Document Images Indexing with Relevance Feedback : an Application to Industrial Context

O. Augereau, N. Journet, J.-P. Domenger
Laboratoire Bordelais de Recherche en Informatique (LaBRI)
Université de Bordeaux, 351 Cours de la Libération
Talence, France
{*augereau,journet,domenger*}*@labri.fr*

***Abstract*—This article presents a new method to index document images. This work is done in an industrial context where thousands of document images are daily digitized, these images have to be sorted in different classes like payroll, various bills, information letters. We propose a software method which aims to accelerate this task. Usually, the number of document classes is *a priori* unknown. In this paper, we propose an automatic estimation of this class number. According to this class number, we use a clustering algorithm in order to group document images. After this step, we propose an assisted classification tool based on content based image retrieval method (CBIR). For each cluster, a *reference image* is automatically selected then considering a similarity measure, the other images are sorted and shown to the user. By interacting with the process, the user can reject wrong images. The user feedback is automatically taken into account to enhance the similarity measure by selecting features.**

The first tests show that, on average, databases are indexed 3 times faster with our assisted classification method than with a standard manual classification process.

***Keywords*-document image clustering; document retrieval; feature selection; relevance feedback; industrial application;**

I. INTRODUCTION

The work presented in this paper lies in an industrial context, where several thousands of documents such as human resource documents are daily scanned and manually indexed. Our goal is to simplify and accelerate the manual indexing of documents. As recently pointed out by Saund [1], this problematic is an important economic issue for document scanning companies.

Indexation in an industrial context is a real challenge because the document in the database are sometimes unknown in advance. Usually, classification plan (*i.e.* the number of classes and how to identify the different document classes) associated with the database is not clearly defined.

The difficulty of using document image classification methods of the state of art like those surveyed by Chen and Blostein [2], is that most of these techniques are supervised. This implies firstly to know the number of classes, and secondly to be able to build a representative ground truth of the distribution of the database that will be large enough to allow efficient learning. These two conditions are particularly difficult to satisfy when databases are large. An example of supervised approach is proposed by Shin *et al* in [3]. Authors suggest an approach based on decision trees. A decision tree is built using a part of the ground truth. In this learning step, documents are taken in order to create a sample as representative as possible. Then, this tree is used to rank the leftovers of the database. With a database of forms made of 5590 images (divided in 20 classes) and using a ground truth of 2000 images, the authors obtained an accuracy of 99.7%. In [4], authors suggest an approach based on neural networks. The database consists of 600 forms belonging to 5 classes. The ground truth is composed of 305 images. An accuracy of 92% is obtained. These two examples demonstrate that it is necessary to label approximatively half of the database before classifying the remaining documents. Even if accuracy of existing systems is very good, the classification results must be checked manually to obtain an accuracy precision of 100% which is an industrial constraint.

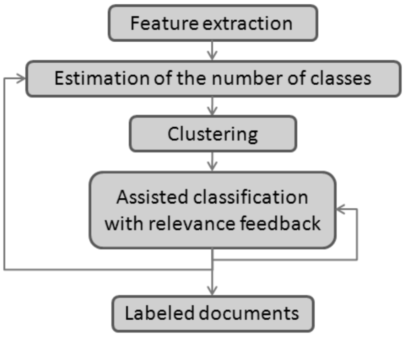
Feature extraction
Estimation of the number of classes
Clustering
Assisted classification with relevance feedback
Labeled documents

Figure 1. Indexing methodology in 4 steps : 1) extraction of set of descriptors, 2) estimation of the number of classes k, 3) clustering and determination of *reference images*, 4) assisted classification with relevance feedback.

To settle the problems mentioned above, we propose in this paper the following methodology (see figure 1). At first a set of descriptors is extracted, *e.g.* those of [3]. Many descriptors are extracted because we do not know in advance which ones will be relevant or not. Then the number of classes k of the database is estimated. As pointed out by Sugar and James [5], finding the "true" number of classes

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.240

in a database is one of the most difficult problem in cluster analysis field. To solve this problem, we use a method based on the study of silhouette [6] after an unsupervised classification algorithm. For a considered k, it is possible to use unsupervised classification algorithm like PAM [7] or K-means [8]. These algorithms build a clustering where the mean silhouette is calculated. The k that produces the best silhouette is selected. Once k is estimated, *reference images* are extracted. A *reference image* is the center of a cluster for K-means or the medoid for PAM. After that, we propose an assisted classification module based on CBIR techniques where query images are the *reference images* automatically extracted previously. For each query image, similar images are presented to the user by using a similarity measure between the *reference image* and the other images of the database. The similarity measure is obtained from clustering step. By this way, manual labeling is made simpler and faster. In order to improve the quality of the similarity measure between documents, some features are selected according to a relevance feedback learning. An introduction about feature selection could be read in [9] and [10].

Finally, when documents belonging to classes of *reference images* are labeled, unlabeled remaining documents are processed by looping previous steps one more time (see figure 1).

Using our assisted classification method, first tests show that a database is on average labeled 3.4 times faster than with a standard manual classification (see table III).

In the section II, an algorithm for estimating the numbers of classes in real industrial databases will be described. Section III details our assisted classification method and the use of relevance feedback. We also explain how it is possible to classify all database by iteratively compute an other k value and applying an other assisted classification on remaining documents. Finally, we will conclude in section IV by discussing about leads to improve indexing for companies.

II. Estimation of the number of classes

When a scanning company receive a database for the first time, the number of classes is not necessarily known. A first step is to estimate the number of classes in the database. In order to automatically estimate this number, measures of clustering quality like homogeneity (intra-cluster distance) and separation (inter-cluster distance) can be used.

The criterion of the mean silhouette described in [6] and [11] is a relevant measure for evaluating clustering quality. Each images are represented by a numerical vector computed from features described in [3]. The silhouette of an element x is calculated from the means of distances between x and the others elements $a(x)$ of the same cluster C_x. The minimum of mean distances between x and the others cluster $b(x)$ is calculated. The silhouette of x is then : $silh(x) = \frac{b(x)-a(x)}{\max(a(x),b(x))}$. Once the silhouette have been calculated for each element, mean silhouette can be calculated for a cluster: $S_{C_i} = \frac{\sum_{j\in C_i} silh(j)}{Card(C_i)}$. Finally, the mean of all clusters mean silhouettes is calculated $GS = \frac{\sum_{i=1}^{k} S_{C_i}}{k}$. If GS is near to 1, the clustering have a better quality because it have a high inter-cluster variability and a little intra-cluster variability. In order to select the number of clusters, GS is calculated for every values of k from 3 to K, where K is specified by the user (for the tests, K have been fixed to $\sqrt(L/2)$, where L is the number of documents in the database). The number k is chosen in order to maximize GS. Tests have been done on 5 databases extracted from industrial productions which were manually labeled. Databases $DB1$ and $DB3$ are made of invoices from different companies. For these databases, the number of clusters is equal to the numbers of companies. Databases $DB2$, $DB4$ and $DB5$ are made of various human resource documents like employment contract, performance appraisal, medical certificate, administrative forms, mutual organization papers, payroll, etc. Table I illustrates the pertinence of silhouette criteria to automatically estimate the number of clusters. We can see that for databases $DB1$ and $DB3$, the estimated k is close to the real k because invoices of a company have few variations so the distance intra-cluster is short. For databases $DB2$, $DB4$ and $DB5$, k is under-evaluated because some documents from different clusters are similar. The figure 2 shows the evolution of silhouette versus k for databases $DB1$ and $DB5$.

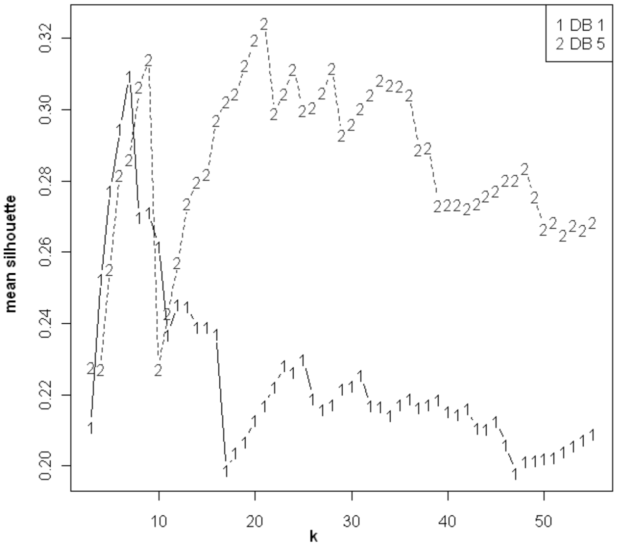

Figure 2. Silhouette versus the number of clusters k for $DB1$ (1509 documents, 7 classes) and $DB5$ (5962 documents, 48 classes). The estimate number of classes is determined by the maximum of mean silhouette, in this case: $k = 7$ for $DB1$ and $k = 21$ for $DB5$. It can be pointed out that mean silhouette of $DB5$ have a local maximum for $k = 48$.

F_1score is a measure used in order to check the accuracy of the clustering, it is based on precision p and recall r such as : $F_1score = 2 \cdot \frac{p*r}{p+r}$. F_1score (in table I) shows that the clustering accuracy is similar with estimated k and with real k. It can be explained by the fact that the clustering algorithms such as PAM or K-means tend to split large classes and to merge small classes to other classes.

Table I
ESTIMATION OF THE NUMBER OF CLASSES WITH SILHOUETTE AND ACCURACY OF PAM CLUSTERING.

		Real k		Estimated k	
Database	images	k	PAM F_1score	k	PAM F_1score
DB1	1509	7	0.7469	7	0.7469
DB2	883	19	0.6746	11	0.6668
DB3	2574	33	0.5778	35	0.5864
DB4	3352	30	0.4371	16	0.5660
DB5	5962	48	0.5823	21	0.5914

III. ASSISTED CLASSIFICATION WITH RELEVANCE FEEDBACK

The automatic estimation of the number of classes k enable to extract k *reference images* that best represent each cluster. These images are used as query images. The assisted classification (see figure 3) is carried out by showing to the user a query image accompanied by n_{Im} images which are most similar. The distance between feature vectors is computed with an Euclidean distance. User can indicates images which do not belong to the same class as the query image among the n_{Im} that are presented. Then, the next n_{Im} images are displayed (we call it a new iteration). During the interactive process, when more than n_{FS} cumulated images have been selected, a feature selection algorithm is executed. Therefore, after each user interaction the best features are selected.

Boruta [12] features selection algorithm is used in order to chose discriminant features among a whole set of features. Boruta is based on random forest construction. The algorithm iteratively removes the features which are less relevant than random variables. In practical terms, selected features are associated to a weight ”1” and unselected features are associated to a weight ”0”.

Considering that feature selection needs more than one element to be processed, we experimentally chose $n_{FS} = 5$. Finally, if more than n_{Wrong} images are considered by the user as wrong images, the next *reference image* is displayed. The same process is then executed until each *reference image* have been proposed. For the tests we chose: $n_{Im} = 50$ and $n_{Wrong} = 19$. The aim of relevance feedback learning with feature selection is to decrease distances between similar documents. Thereby, the assisted classification could propose more relevant documents and the percentage of labeled documents increase.

Table II and III summarize the results obtained in the classification of 5 human resource documents databases. The features consist mainly of statistics (sum, mean, median, standard deviation, minimum, maximum) of several characteristics like the area and perimeter of connected component, bounding box of connected component, blocs of text, table and image; but also the number of horizontal and vertical lines.

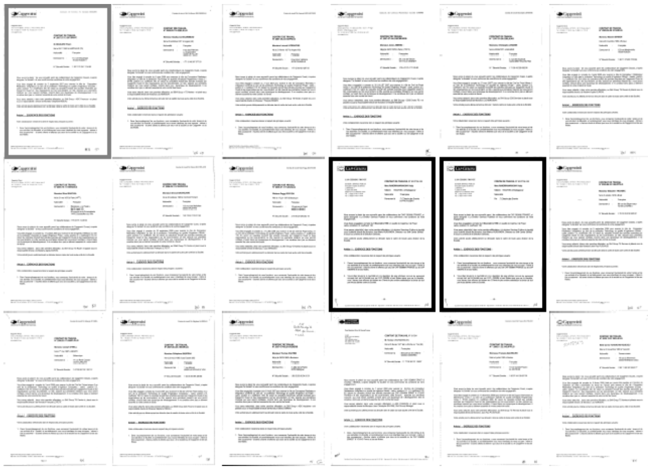

Figure 3. Assisted classification. Documents are sorted by increasing distances. The first image in a gray frame is the *reference image*. The two images in black frames are selected by the user because they do not belong to the same class. Images are extracted from $DB4$ when labeling the fourth *reference image* at the second iteration.

As we can see in figure 4, feature selection is useful for classes with a large number of images. This figure is the illustration of what happening precisely during the labeling of one *reference image* with and without feature selection. The x axis represents the number of iterations *i.e.* the number of times that n_{Im} images where displayed. Four values are displayed. The first one represents the number of remaining images that belonging to the class of the current *reference image*. The second and third ones represents the number of similar and dissimilar images displayed to the user. The number of ”wrong” images is the number of images selected by the user. If n_{Wrong} images are selected, the process is stopped. The process is also stopped if all images of the class are labeled. When there are more than n_{FS} dissimilar images feature selection will be launched and distances will be recalculated for the next iteration. If there are no new dissimilar images, the last feature selection is kept. When feature selection is activated, a cross is drawn on the fourth curve of the graph. For example, at the second iteration there are 46 documents of the same classes and 4 documents of an other class. Because there were also 1 selected document at first iteration features selection are used. In this case, 9 features among 99 are selected. Selected images have few differences because employment contract have few variations with years. For example one of selected features is the number of paragraphs which is larger in selected images than in images that belonging to the class of the *reference image*.

Table II
RESULTS OF ASSISTED CLASSIFICATION WITH AND WITHOUT FEATURE SELECTION. A MEAN GAIN OF 9.2760% IS OBSERVED.

		Percentage of labeled database	
database	images	without FS	with FS
DB1	1509	57.3227	85.6858
DB2	883	78.2559	81.0872
DB3	2574	76.8453	80.9634
DB4	3352	73.4486	79.8031
DB5	5962	60.2314	64.9446

Table III
TIME CONSUMED FOR LABELING DATABASES. ASSISTED CLASSIFICATION WITH FEATURE SELECTION ALLOWS TO DIVIDE LABELING TIME BY 3.4 ON AVERAGE.

		Time consumed for labeling (minutes)	
database	images	manual classification	assisted classification
DB1	1509	201.2	42.1
DB2	883	117.7	38.5
DB3	2574	343.2	101.1
DB4	3352	447.0	131.1
DB5	5962	794.9	338.7

In order to determine the relevance of using feature selection algorithm, we compare the labeling process with or without using feature selection for one loop of the whole process. The results in table II shows that the feature selection systematically increase the percentage of labeled documents. For each database, two values are calculated. The first value $without\ FS$ represents the percentage of documents labeled by assisted classification without using feature selection. The second value is $with\ FS$ and represents the percentage of documents labeled by assisted classification using feature selection. Thanks to this feature selection step, it is possible on average to increase the classification rate of more than 9%.

For companies, practical consequences is that assisted classification allows to label more quickly documents. Tests realized in production on several thousand images and several operators showed that, on average, it takes 8 seconds for a professional operator to label a random image from the database. With assisted classification, it takes an average of 25 seconds to select similar images to a query among the 50 images that are proposed. Table III shows that using assisted classification can divide the time for labeling document by more than 3 for one loop of the whole process.

After this first step of assisted classification process, about 20% of the database is not labeled. Unlabeled documents are expected to be classified by looping the whole process. k is estimated one more time but only with unlabeled documents. The process is looped until all documents are labeled.

Figure 5 shows the evolution of the percentage of labeled documents versus the number of loops. All databases are labeled in two or three loops. For example, 60% of the database is labeled in the first loop, 91% in the second loop and finally the whole database is labeled in the third loop.

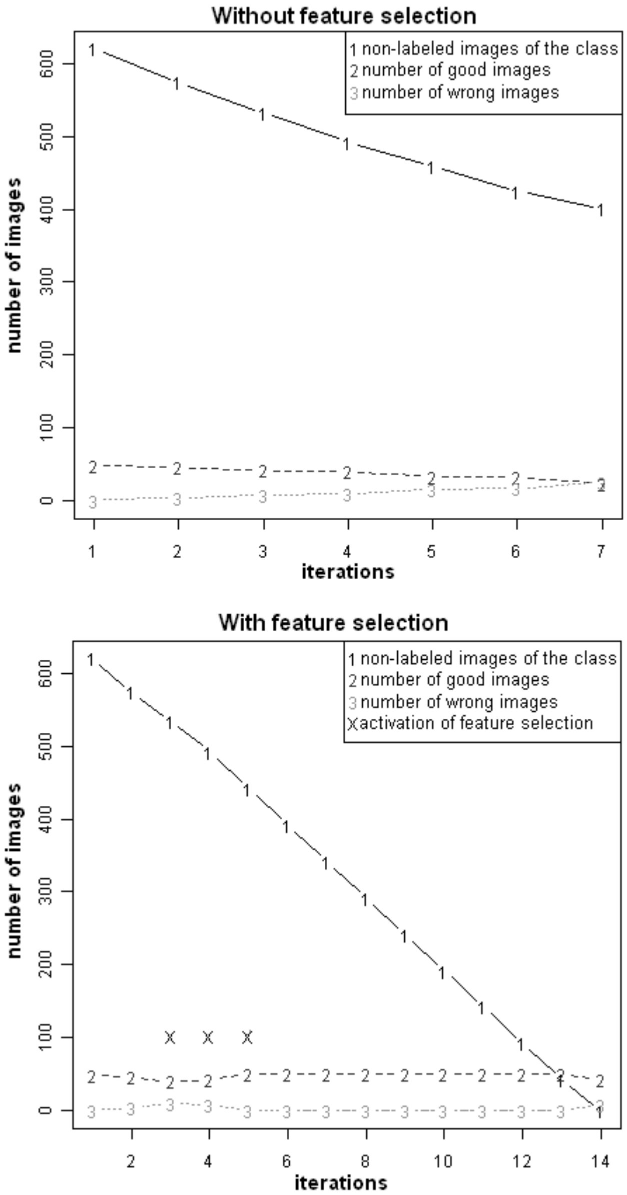

Figure 4. Usefulness of feature selection. The curves are plotted from the labeling of documents similar to the fourth medoid of $DB4$. On the first plot, no feature selection is used and 401 documents have not been labeled. On the second, feature selection is activated. All document of the class have been labeled. Distance between document are re-calculated in iteration number 3,4 and 5. After the fifth iteration feature selection are not computed again, distance remain the same as for the last feature selection.

It should be noted that each image is associated with a reference image corresponding to the same class. However, in case of over-segmentation, several *reference images* represent the same class of document. So this references must be manually merged together in order to assign them the same label in the end. Under-segmentation will imply at least one more loop because missing class will not be labeled this

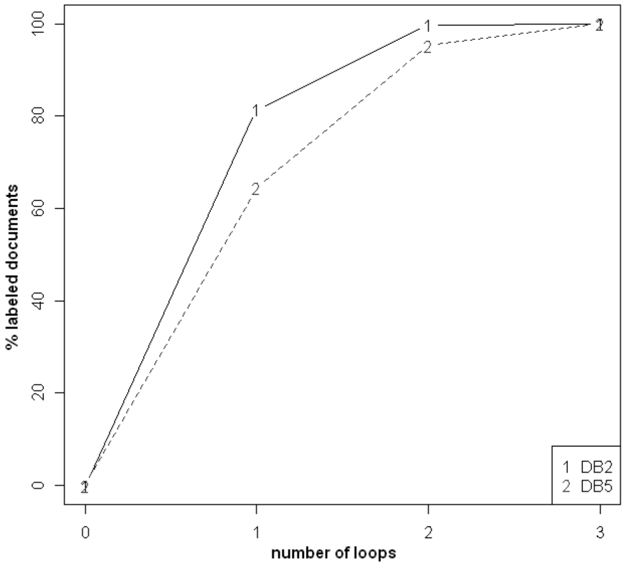

Figure 5. Percentage of labeled documents versus the number of loops. Databases are fully labeled in 3 loops, even $DB5$, the largest database.

time.

IV. Conclusion and future work

This paper presents a new methodology for classifying document images such as human resources documents. The first contribution of our proposal is to estimate the number of categories of documents that make up a database, where the methods of the state of art consider that this information is given by the user. The second contribution is to have established a system of indexing document images based on ”query by example” in which human is in the heart of the system. As the user labeled images in the database, our system allows the human operator to quickly index large amount of documents. The tests highlight a time saving process of indexing to the order of a factor 3.

In our future work, we would like to enhance the k estimation accuracy. For this, we plan to combine our current estimator with other estimators such as BIC [13]. We also would like to go further in the way of feature selection, by providing tools that can enable to select elements which are discriminating e.g. a logo, a table *etc.*.

Acknowledgment

We thank the company Gestform [1] for his involvement in this work. Gestform has provided more than 10,000 images from its production. Gestform also provided statistics about the time of manual classification. Now, our assisted classification software is used by Gestform for industrial application.

References

[1] E. Saund, “Scientific Challenges Underlying Production Document Processing,” *Proceedings of Document Recognition and Retrieval XVIII*, vol. 7874, p. 787402, 2011.

[2] N. Chen and D. Blostein, “A survey of document image classification: problem statement, classifier architecture and performance evaluation,” *International Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 1–16, 2007.

[3] C. Shin, D. Doermann, and A. Rosenfeld, “Classification of document pages using structure-based features,” *International Journal on Document Analysis and Recognition*, vol. 3, no. 4, pp. 232–247, 2001.

[4] F. Cesarini, M. Lastri, S. Marinai, and G. Soda, “Encoding of modified XY trees for document classification,” in *icdar*. Published by the IEEE Computer Society, 2001, p. 1131.

[5] C. Sugar and G. James, “Finding the Number of Clusters in a Dataset,” *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 750–763, 2003.

[6] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[7] L. Kaufman and P. Rousseeuw, “Finding groups in data: an introduction to cluster analysis,” *NY John Wiley & Sons*, 1990.

[8] J. e. a. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281-297. California, USA, 1967, p. 14.

[9] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[10] W. Jiang, G. Er, Q. Dai, L. Zhong, and Y. Hou, “Relevance feedback learning with feature selection in region-based image retrieval,” in *Proc. of IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, 2005, pp. 509–512.

[11] K. Pollard and M. Van Der Laan, “A method to identify significant clusters in gene expression data,” *Invited Proceedings of Sci2002*, vol. 2, pp. 318–325, 2002.

[12] M. B. Kursa and W. R. Rudnicki, “Feature selection with the boruta package,” *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[13] Q. Zhao, V. Hautamaki, and P. Franti, “Knee Point Detection in BIC for Detecting the Number of Clusters,” in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2008, pp. 664–673.

[1] www.gestform.com