TWO METHODS FOR FILLED-IN DOCUMENT IMAGE IDENTIFICATION USING LOCAL FEATURES

Diego Carrion-Robles, Vicent Castello-Fos, Juan-Carlos Perez-Cortes and Joaquim Arlandis*
Institut Tecnològic d'Informàtica, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain

Keywords: Document identification, Filled-in documents, Local features, Reject option.

Abstract: In this work, the task of document image classification is dealt with, particularly in the case of pre-printed forms, where a large part of the document can be filled-in with the result of a potentially very different image. A method for the selection of discriminative local features is presented and tested along with two different classification algorithms. The first one is an incremental version of the method proposed in (Arlandis et al., 2009), based on similarity searching around a set anchor points, and the second one is based on a direct voting scheme ((Arlandis et al., 2011)). Experiments on a document database consisting of real office documents with a very high variability, as well as on the NIST SD6 database, are presented. A confidence measure intended to reject unknown documents (those that have not been indexed in advance as a given document class) is also proposed and tested.

1 INTRODUCTION

A common and practical task where a specific problem of document classification arises is the organization of bills, forms, invoices, legal, medical or administrative documents for processing (e.g. OCR), storing or archiving. Since these documents can be categorized (i.e. "Bill from supplier A", "Tax form number X", etc.), a typical classification method could in principle be applied, but in this case, only a part of the document is kept the same, and the rest changes in every instance. The conserved part can be different from document to document and significantly smaller than the variable area that can be composed of large handwritten, typed or stamped regions.

Traditional approaches of document categorization have addressed the problem as a clustering task, where documents having a certain degree of semantic similarity are assigned to the same class or category. In our case, the task is one of supervised classification, since we need to identify the class of the image among a number of known document classes.

The use of textual data from OCR or the global image structure is also not adequate in our case, since the variable information can significantly alter these features.

Another classical approach relies on the segmentation and the analysis of the layout, but large marks or filled-in areas can introduce changes and errors in that step, so we propose to use only visual features and not the results of structural layout analysis.

A typical, document image consists of white background pixels and black foreground pixels, although other combinations like gray-scale, colour, or complex backgrounds and foregrounds can occur. The foreground is mostly composed of text (in many cases having different appearances like typed fonts, handwriting styles, case letters, bolded text, sizes, etc.), although other objects like images, graphics, logos, or frames are frequent, too. Usually, the text areas also include background patterns interleaved, and some background pattern can also be present in most of the surface of a document.

In summary, a filled-in document can be seen as an image having static (fixed, pre-printed) and variable contents (machine printed, handwritten, marked, stamped, covered with adhesive labels, etc.). Under this definition, a category or class of documents is defined as the set of images having different static content from the other classes and a specific, approximately equal, intra-class static content. The variable content, as has been pointed out, can signifi-

*Work partially supported by the Spanish MICINN grants TIN2009-14205-C04-02 and Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and by IMPIVA and the E.U. by means of the ERDF in the context of the R+D Program for Technological Institutes of IMPIVA network for 2011.

Carrion-Robles D., Castello-Fos V., Perez-Cortes J. and Arlandis J..
TWO METHODS FOR FILLED-IN DOCUMENT IMAGE IDENTIFICATION USING LOCAL FEATURES.
DOI: 10.5220/0003884004810487
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods* (IATMLRP-2012), pages 481-487
ISBN: 978-989-8425-98-0

cantly vary in size and content for different documents within a class. In Figure 1, some filled-in document types are shown.

Given the specific nature of the task, the approaches proposed and compared in this work use local representations to describe the document classes. One of them is based on automatically finding a number of adequate anchor points for each class, and the other uses a direct voting scheme of local-feature vectors using a k-nearest neighbors classifier. A common step previous to both techniques is the selection of candidate points. The experiments carried out test the robustness of the approaches, taking into account that no filled-in contents or representations are used in the training phase.

2 RELATED WORK

The image features proposed in the literature of document analysis and classfication are many and very different. Some are related to the document layout, frame detection, salient visual features, character recognition, texture primitives, shape codes, global image transformations and projections, or semantic block structure detection.

Within works in the domain of Information Retrieval, where the concept of static content against filled-in data is not dealt with, document identification is referred to as a duplicate detection task. In that case, the approaches focus on the correct classification in spite of differences among document instances, like resolution, skew, distortions and image quality. Speed and robustness are key elements, as well as the ability to handle very large databases.

Most works dealing with filled-in documents are related to form identification. Many of them are based on analyzing global and local structures. Structural features are usually limited to documents having frames, cells, lines, blocks, or similar items, and they may fail when different documents have very similar structures. Other works rely on using character and string codes to achieve the document identification (Sako et al., 2003), as well as, on computing pixel densities from image regions (Heroux et al., 1998). Within form-type documents, specific applications are addresed to coupons (Nagasaki et al., 2006), banking (Ogata et al., 2003) or business (Ting and Leung, 1996) form identification.

More recent and closely related works, are the ones presented by Parker (Parker, 2010) and Sarkar (Sarkar, 2006), (Sarkar, 2010). Sarkar (Sarkar, 2006) presents a methodology to select and classify anchor points from document images. The anchor

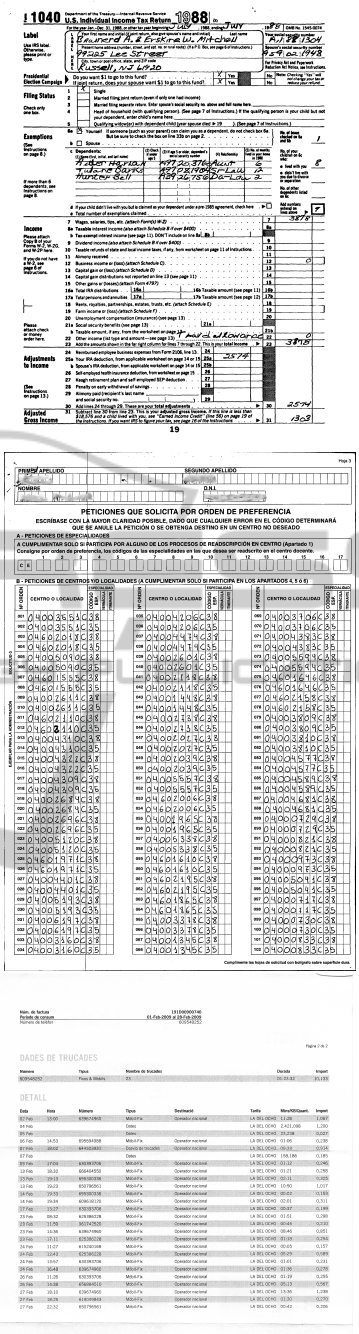

Figure 1: Document examples. The first one is a form where the static contents encompass most of the document. The second is a form page with a large number of cells that can be filled-in or not. The third is a business document with few structural patterns and static contents (located in the header), while the variable part can cover the rest of the image.

points selection is based on the use of thresholded Viola&Jones rectangular salient visual features in the luminance channel (Viola and Jones, 2001). For each document class, a probability distribution of the list

of local features (including global location coordinates) is obtained by a latent conditional independence (LCI) model. An image is classified by matching its resulting feature list to category-specific generative models by means of a maximum likelihood criterion, and it is assigned to the category whose distribution is closest, in the Kullback-Liebler sense, to the empirical distribution. This correspondence is well known in the text categorization/retrieval community where observations are variable-length lists of words. Recently, Sarkar (Sarkar, 2010) proposed a complete methodology to select anchor points based on randomly picked sub-images and aplying succesive refinements by expanding and ranking the candidates using two alternative quality measures.

Parker (Parker, 2010) proposes and compares three methods for selecting anchor points. The first is based on two criteria: "graphical action" and intra-class distance minimization. The second and third methods try to select the anchor points that maximize the KL-divergence (Kullback-Leibler divergence) function, a measure of the separation of two distributions: one of the distances among anchor points within a sample of a given document class, and the other one of the distances from those anchor points to documents of different classes. Parker claims that the performance of the proposed form identification system can be estimated in a theoretical way by using the KL-divergence. He shows the results of experiments of the three methods using a customized database of forms extracted from the IRS (Internal Revenue Service, the revenue service of the United States federal government), where only one document type having filled-in data was used and ten completed forms were used to train the system. The main conclusion of the experiments is that the use of inter-class information to select the anchor points of a class improves the performance of the system (estimated by means of the KL-divergence). This method implies the use of several documents of each class to train the system, and a high number of correlation operations can be required to select anchor points to be robust against image translations, as needed in the operating phase.

3 APPROACHES

A method to select a set of potentially discriminant reference points to be used in a document identification task is presented in section 3.1. This method has been used to extract features and classify documents by means of two approaches: a new incremental version of the method proposed in (Arlandis et al., 2009), based on the cross matching between pairs of documents (section 3.2), and the method proposed in (Arlandis et al., 2011), which relies on the combination of the evidence contributed by multiple local features and a direct voting scheme (section 3.3).

3.1 Reference Point Selection

The goal of this phase is to obtain an ordered list of small sub-images of a fixed size from the reference image of each class. These sub-images should be representative of that image. Thus, a selection criterion is necessary to ensure that these sub-images are located in the most informative regions of the reference image in order to retain the areas with clear graphical content such as text or any other potentially discriminative pattern, avoiding uniform areas or uninformative background regions. This decision can be made on the basis of image contrast, or variance, or on more complex operators, like textures, corner detection or specific filters.

In order to avoid uniform areas, a good approximation can be obtained by sorting by variance all the possible sub-images of the desired size, possibly using subsampling to reduce the computational cost. The problem of using this method alone is that some uninformative regions, like borders between very dark and very light areas of the document, usually have a high variance, ending up in the top positions of that list. This kind of patterns are undesirable to be used as discriminative local features because many different documents are bound to have them, for instance, borders of tables, scanning artifacts like shadows or edges, etc.

To avoid this, a more sophisticated second pass can be carried out, using specific features, like Haralick descriptors (Haralick et al., 1973). Particularly, establishing a limit in the autocorrelation value and the entropy difference has been found to be helpful in eliminating the sub-images that have high variance but a low discriminative potential. In Figure 2, some examples of selection criteria are shown.

3.2 Approach 1: Cross Matching of Document Pairs

This is an incremental version of the method presented in (Arlandis et al., 2009). It reduces drastically the time needed to train a high number of document classes of that method, particularly when the documents to be indexed are not very similar. This is because, on one hand, candidate images are pre-selected (as explained in the former section), which increases the likelihood of finding a discriminant feature faster.

Figure 2: Examples of selection of candidate sub-images. The first figure is a detail from the original reference image for a class. The second represents the 80x30 sub-images that have higher variance in that portion of the reference image. Note that most of them are in borders between dark and light sectors. The third image represents the 80x30 sub-images that had higher variance *and also* passed certain entropy difference criteria.

On the other hand, in the presented method, a local feature is selected based on its discriminative power between pairs of images, instead of requiring to find it discriminitive among the rest of the classes.

Similarly to (Arlandis et al., 2009), this approach relies on the fact that the discriminant power of a sub-image of a fixed size $I^c_{x,y}$, on the reference image of class c with respect to the reference image of another class c' can be expressed as:

$$r^{cc'}_{x,y} = \min_{\substack{-w_x \le i \le w_x \\ -w_y \le j \le w_y}} d(I^c_{x,y}, I^{c'}_{x+i,y+j})$$

where d is a distance function used to measure dissimilarity beteween two sub-images and w is the half size of the search window, i.e, the matching area around (x,y) needed to compensate for image distortions and translations.

Then, $r^{cc'}_{x,y}$ represents the distance from $I^c_{x,y}$ to the most similar sub-image found in an image from the class c' around (x,y). Therefore, r is a good estimator of the minimum distance that is expected to be found when comparing $I^c_{x,y}$ to an image belonging to class c'. That suggests that higher values of $r^{cc'}_{x,y}$ give rise to a higher discriminant power of $I^c_{x,y}$ with respect to c'.

The set of final δ-landmarks (or anchor points) of a class c, L^c, can be defined as the set of the sub-images $I^c_{x,y}$ that have a significant dissimilarity with respect to *each and every one* of the other known classes,

$$L^c = \{I^c_{x,y} | r^{cc'}_{x,y} > \mathcal{T}_r\}, \qquad c \neq c'$$

where the threshold $\mathcal{T}_r$ can be empirically set. The cardinality of this set will depend on two factors:

- A sub-image of a class may be discriminant enough ($r^{cc'}_{x,y} > \mathcal{T}_r$) with respect to more than a class, and δ-landmarks from differents pairs of classes can be shared.
- It is possible to enforce a minimum number of δ-landmarks for each pair of classes by appropriately tuning $\mathcal{T}_r$.

Training

To find the sub-images that discriminate between two given classes, the candidate features of the new class found in the δ-landmark selection phase are tested for minimum normalized distance in a search window (relative to the coordinates of the feature) of the other class. If the distance found is less than a predetermined threshold $\mathcal{T}_r$, that feature is annotated as discriminating between the two classes. Afterwards, the roles of the classes are reversed and the process is repeated (notice that $r^{cc'} \neq r^{c'c}$). Finally, the selected features are consolidated by eliminating the repeated ones, which happens when a candidate feature has been found to be discriminative among more than a pair of classes.

Test

Testing in this case is straightforward: each local feature from each training class is compared against the test document in a search window around the feature point coordinates to find the minimum distance. From this, an average distance for each class can be computed. Finally, the test document is assigned to the class that gets the minimum average distance.

3.3 Approach 2: Direct Voting Scheme

The second approach tested was proposed in (Arlandis et al., 2011). In this case, the identification of a test document relies on the combination of the evidence contributed by a high number of local features (sub-images).

In the training phase, a high number of sub-images from each class are selected from its reference images, and a local feature vector from each sub-image is obtained expanding the gray values of pixels in a row vector. The sub-image selection criterion should retain areas with potentially discriminative content, as

explained in section 3.1. In the test phase, a high number of sub-images are also selected from an test image and used to classify it. In this case, a higher number of sub-images is required since the reference images are filled-in documents, and they have more potentially discriminative content.

The discriminative power of the local features extracted from the selected sub-images is improved by taking into account non-local, or *global*, geometric information. Thus, the coordinates of each sub-image are added as two new components to the feature vectors, and properly normalized before classification. To tune the effect of these global features with respect to the rest of the components, two weighting factors (α_x, α_y) empirically estimated are applied.

Each vector is classified according to the k-nearest neighbors rule, and finally, the class with the largest number of votes is obtained. More formally, the classification procedure used is related to the methods often referred to as *direct voting schemes* (Mohr et al., 1997). Given a prototype set representing the reference classes, and a set of feature vectors $m_Y = \{y_1, \ldots, y_m\}$ extracted from a test image Y, the classifier can be written as a linear combination of m_Y classifiers, each one from every feature vector of Y (Kittler et al., 1998). The so called *sum rule*, often used in practical applications, can be used to optimally classify an image Y in a class $\hat{w}$:

$$\hat{w} = \arg\max_{1 \le j \le d} \sum_{i=1}^{m_Y} P(\omega_j | y_i),$$

Assuming that the number of vectors of each class in the prototype set is fixed according to the *a priori* probabilities of the classes, the following classification rule can be used:

$$\hat{w} = \arg\max_{1 \le j \le d} \sum_{i=1}^{m_y} k_{ij}$$

where k_{ij} is the number of neighbors of y_i belonging to the class ω_j provided by the k-nearest neighbors rule. That is, a class $\hat{w}$ with the largest number of votes accumulated over all vectors extracted from the test image is selected.

Note that the selection criterion, as proposed in section 3.3, is applied within a class. Therefore, it is not guaranteed that similar sub-images from different classes can be found at similar locations, although it is expected that the probability of matching an "extraneous" vector (*casual matchings*) will be distributed among the different classes when using a large number of vectors.

4 EXPERIMENTS

Two different sources of documents have been used to test the accuracy of the proposed methods. The first is the SD6 NIST database (Dimmick and Garris, 1992) consisting of 5590 binary filled-in forms, 300 dpi, US Letter size, from 20 different classes. The second one (IDF1) is a document collection obtained from an actual office setting. It consists of 683 binary and gray-scale documents from 47 classes including invoices, bank documents, personal documents, and a variety of forms of different sizes and aspect ratios, but mostly 300 dpi DIN A4 size, portrait orientation, and various amounts of filled-in contents.

From these databases, two data sets have been built:

- SD6. Used as a baseline for comparison against other approaches (see (Sarkar, 2006), (Sarkar, 2010)). Experiments with different number of reference images were carried out.
- IDF1+SD6. Composed by the union of both databases, it is used to obtain results using the maximum number of classes available (67 classes). One reference image per class was selected for the training set, while 6 to 10 images per class were used for testing. The reference images have been checked and, if necessary, manually cleaned to remove filled-in contents.

Several preprocess techniques have been tested with all image sets. Automatic orientation normalization was applied to the documents and each image was size normalized (to an equivalent of an A4 page at 300 dpi area) preserving its original aspect ratio. Finally, to reduce the processing time, the images were scaled by a factor of 0.25.

The list of candidate reference points was obtained and ordered as explained in section 3.1. To select candidate sub-images, several textural features were computed on local windows of 80×30 pixels. Some tests were performed to measure the capability of several textural features to select the most discriminative sub-images, and it was found that variance, autocorrelation coeficient and entropy difference worked better than the rest. The entropy difference was finally used in the experiments.

In the particular case of the Cross Matching classifier, to account for potential translations of the test image relative to the reference images, a search area of 200×200 pixels around the landmark coordinates was used. The threshold $\mathcal{T}_r$ was empirically set, and the first two landmarks of the candidate list having a correlation index over $\mathcal{T}_r$ were selected for each pair of classes, which led to a total average number of 29.9

(SD6) and 52.2 (IDF1+SD6) landmarks per class. Using this setting, all documents from both SD6 and IDF1+SD6 test sets were correctly classified.

In the case of the Direct Voting Scheme classifier, a sub-sampling was applied in order to ensure that any selected sub-image having static contents from a test image was included in the training set. After an empirical initial evaluation, a fixed number of 300 reference points were selected from each training image and 400 points from each test image. A PCA dimensionality reduction was applied to the local features selected, resulting on 15-dimensional vectors. Four nearest neighbours were considered in the sum classification rule described in section 3.3. A *kd*-tree data structure, provided fast approximate *k*-nearest neighbor search. All documents from the combined IDF1+SD6 test set were correctly classified using two reference images.

4.1 Reject Option

For a given test set, the distribution of the reliability indices of well classified documents should not overlap with the distribution of the reliabilities of misclassified and non-indexed documents (*unknown documents*). Obviously, the more separated both distributions are, the better generalization is to be expected.

Thus, 205 randomly selected document images, mostly forms, not belonging to any of the reference images, was collected and used as a test set of unknown documents. The reliablity of a class for a given document was computed as follows:

- Cross Matching classifier (CM): The mean of the correlation index obtained for all the landmarks of the class.
- Direct Voting Scheme (DVS): The class posterior probability provided by the sum rule classifier.

Table 1 shows the results obtained on the SD6 and IDF1+SD6 databases, along with the unknown document set. Using the reliability indices defined, the recall at 100% precision, and the KL-divergence obtained are shown, as well as the error rate (no unknown documents considered). The KL-divergence was computed using the abovementioned reliability distributions. Because of the non-symetric quality of the KL-divergence, the minimum value of the two disssimilarity functions between both reliability distributions is shown.

On one hand, the results show that the combination of the reference point selection method used, along with the two classifiers described, provided a 100% recognition rate on the two sets tested. On the other hand, the recall, precision and KL-divergence values obtained suggest that the reliability measure provided by both classifiers is able to correctly rank the known and unknown documents, and therefore, allows the rejection of the unknown ones without significantly affect the rejection of indexed documents.

Table 1: Error rate, Recall at 100% precision, and KL-divergence measured on the SD6 and IDF1+SD6 databases for the best parameter sets.

	Error Rate		Recall 100%		KL-diverg	
	CM	DVS	CM	DVS	CM	DVS
SD6	0	0	100	99.8	40.8	36.7
Both	0	0	99.9	99.9	38.0	37.3

The processing speed measured on an AMD 64-bits 4 CPU 3 GHz machine for the DVS method was 1.6 doc/s in both data sets. In the case of CM, the speed was 0.47 doc/s for the IDF1+SD6 database and 1.02 doc/s for the SD6 database.

5 CONCLUSIONS

Two approaches to deal with the task of classifying documents with total flexibility of designs, layouts, sizes, and amount of filled-in contents in an efficient way have been tested. A common method for selecting the best reference points in the document images has been used to improve the results.

Experiments on document identification were carried out, and all the documents from both SD6 and the combined database were correctly classified, and good performances on the rejection rates of non-indexed document images were also achieved. Training and test computation times were within the demands of a real workflow in document processing.

REFERENCES

Arlandis, J., Castello-Fos, V., and Pérez-Cortes, J. C. (2011). Filled-in document identification using local features and a direct voting scheme. In Vitrià, J., Sanches, J. M. R., and Hernández, M., editors, *IbPRIA*, volume 6669 of *Lecture Notes in Computer Science*, pages 548–555. Springer.

Arlandis, J., Perez-Cortes, J.-C., and Ungria, E. (2009). Identification of very similar filled-in forms with a reject option. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 246 –250.

Dimmick, D. L. and Garris, M. D. (1992). Structured forms database 2, nist special database 6.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610 –621.

Heroux, P., Diana, S., Ribert, A., and Trupin, E. (1998). Classification method study for automatic form class identification. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 926 –928 vol.1.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226 –239.

Mohr, R., Picard, S., and Schmid, C. (1997). Bayesian decision versus voting for image retrieval. In *IN PROC. OF THE CAIP-97*, pages 376–383.

Nagasaki, T., Marukawa, K., Kagehiro, T., and Sako, H. (2006). A coupon classification method based on adaptive image vector matching. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 280 –283.

Ogata, H., Watanabe, S., Imaizumi, A., Yasue, T., Furukawa, N., Sako, H., and Fujisawa, H. (2003). Form-type identification for banking applications and its implementation issues. In *DRR'03*, pages 208–218.

Parker, C. (2010). Anchor point selection by kl-divergence. In *Image Processing Workshop (WNYIPW), 2010 Western New York*, pages 42 –45.

Sako, H., Seki, M., Furukawa, N., Ikeda, H., and Imaizumi, A. (2003). Form reading based on form-type identification and form-data recognition. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ICDAR '03, pages 926–, Washington, DC, USA. IEEE Computer Society.

Sarkar, P. (2006). Image classification: Classifying distributions of visual features. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 02*, ICPR '06, pages 472–475, Washington, DC, USA. IEEE Computer Society.

Sarkar, P. (2010). Learning image anchor templates for document classification and data extraction. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3428 –3431.

Ting, A. and Leung, M. (1996). Business form classification using strings. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 2, pages 690 –694 vol.2.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:511–518.