Document Image Recognition and Classification

Jean-Philippe Domenger
University of Bordeaux

Document Image Recognition and Classification

Olivier Augereau
Supervisors : Jean-Philippe Domenger, Nicholas Journet

augereau@labri.fr
www.labri.fr/perso/augereau/index.php

Abstract. The subject of the thesis takes place in research field of image analysis and more particularly in document analysis. The goal is to explore new techniques for document recognition and classification by analyzing document image features and not using OCR. One particularity of this thesis is that it is link to a digitizing company. It give the benefit of using resources of the company like the millions of documents that are digitized each month. Furthermore, researches will be applied to an industrial context so it will be essential to understand and to ensure appliance to industrial constraints. The global problematic of the thesis is to find a way to accelerate or automatize identification and classification of documents.

Keywords: document image, classification, indexation, retrieval, recognition, industrial application

1 Features and classification algorithms

The first step of the thesis was to research techniques for image analysis i.e. document image feature extraction. Three main types of features can be set apart: visual features, structural features and textual features [1], see figure 1.

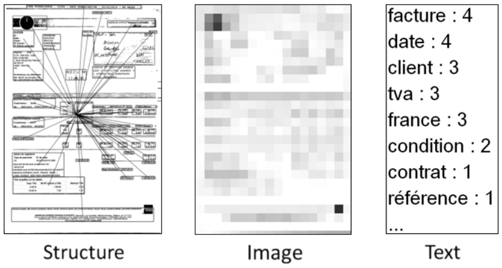

Fig. 1. Three examples of features.

Specific features must be extracted in order to analyze document zones like text, drawing, halftones, speckles, table, logo, etc. Keysers et al. [2] extract a list of nine features in order to classify image zones. Layout extraction and comparison with the cyclic polar page layout representation [3] is very interesting but do not provide very good result on my databases. Results show that similar documents have slightly different layout and different documents have similar layout. Maybe better result will be obtain if we could teach which blocks are fixed, which blocks move and which are significant or not. Visual, structural or textual approaches are not exclusive and can be combined to obtain better results.

Chen and Blostein [1] offer a detailed survey about document image classification. By studying this survey, it could be pointed out that most of classification techniques are supervised. The problem with supervised techniques is that they need to know in advance the number of classes and the type of documents in order to annotate half of the database to use it as data learning. This implies that labeling time can not be accelerate by more than two. Moreover, even if techniques give excellent results with very few mistakes, company must control all the images one by one to correct the wrong labeled images. Finally, saving time will be very low.

2 Classification of an unknown database

Instead of trying to automate classification task, we propose in [4] to help user by sorting document by similarity. Methodology is sum up on figure 2 At first a set of descriptors is extracted, *e.g.* those of [5]. Many descriptors are extracted because we do not know which ones will be relevant or not. Then the number of classes k of the database is estimated.

The criterion of the mean silhouette described in [6] and [7] is a relevant measure for evaluating clustering quality. Each images are represented by a numerical vector x computed from features described in [5]. For a given clustering, 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. PAM (Partitioning Around Medoid) [8] algorithm is used for clustering, the input is the number of the classes k (like k-Means). 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. $F_1 score$ is used in order to check the accuracy of the clustering. It is based on precision p and recall r such as : $F_1 score = 2 \cdot \frac{p*r}{p+r}$. $F_1 score$ shows that the clustering accuracy is quite similar with estimated k and with real k. Furthermore, sometimes an estimated k with a better silhouette provide better $F_1 score$ than the real k.

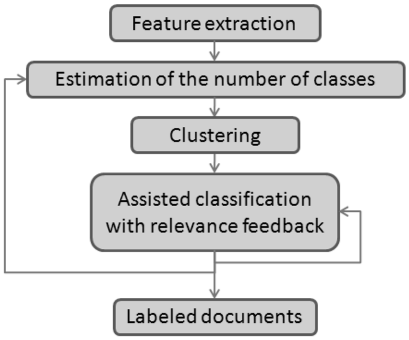
Feature extraction
Estimation of the number of classes
Clustering
Assisted classification with relevance feedback
Labeled documents

Fig. 2. Classification methodology from [4].

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 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 [9] 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 $n_{FS} = 5$. Benefits of feature selection could be observed on figure 3.

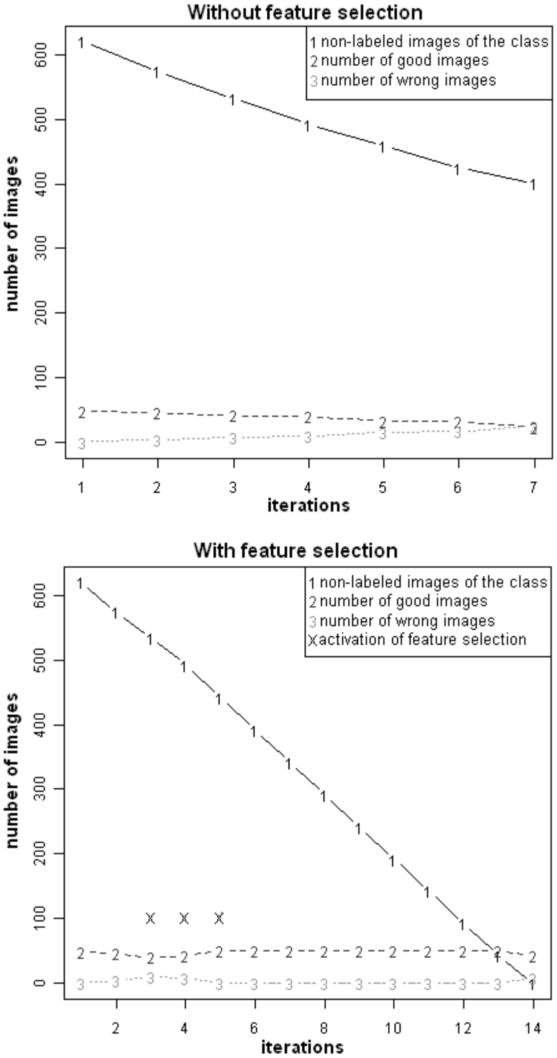

Fig. 3. Impact of feature selection on classification.

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, assisted classification propose more relevant documents and the percentage of labeled documents increase.

First tests using our assisted classification method show that a database is on average labeled 3.4 times faster than with a standard manual classification.

3 Identification and extraction

In this section, a method for matching semi-structured document images is presented (issued from [10]). The aim is to recognize and to localize precisely document image. One of the difficulties is that documents could have geometric variations like translation in plane, rotation perpendicular to the plan, uniform zoom, but also missing information owing to noise, deformation, crop, stains, etc. One image can contain several documents.

The presented methodology could be divided in 4 major steps. 1) In a first time, a *model* image is created. It correspond to an example of the document image researched. More information about a *model* will be given a little lower in text. 2) Interest points are extracted and described by using SURF [11]. 3) Interest points of the model are matched with those of query images by using FLANN [12], a fast and approximative nearest neighbor algorithm. 4) Finally, the transformation which is a 4-parameters model is estimated with RANSAC [13]. If enough matchings are validated by RANSAC, the document is considered to be present on the image and it is localized precisely.

1) A *model* is created from an existing image. Better result will be obtained if the quality is not too poor in order to found interest points in non-variant zones. Variant zones could be removed from the *model*. It will improve time processing by reducing the number of interest points to find and to match. Furthermore it will minimize wrong matchings and self-similarity.

2) Interest points are detected using SURF. An integral image is used in order to speed up calculations, then interest points are found on high intensity variation zones by using Hessian matrix :

$$H(f(x,y)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

Points are then described with SURF, this step is based on Haar wavelet and provide a 64 dimensions vector describing each point, an orientation and a scale value.

3) Each interest points of the model are then matched to the most similar interest point of the query image. Euclidean distance between the vector of 64 dimension of the model point and the query point is computed. The position information is not used here. The k-NNN (nearest neighbor) research could be very long if it is exhaustive. KD tree improve the research time but KD tree are efficient only in low dimensions. [14] propose to use multiple randomized kd-trees in order to speed up matching. FLANN provide an implementation of this algorithm where multiple trees are built in 5 random dimensions. After matching, 2 filters are applie like in SIFT [15]. First if the 2nd NN is too close to the 1st NN (in 64 dimension space) the matching is considered to be wrong. Secondly, each feature votes for an orientation and a scale using Hough with coarse bins.

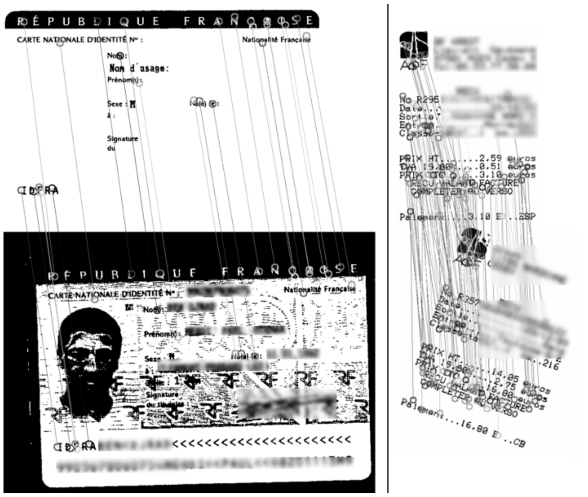

Fig. 4. Matching a model to another document.

4) The aim of this step is to separate the set of matchings between good matching (inliers) and mismatchings (outliers). RANSAC algorithm is used in order to find the following 4 parameters matrix :

$$M_t. \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha.cos(\theta) & -sin(\theta) & T_x \\ sin(\theta) & \alpha.cos(\theta) & T_y \\ 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

The number of trying iterations in RANSAC is fixed to 200. A matching is validated if the transformation of the interest point of the model is at a distance less than 10 pixels from the point matched in the query. In order to not find only very local inliers, a minimum distance of 5 pixels between new matching is needed to validate it. If the numbers of inliers is superior to a fixed threshold t, the document is considered as present, and is extracted. The value have been empirically fixed to $t = 10$.

Since document could be identified and precisely localized, if several documents are present on the same image, multi-detection is done by deleting matched document and process it again. Figure 5 illustrate this principle.

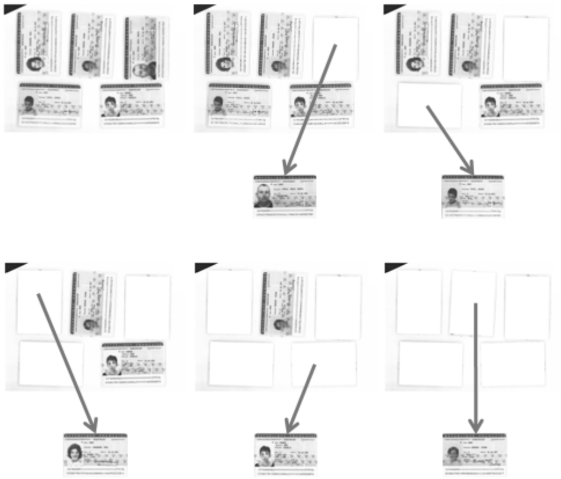

Fig. 5. Multi-detection principle applied to ID cards.

This method have a precision of 100% and a recall of 96% for detecting 483 French ID card on documents. Documents are scanned and printed freely by users, posted and then scanned and binarized by a company at 200 dpi.

4 Indexation based on user defined rules

The next step will be to study user interactions in order to improve identification of documents. Feedback is a good way, but another way is to provide tools to user in order to allow him to specify his query like areas or content of text, tables, pictures, logos, layout, etc.

Acknowledgment

We thank the company Gestform for his involvement in this thesis and Mr. Jean-Marc Nahon, director of computer science studies of the company. Gestform has provided many images from its production and industrial knowledge about digitization.

References

1. 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.
2. D. Keysers, F. Shafait, and T. Breuel, “Document image zone classification - a simple high-performance approach,” in *2nd Int. Conf. on Computer Vision Theory and Applications*, 2007, pp. 44–51.
3. A. Gordo and E. Valveny, “A rotation invariant page layout descriptor for document classification and retrieval,” in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition-Volume 00*. IEEE Computer Society, 2009, pp. 481–485.
4. O. Augereau, N. Journet, and J. P. Domenger, “Document images indexing with relevance feedback: an application to industrial context,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1190–1194.
5. 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.
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. 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.
8. L. Kaufman and P. Rousseeuw, “Finding groups in data: an introduction to cluster analysis,” *NY John Wiley & Sons*, 1990.
9. 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.
10. O. Augereau, N. Journet, and J. P. Domenger, “Reconnaissance et extraction de pièces d’identité,” in *CIFED*, Bordeaux, France, 2012.
11. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
12. M. Muja and D. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *International Conference on Computer Vision Theory and Applications (VISSAPP 09), pages 331*, vol. 340. Citeseer, 2009.
13. M. Fischler and R. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
14. C. Silpa-Anan and R. Hartley, “Optimised kd-trees for fast image descriptor matching,” *Computer Vision and Pattern Recognition*, vol. 0, pp. 1–8, 2008.
15. D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.