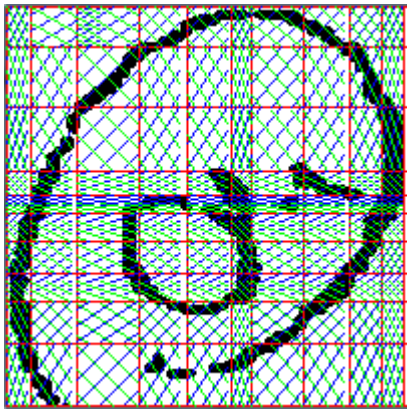


Euclidean distance to match the feature vectors extracted by two different logos has been adopted.

Handwriting signature is considered as a biometric trait which has been culturally accepted to verify automatically the identity of the people over the years (Plamondon and Lorette, 1989; Pirlo and Impedovo, 2013; Pirlo et al., 2015; Galbally et al., 2015). For signature analysis, it is acquired manually and the *equimass grid* for the segmentation, *intersection with lines* for the feature extraction and *cosine similarity* for the matching have been implemented. More specifically, after binarisation and normalisation of the signature, both the *median noise removal* algorithm for noise removal and the equimass grid technique (Impedovo et al., 2012) have been applied.

In the feature extraction step, in order to extract the numerical feature vector by pre-processed signature, the intersection with lines method has been considered. Four different grids (horizontal, vertical, diagonal +45° and –45°) have been superimposed on the signature image of Figure 9, which shows an example of pre-processed signature for feature extraction step.

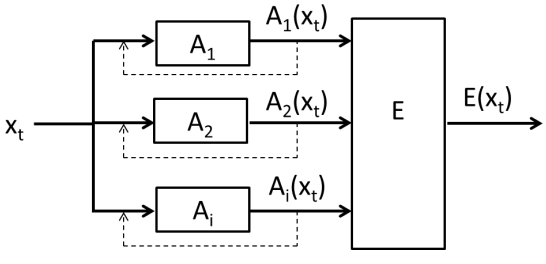
Figure 9 Example of pre-processed signature (see online version for colours)



To match the feature vectors extracted by two different signatures, *cosine similarity* algorithm has been used.

Finally, the multi-expert parallel system has been designed respect to the three different domains, named A_1 for layout analysis, A_2 for logo analysis and A_3 for signature analysis. The Borda count method has been considered in order to combine each list provided by the experts (Pirlo et al., 2014). Figure 10 shows the multi-domain intelligent system for document image retrieval.

Figure 10 Multi-domain intelligent system



In this multi-expert scenario, both the acceptance threshold and the weight associated at each expert play a crucial role in the selection of new patterns to be added to the experts' knowledge base. Thresholds (τ_1, τ_2, τ_3) are set to 0.50, 1.00 and 0.15 for layout, logo and signature, respectively chosen through a trial-and-error process, applied to the distances between two different z-score normalised images. While, for each expert, the weights have been calculated according to the equation know in literature (Polikar, 2007):

$$w_i = \left| \log\left(\frac{1}{\beta_i}\right) \right| \qquad \beta_i = \frac{\varepsilon_i}{1-\varepsilon_i} \tag{6}$$

So, after the analysis of each domain (layout, logo and signature), the following weights have been detected: $w_1 = 0.40$, $w_2 = 0.80$, $w_3 = 0.10$ for layout, logo and signature, respectively.

5 Results

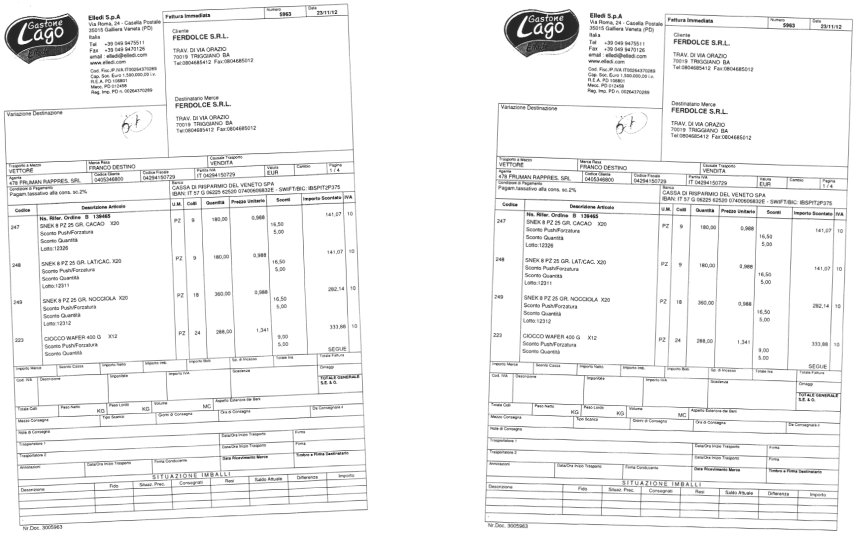
In the testing phase, 30 real commercial forms, belonging to 13 different companies (see an example of these documents in Figure 2 and their categorisation in Table 1) have been considered.

Table 1 Dataset of real documents

Company	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Number of documents	8	5	4	1	2	1	1	1	3	1	1	1	1

The same documents, rotated of –2° and +2° respect to the original form, are used to verify the efficacy of the multi-domain system (see an example of these rotated documents in Figure 11). The rotation is applied to simulate a real case of unaligned sheet on the scanner glass.

Figure 11 Example of rotated document of –2° and +2°



Finally, in order to estimate the quality of the ranked list provided by the system, for a given query, we considered the average normalised rank (ANR) that is defined as follows (Huang et al., 2005):

$$ANR = \frac{1}{N * N_w} * \sum_{i=1}^{N_w} \left(R_i - \frac{N_w + 1}{2} \right) \tag{7}$$

where

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

Of course, ANR values are in the range [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 of the ranked list (wrong position).

For each domain, Table 2 shows the selected samples respect to the instance selection strategy hereafter described in Algorithm 1. The order of entry documents starts from company 1 and ends to company 13.

Table 2 Training set 1

Company	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Layout	6	2	4	1	1	0	0	1	1	0	1	0	0
Logo	5	2	4	1	2	1	1	1	3	1	1	1	1
Signature	7	3	3	0	1	0	1	0	0	0	1	0	0

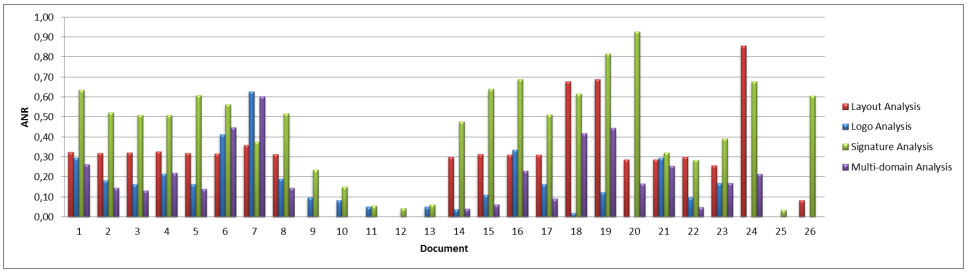
While Table 3 shows the selected samples, for each domain, from company 13 to company 1 in the training phase.

Table 3 Training set 2

Company	*13*	*12*	*11*	*10*	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
Layout	1	1	1	0	1	1	0	1	2	1	1	0	2
Logo	1	1	1	1	1	0	0	1	1	0	1	1	3
Signature	1	1	1	1	1	1	1	0	1	0	3	2	1

The final results, obtained by the weighted multi-domain system, have been considered respect to each training set (training 1 and training 2) and test set (rotated documents of –2° and rotated documents of +2°). So, the mean values of ANR are shown in Figure 12. Moreover, the original documents: d18, d21, d22 and d27 have been deleted. For final evaluation, this is useful because the aforementioned documents have not been stored in any domain's knowledge base in training phase.

Figure 12 Mean ANR in weighted multi-domain system (see online version for colours)



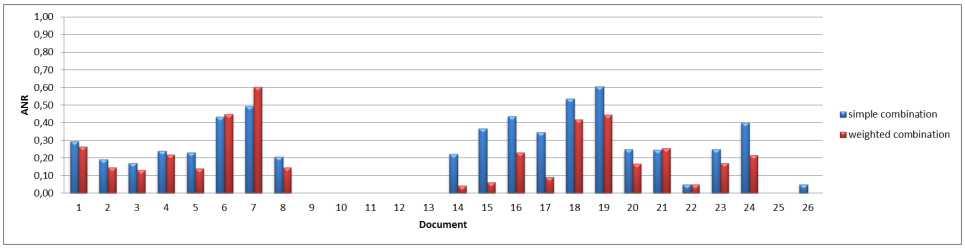
In this case, the weighted multi-domain system's performance are:

- more efficient of each single domain for retrieving of ten documents (mean ANR 0.15)
- equal at last one of the single domain for retrieving of other ten documents (mean ANR 0.04)
- less efficient of layout analysis for retrieving of other two documents
- less efficient of logo analysis for retrieving of other four documents.

When the ANR is close to 0, it is possible to observe that multi-domain approach is similar to the single domain analysis. In this case, the use of the single domain could decrease the computational costs.

Finally, Figure 13 shows the efficacy of the weighted combination, in a multi-domain system, respect to the simple combination.

Figure 13 Simple and weighted combination (see online version for colours)



More specifically, the mean ANR in the simple multi-domain system is equal to 0.23, while the mean ANR in a weighted multi-domain system is equal to 0.16 with a system's performance improvement of 7%.

6 Conclusions

A new multi-expert intelligent system has been introduced for document retrieval, according to three different domains (layout, logo and signature) extracted from real commercial documents. More specifically, the adopted methodology analyses every single decision provided by multi-domain system so that, in the training phase, a new sample classified with a dissimilar confidence to the previous trained samples is used to update the system.

The experimental results show that document selection depends on the: *acceptance threshold*, *classification algorithm* and *training set at the previous step*.

The results demonstrate that the weighted multi-domain system is equal or more efficient respect to the single-domain analysis of 77%, for the retrieving purpose.

More specifically, the multi-domain intelligent system is more efficient for the retrieving of ten documents with an average normalised rank (ANR) of 0.15 respect to the single-domain analysis, it is equal for the retrieving of ten documents (mean ANR of 0.04) respect to the analysis of at least one domain and it is less efficient for the retrieving of two and four documents respect to the layout-based and logo-based analysis, respectively.

Finally the weighted multi-domain system improves the performance, in terms of the average normalised rank (ANR), of 7% respect to the simple (unweighted) multi-domain system.

In the future, this multi-domain intelligent system will be more investigated according to a large dataset to solve problems related to the big data.

Acknowledgements

Work supported by the project 'Ricerca Scientifica per l'implementazione di un sistema multi-dominio per il document retrieval' funded by the D.M. 593 – 8 August 2000 – Art. 14.

References

Barbuzzi, D., Mangini, F.M., Impedovo, D. and Pirlo, G. (2013) 'Sistema Multi-Esperto Intelligente per la Diagnosi del Tumore al Seno', *CONGRESSO NAZIONALE AICA 2013 'Frontiere Digitali: dal Digital Divide alla Smart Society'*, 18–20 Settembre 2013, Salerno, Italy, pp.97–106, Università degli Studi di Salerno, Fisciano (Salerno), ISBN: 9788898091164.

Barbuzzi, D., Pirlo, G. and Impedovo, D. (2014) 'About retraining rule in multi-expert intelligent system for semi-supervised learning using SVM classifiers', *International Journal of Signal and Imaging Systems Engineering*, Vol. 7, No. 4, pp.245–251.

Barbuzzi, D., Pirlo, G., Uchida, S., Frinken, V. and Impedovo, D. (2015) 'Similarity-based regularization for semi-supervised learning for handwritten digit recognition', *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR 2015, Tunis)*, pp.101–105, Nancy, France.

Duygulu, P. and Atalay, V. (2002) 'A hierarchical representation of form documents for identification and retrieval', *International Journal on Document Analysis and Recognition*, Vol. 5, No. 1, pp.17–27.

Erol, B., Antúnez, E. and Hull, J.J. (2008) 'HOTPAPER: multimedia interaction with paper using mobile phones', *Proceedings of the 16th ACM International Conference on Multimedia*, October, pp.399–408, ACM.

Galbally, J., Diaz-Cabrera, M., Ferrer, M.A., Gomez-Barrero, M., Morales, A. and Fierrez, J. (2015) 'On-line signature recognition through the combination of real dynamic data and synthetically generated static data', *Pattern Recognition*, September, Vol. 48, No. 9, pp.2921–2934.

Hiemstra, D. and Baeza-Yates, R. (2009)' Structured text retrieval models', *Encyclopedia of Database Systems*, pp.2868–2871, Springer, US.

Huang, M., DeMenthon, D., Doermann, D., Golebiowski, L. and Hamilton, B.A. (2005) 'Document ranking by layout relevance', *Proceedings Eighth International Conference on Document Analysis and Recognition*, August, pp.362–366, IEEE.

Impedovo, D., Pirlo, G., Sarcinella, L., Stasolla, E. and Trullo, C.A. (2012) 'Analysis of stability in static signatures using cosine similarity', *2012 International Conference on Frontiers in Handwriting Recognition (ICFHR)*, September, pp.231–235, IEEE.

Impedovo, S. and Barbuzzi, D. (2014) 'Instance selection for semi-supervised learning in multi-expert systems: a comparative analysis', *Journal of Next Generation Information Technology*, Vol. 5, No. 4, p.61.

Jones, K.S., Walker, S. and Robertson, S.E. (2000) 'A probabilistic model of information retrieval: development and comparative experiments: Part 2', *Information Processing & Management*, Vol. 36, No. 6, pp.809–840.

Ko, Y. (2012) 'A study of term weighting schemes using class information for text classification', *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM.

Liu, Q. and Liao, C. (2011) 'PaperUI', *4th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR 2011)*, Springer Berlin Heidelberg, pp.83–100.

Liu, X. and Doermann, D. (2008) 'Mobile retriever: access to digital documents from their physical source', *International Journal of Document Analysis and Recognition (IJDAR)*, Vol. 11, No. 1, pp.19–27.

Pirlo, G. and Impedovo, D. (2013) 'Verification of static signatures by optical flow analysis', *IEEE Transactions on Human-Machine Systems*, Vol. 43, No. 5, pp.499–505.

Pirlo, G., Chimienti, M., Dassisti, M., Impedovo, D. and Galiano, A. (2014) 'A layout-analysis based system for document image retrieval!', *Mondo Digitale*, Vol. 13, No. 49(1), pp.1–16.

Pirlo, G., Cuccovillo, V., Diaz-Cabrera, M., Impedovo, D. and Mignone, P. (2015) 'Multidomain verification of dynamic signatures using local stability analysis', *IEEE Transactions on Human-Machine Systems*, Vol. 45, No. 6, pp.805–810.

Pirlo, G., Impedovo, D. and Barbuzzi, D. (2013) 'Learning strategies for knowledge-base updating in online signature verification systems', *New Trends in Image Analysis and Processing – ICIAP 2013*, pp.86–94, Springer, Berlin Heidelberg.

Pirlo, G., Trullo, C.A. and Impedovo, D. (2009) 'A feedback-based multi-classifier system', *10th International Conference on Document Analysis and Recognition 2009, ICDAR'09*, July, pp.713–717, IEEE.

Plamondon, R. and Lorette, G. (1989) 'Automatic signature verification and writer identification – the state of the art', *Pattern Recognition*, Vol. 22, No. 2, pp.107–131.

Polikar, R. (2007) 'Bootstrap-inspired techniques in computational intelligence', *IEEE Signal Processing Magazine*, Vol. 24, No. 4, pp.59–72.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Communications of the ACM*, Vol. 18, No. 11, pp.613–620.

Turtle, H. (1994) 'Natural language vs. Boolean query evaluation: a comparison of retrieval performance', *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.212–220, Springer-Verlag Inc., New York.

Tzacheva, A., El-Sonbaty, Y. and El-Kwae, E.A. (2001) 'Document image matching using a maximal grid approach', *Electronic Imaging 2002*, December, pp.121–128, International Society for Optics and Photonics.

Wei, X. and Croft, W.B. (2007) *Modeling Term Associations for Ad-Hoc Retrieval Performance within Language Modeling Framework*, pp.52–63, Springer, Berlin Heidelberg.