Universitat Politècnica de València

Departament de Sistemes Informátics i Computació

Trabajo Final de Máster

8 de septiembre de 2011

Filled-in document image identification using landmarks

Presentado por: Diego Carrión Robles
Dirigido por: Joaquim Francesc Arlandis Navarro / Juan Carlos Pérez Cortés

Contents

Chapter 1

Introduction

Nowadays, a small business can receive or generate several hundred documents per day, all of which must be organized and stored. With a high-speed scanner, the documents can, in a small amount of time, be converted to electronic form. In some cases, each of the documents must still be categorized and moved to the proper location. So we have a classification problem. Namely, about 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 or identification 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 this 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 this 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 this proposal is 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 significantly vary in size and content for different documents within a class. In Figure 1.1, some filled-in document types are shown.

Given the specific nature of the task, the approach proposed in this work use local representations to

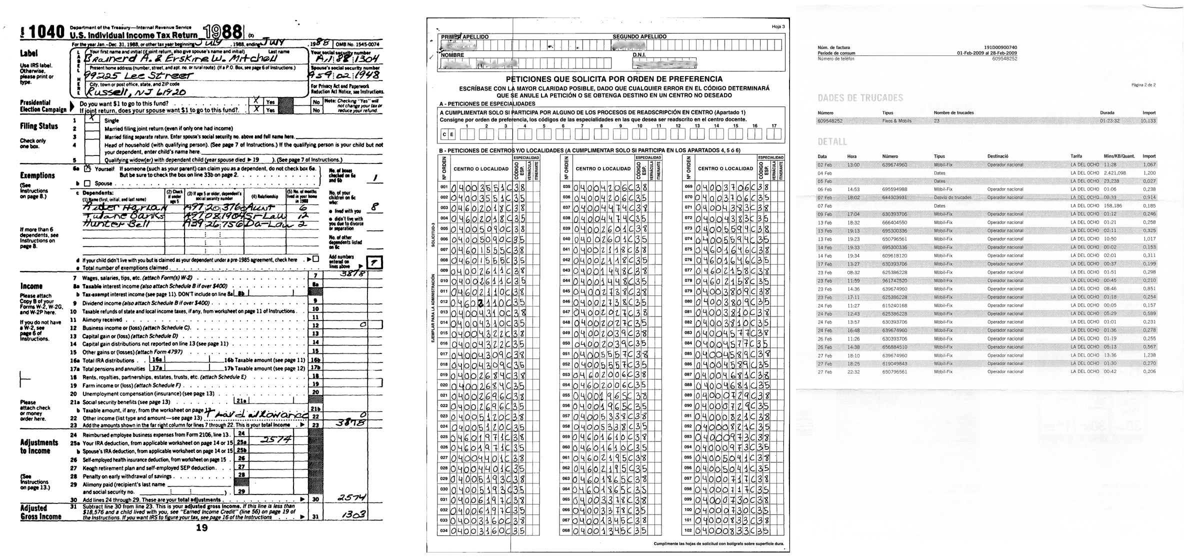


Figure 1.1: Document examples. The first one is a form where the static contents encompass most of the document. The second is a form page with a large number of cells that can be filled-in or not. The third is a business document with few structural patterns and static contents (located in the header), while the variable part can cover the rest of the image.

describe the document classes. This approach is based on automatically finding a number of adequate anchor points for each class, after a previous selection of the candidate points. The experiments carried out test the robustness of the approach, taking into account that no filled-in contents or representations are used in the training phase.

Chapter 2

State of the art in document identification

In the classification of documents, traditionally, a significant amount of effort has been dedicated to develop approaches based on group documents with a certain degree of semantic similarity as belonging to the same class or category. However, in some applications, as those related to digitization and data mining, among others, the classes must be defined as representing particular types of documents. In this case, the task is commonly known as "document identification" and the grouping or clustering methods are not suitable. In most of these applications, the identification of the image of a document is required first, before any specific process.

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.

In the scope of Information Retrieval, when no filled-in contents exist, document identification can be seen as a duplicate detection task [Doermann 98]. In this case, the approaches have to tackle with differences among document instances, like resolution, skew, distortions and image quality, speed and robustness, as well as, handling 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 [Fan 01], [Ohtera 04], [Mandal 05]. Structural features are usually limited to documents having frames, cells, lines, blocks, or similar items, and it may not help with different types of documents having similar structure. Other works rely on using character and string codes to achieve the document identification [Sako 03], as well as, on computing pixel densities from image regions [Heroux 98]. Within form-type documents, specific applications are addresed to coupons [Nagasaki 06], banking [Ogata 03] or business [Ting 96] form identification.

More recent and closely related works, are the ones presented by Parker [Parker 10] and Sarkar [Sarkar 06], [Sarkar 10]. Sarkar [Sarkar 06] presents a methodology to select and classify anchor points from document images. The anchor points selection is based on the use of thresholded Viola&Jones rectangular salient visual features in the luminance channel. 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 10] 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 10] 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 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 IRS, 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.

In fact, no public suitable scientific databases, in terms of high number of document types having a certain layout variability and filled-in contents, have been found.

Chapter 3

Overview of the system

3.1 Introduction

This work is, at its essence, an incremental version of the method presented in [Arlandis 09], adapted to be suitable for a high number of document classes. It drastically reduces the computational complexity of that method, particularly when the documents to be indexed are not very similar. In this case, local features are selected based on their discriminative power between pairs of images.

Given a set of C reference images, representing C different document classes, we define a discriminant landmark area, or δ-landmark of an image, as a sub-image which can be used to describe its own class, discriminating it from all other classes.

δ-landmarks can be seen as salient visual features related to a location on the image. The δ-landmarks of a class should be different, at least to a certain degree, from sub-images found at approximately the same location in documents of any other class, and, at the same time, they should always be found at their respective locations in images belonging to its own class.

δ-landmarks can be automatically learned from a set of class reference images. For each class, its δ-landmarks can be computed by scanning the reference image, searching for sub-images having features with high discriminant power with respect to sub-images found in the same region in the rest of the classes. The dissimilarities can be computed by using a correlation or a dissimilarity function among local feature vectors of the images. After this training, a number of δ-landmarks can be taken as the basis of each document class model.

Subsequently, δ-landmarks can be used to classify, or reject, new images under a certain criterion, by matching each δ-landmark against sub-images extracted from a corresponding window on the test image at the δ-landmark location. No restrictions are placed on the layout and contents of the input images, since either text or non-text image regions can be automatically selected as discriminant landmark areas.

3.2 δ-landmark matching

In [Arlandis 09], the approach relied 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 images of the rest of the classes could be expressed be expressed as:

$$r^c_{x,y} = \min_{\substack{c' \neq c \\ -w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^c_{x,y}, I^{c'}_{x+i,y+j}) \tag{3.1}$$

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. $r^{cc'}_{x,y}$ represents the distance from $I^c_{x,y}$ to the most similar sub-image found in any other reference image around (x, y).

So that approach aimed to find, for each class, a single sub-image which discriminated that class from *each and every one* of the other classes. This works well with a limited number of classes even when they have very similar reference images, because the computational cost is quadratic with the number of classes.

Furthermore, that method does not lend to a naturally scalating system because for every new class that is added, the entire system needs to be retrained. That is natural, because it is very plausible that, for a given class, the feature that discriminated it from the previous classes now is not discriminative against the new class and thus it must be discarded and a new one needs to be found.

In order to keep previously computed features, a different strategy would be needed. If, instead of aiming to find a single sub-images that discriminates between a class and every other, the objective was to find features that discriminate between classes in pairs, equation 3.1 could read like:

$$r^{cc'}_{x,y} = \min_{\substack{-w_x \leq i \leq w_x \\ -w_y \leq j \leq w_y}} d(I^c_{x,y}, I^{c'}_{x+i,y+j}) \tag{3.2}$$

Here, $r^{cc'}_{x,y}$ 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 any of the other known classes,

$$L^c = \{I^c_{x,y} \mid r^{cc'}_{x,y} > \mathcal{T}_r\}, \qquad c \neq c'\} \tag{3.3}$$

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.

The second factor gives an absolute minimum for the number of final δ-landmarks that a class will have; whereas the first factor implies that if good candidate landmarks have been extracted as explained in section 3.4, the number of final δ-landmarks for a given class is bound to remain low.

3.3 Feature selection method

In this method δ-landmarks can be used directly as features. A way to accelerate the process is to preselect the features that, owing to a certain measure of quality, are more likely to have good intrinsic discriminative power. This way, instead of an intensive search for pairs of features that rise above $\mathcal{T}_r$, a smaller subset of them would be tested. The techniques for this are detailed in section 3.4.

3.3.1 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 maximum normalized correlation in a search window (relative to the coordinates of the feature) of the other class. If the correlation found is less than a predetermined threshold $\mathcal{T}_r$, that feature is annotated as discriminating between the two classes. This process can be repeated a number of times to ensure a minimum number of discriminating features between the two classes. Afterwards, the roles of the classes is reversed and the process is repeated (notice that $r^{cc'} \neq r^{c'c}$).

The whole process is then applied to each pair of classes, when enough discriminating features are found between the new class and each and every one of the other preexisting classes, the features found are consolidated by eliminating the repeated ones. This happens when the same candidate feature of a class has been found to be discriminating between it and two or more other classes. In the end, the class is represented by the unique features that remain; which have been found to be enough to discriminate between it and the other classes.

3.3.2 Test

Testing is straightforward: each local feature from each training class is correlated against the test document in a search window around its point coordinates to find the best correlation. From this, an average correlation for each class can be computed. Finally, the test document is assigned to the class that obtains the best average correlation.

Furthermore, a rejection threshold can be determined experimentally by comparing the results obtained by test documents of known classes with the results obtained by unknown documents. This is elaborated upon in section 3.5.1.

3.4 Preprocessing and feature extraction

In this section, the image preprocessing applied to either training and testing, as well as the process of selection of candidate features in training are described. The use of preprocessing techniques is justified in terms of improving the system's performance, both in speed and error rate.

3.4.1 Image preprocessing

The corpus' documents present several problems which must be solved before making any experiment. As seen in section 4.1, the document images can be rotated due to mechanical defects when scanning the documents. Also, different documents from the same class can have brightness and contrast differences due to problems in the scanning hardware.

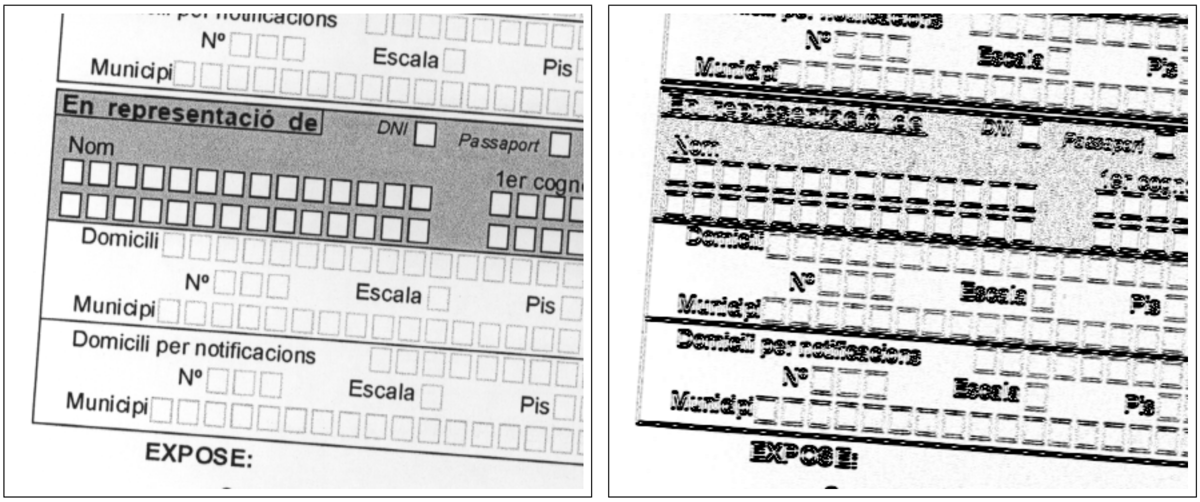


Figure 3.1: Image before and after applying Sobel.

Rotation correction

The method used for correcting the rotation follows the procedure described in [Andreu-Cerezo 10].

The problem when correcting possible rotation defects is that the rotation angle is unknown. This defect is caused by mechanical failures when sliding the document along the scanner feeder. Sometimes, the document is perfectly scanned, sometimes it suffers random deviation in both sides and with different magnitudes. Thus the problem is reduced to detecting the document's rotation angle.

In order to calculate the angle, first the horizontal Sobel operator for border detection is applied. This, applied to a grayscale image, has the effect of highlighting the horizontal (or close to horizontal) lines. In figure 3.1 the results of applying Sobel in an original image can be appreciated.

Then the Hough Transform was used to estimate the angle of rotation of the majority of the horizontal lines in the image. To reduce the computational cost, only lines at +5/-5 degrees of rotation are searched, and filtering is used to reduce small horizontal line segments (less than 10 pixels) to a single pixel. The results can be appreciated in figure 3.2.

Finally, after obtaining the rotation angle, the image is straightened using Paeth rotation (figure 3.3).

3.4.2 Candidate feature extraction and ordering

The goal of this phase is to obtain an ordered list of δ-landmarks 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. This section explores ways of ordering this sub-images, by using different texture values for the ordering of the features.

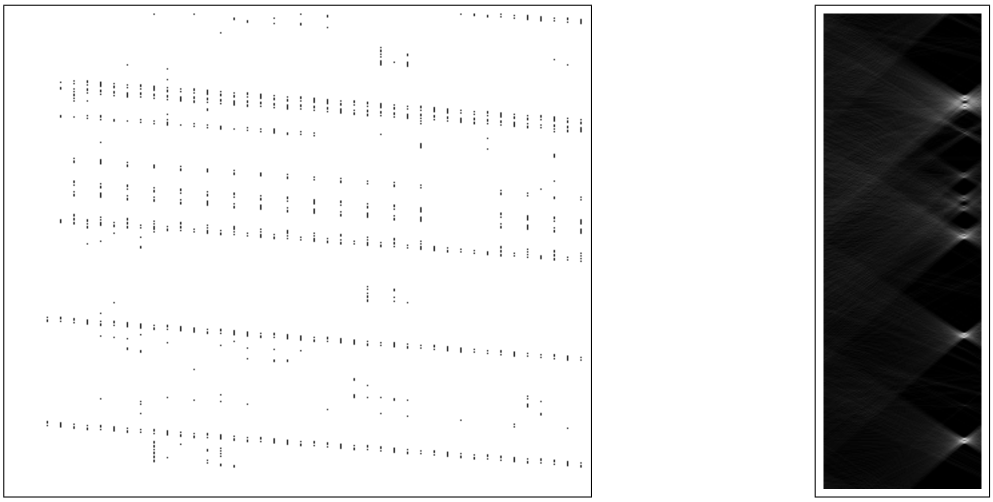
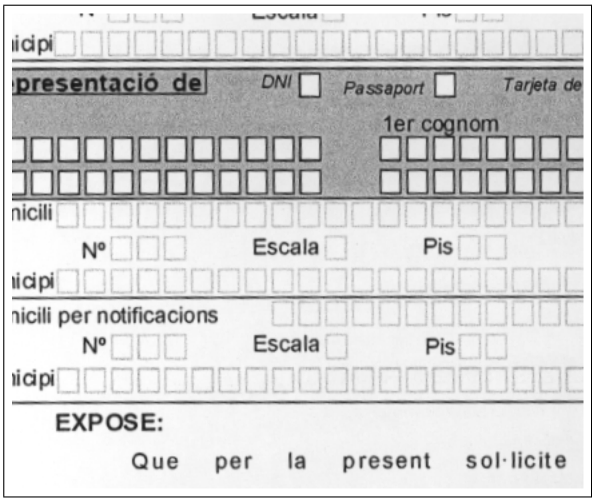


Figure 3.2: Filtered image and resulting Hough map.



icipi

presentació de DNI □ Passaport □ Tarjeta de

1er cognom

nicili

Nº □□□ Escala □ Pis □□

icipi

nicili per notificacions

Nº □□□ Escala □ Pis □□

icipi

EXPOSE:

Que per la present sol·licite

Figure 3.3: Resulting image

Variance filter

A variance filter seems an obvious choice for extracting representative sub-images from a reference image. Uniform areas tend to have lower variance than non-uniform areas. As uniform areas tend to exist in all kinds of documents (white or black areas in borders, white space between paragraphs...) it can be assumed that non-uniform areas should contain more informative sub-images.

Considering a sub-image as a vector of pixel values with of n dimensions $\vec{x} = (x_1, ..., x_n)$, the variance σ^2 comes determined by the formula:

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

where μ is the arithmetic mean of the components of the vector $\vec{x}$

$$\mu = \sum_{i=1}^{n} \frac{x_i}{n}$$

The variance function can be simplified to look like this:

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i)^2}{n} - \mu^2 \tag{3.4}$$

Integral images

Using the concept of integral images popularized in [Crow 84] and [Viola 04] it is possible, as shown in [Shafait 08], to compute the local variance much more efficiently. Specifically, if we define the integral image I for each pixel as the sum of all the pixels above and to the left of that pixel in the original image i:

$$I(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x', y')$$

then I can be computed efficiently with a single pass through the original image taking advantage of the fact:

$$I(x, y) = i(x, y) + I(x - 1, y) + I(x, y - 1) - I(x - 1, y - 1)$$

With a precomputed integral image, the mean $\mu(x, y)$ for a window w pixels wide and h pixels high with center in (x, y) can be computed by only two addition and two subtraction operations:

$$\begin{aligned} \mu(x, y) = (&I(x + w/2, y + h/2) + I(x - w/2, y - h/2) - \\ &I(x + w/2, y - h/2) + I(x - w/2, y + h/2))/(w * h) \end{aligned} \tag{3.5}$$

And to compute the variance in that window one could compute:

$$\sigma^2(x, y) = \frac{1}{w * h} \sum_{j=x-w/2}^{x+w/2} \sum_{i=y-h/2}^{y+h/2} g^2(x, y) - \mu^2(x, y)$$

where, in turn, the term $g^2(x, y)$ can also be computed as in equation 3.5 by using an integral image of the squared pixel intensities. This has an implementation issue: as the values of the square integral image can get very large, overflow problems can occur if 32-bit integers are used in this latter case.

Texture filters

There are some documents for which the techniques explained in section 3.4.2 could not be sufficient. Images which contain high contrast frames (for instance, a white square over a black background), or dark borders over a white background will have as top candidate features sub-images with roughly half the pixels white (or light), half black (or at least dark). This is very common in tables and charts, and also a common effect when scanning with parts of the scanner sensor exposed to an exterior light.

An example of this is shown in figure 3.4: the left image shows a form with many frames and tables filled in black over white. The center image shows that the candidate δ-landmarks ordered by variance (the image shows the first 200 of them) would cluster precisely on those black-white borders.

This kind of features are very uninformative, because similar ones can appear in any document, but they have maximum variance and thus are deemed the best candidates when using variance as the selection criterion.

This happens because any sub-image with half the pixels black and half white will have the maximum possible variance. To demonstrate it, let's suppose, for the sake of simplicity, that in a sub-image with N number of pixels, black pixels have a value of 0 and white pixels have a value of 1, and pixels can only be either black or white. If nb is the number of black pixels in the image, a function of the variance of the sub-image depending on the number of black pixels will have the form (following equation 3.4):

$$\sigma^2(nb) = \sum_{i=1}^{N} \frac{(x_i)^2}{N} - \mu^2 = \frac{nb \cdot 0^2 + (N - nb) \cdot 1^2}{N} - \mu^2 = \frac{N - nb}{N} - (\frac{N - nb}{n})^2$$

Differentiating to obtain the maximum:

$$\begin{aligned} \frac{\mathrm{d}\sigma^2(nb)}{\mathrm{d}nb} = -\frac{1}{n} - \frac{2nb}{n^2} + \frac{2}{n} = \frac{1}{n} - \frac{2nb}{n^2} &= 0 \\ nb &= \frac{n}{2} \end{aligned}$$

These findings suggest that some better criteria for selecting features could be used. Criteria that value not only the contrast of the sub-image as a whole, but also the spatial distribution of pixel values over the image, in order to find sub-windows that are more representative.

In [Haralick 73] a set of 14 texture measures is proposed, based on gray-level co-occurrence matrices (GLCM). They form the basis from which to obtain a set of measures that capture the spatial relationship within the gray tones of an image.

Basically, given an image I, its GLCM P_d for a displacement vector $d = (d_x, d_y)$ is a matrix sized $N_G \times N_G$ (being N_G the number of different gray tones in the image) where the entry (i, j) of P_d is the number of occurrences of the pair of gray levels i and j which are a distance d apart. More formally:

$$P_d(i,j) = \mid ((r,s),(t,v)) : I(r,s) = i, I(t,v) = j \mid$$

where $(r, s), (t, v) \in N \times N, (t, v) = (r + dx, s + dy)$, and $\mid . \mid$ is the cardinality of a set.

[Haralick 73] only considers for the proposed measures horizontal, vertical or diagonal displacement vectors, that is, vectors d where $d_x = d_y$, $d_x = 0$ or $d_y = 0$. So those matrices can be defined by an angle (0°, 45°, 90°or 135°) and a distance (which can be negative).

Starting from those definitions, [Haralick 73] proposed 14 measures:

Notation:

$$P_y(j) = \sum_i P(i,j)$$

$$P_{x+y}(k) = \sum_i \sum_{\substack{j \\ i+j=k}} P(i,j), \quad k = 2, 3, \ldots, 2N_G$$

$$P_{x-y}(k) = \sum_i \sum_{\substack{j \\ |i-j|=k}} P(i,j), \quad k = 0, 1, \ldots, N_G - 1$$

1. Angular Second Moment

$$f_1 = \sum_i \sum_j P^2(i,j)$$

2. Contrast

$$f_2 = \sum_{n=0}^{N_G-1} n^2 \{ \sum_i \sum_{\substack{j \\ |i-j|=n}} P(i,j) \}$$

3. Correlation

$$f_3 = \frac{\sum_i \sum_j (i-\mu_x)(j-\mu_y) P(i,j)}{\sigma_x \sigma_y}$$

4. Variance

$$f_4 = \sum_i \sum_j (i-\mu)^2 P(i,j)$$

5. Inverse Difference Moment

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} P(i,j)$$

6. Sum Average

$$f_6 = \sum_{i=2}^{2N_G} i P_{x+y}(i)$$

7. Sum Variance

$$f_7 = \sum_{i=2}^{2N_G} (i - f_8)^2 P_{x+y}(i)$$

8. Sum Entropy

$$f_8 = -\sum_{i=2}^{2N_G} P_{x+y}(i) \log(P_{x+y}(i))$$

9. Entropy

$$f_9 = -\sum_i \sum_j P(i,j)\log(P(i,j))$$

10. Difference Variance

$$f_{10} = \frac{(N_G^2 \sum_i P_{x-y}^2) - (\sum_i P_{x-y})^2}{N_G^4}$$

11. Difference Entropy

$$f_{11} = -\sum_{i=0}^{N_G-1} P_{x-y}(i)\log(P_{x-y}(i))$$

12. Information Measure of Correlation #1

$$H_{XY} = -\sum_i \sum_j P(i,j)\log(P(i,j)) = f_9$$

$$H'_{XY} = -\sum_i \sum_j P(i,j)\log(P(i)P(j))$$

$$H''_{XY} = -\sum_i \sum_j P(i)P(j)\log(P(i)P(j))$$

$$f_{12} = \frac{H_{XY} - H'_{XY}}{\max(H_X, H_Y}$$

13. Information Measure of Correlation #2

$$f_{13} = (1 - e^{-2(H''_{XY} - H_{XY})})^{1/2}$$

14. Maximal Correlation Coefficient.

$$f_{14} = \sqrt{\lambda_2}$$

where:

$$\det(Q(i,j) - \lambda I) = 0$$

$$Q(i,j) = \sum_k \frac{P(i,k)P(j,k)}{P_x(i)P_y(k)}$$

A drawback of these measures is that, for a given distance, a large number of features can be computed ([Tuceryan 98]) and some kind of feature selection method must be used to select the most relevant ones.

In figure 3.4 the difference between using variance as candidate selection criterion or other texture (in this case, entropy, f_9) is shown. On the right image, if entropy is used as criterion, the selected features are mainly groups of letters: these would be more difficult to find on other classes at the same locations.

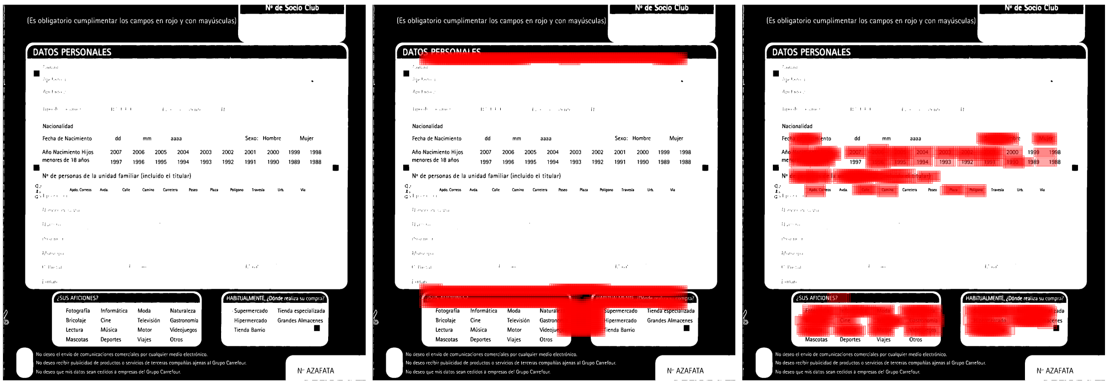

Nº de Socio Club
(Es obligatorio cumplimentar los campos en rojo y con mayúsculas)
DATOS PERSONALES
Nacionalidad
Fecha de Nacimiento
Sexo: Hombre Mujer
Año Nacimiento Hijos menores de 18 años
Nº de personas de la unidad familiar (incluido el titular)
¿SUS AFICIONES?
Fotografía Informática Moda Naturaleza
Bricolaje Cine Televisión Gastronomía
Lectura Música Motor Videojuegos
Mascotas Deportes Viajes Otros
HABITUALMENTE, ¿Dónde realiza su compra?
Supermercado Tienda especializada
Hipermercado Grandes Almacenes
Tienda Barrio
Nº AZAFATA

Figure 3.4: Effect of using a texture measure to select candidate features. From left to right: original image, first 200 selected features using variance as criterion, first 200 selected features using entropy as criterion.

3.5 The reject option

Document retrieval is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.

There is an implicit parallelism between document retrieval and document identification. In the context of this work, it refers to correctly ascertain to which class a test document belongs (relevant documents), or finding out that it does not belong to any of the classes in the system (non-relevant documents). To implement this functionality, the system needs a reject option.

To evaluate the power of the system for rejecting unknown documents, three measures are used in the experiments (section 4): recall at 100% precision, area under the ROC curve and KL-divergence.

3.5.1 Recall and precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need. Its formula is:

$$precision = \frac{relevant\ documents\ retrieved}{documents\ retrieved}$$

Precision can have a value between 0 and 1. In a perfect retrieval, only relevant documents are retrieved and has thus a precision value of 1.

Recall is the proportion of relevant documents retrieved, compared to the amount of relevant documents in the database:

$$recall = \frac{relevant\ documents\ retrieved}{relevant\ documents}$$

It is trivial to achieve a precision of 1 if there is no regard for the recall value, and vice versa. The strategy would simply be to retrieve no documents, or to retrieve all of them, respectively.

Because of this a widely used measure is *"recall at x% precision"*, which indicates the recall ratio that can be attained with a predetermined x% precision. Generally, *"recall at 100% precision"* is used.

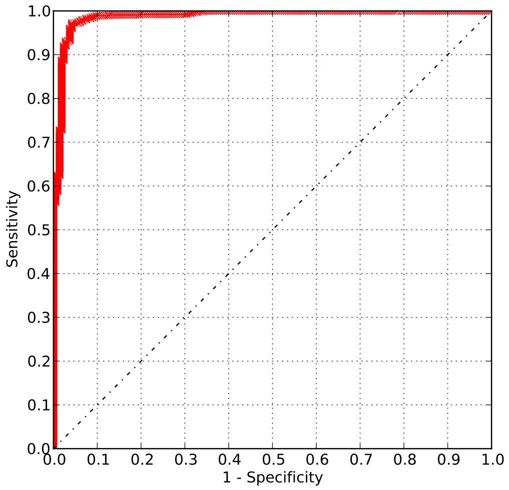


Figure 3.5: A ROC curve from the results given in figure 4.2 (with scale=1:4). The AUC is 0.98819.

In this work, the recall at 100% precision was computed by ordering the results by average correlation value, which can be seen as a *confidence measure* and counting from top to bottom up to the first mistake. In this context, "mistake" can be either a document misclassified as a wrong known class, or an unknown document having a higher value than other known documents. The formula would be:

$$recall\ at\ 100\%\ precision = \frac{known\ documents\ correctly\ classified\ until\ mistake}{known\ documents}$$

That value reaches 100% only when all the known documents score higher than unknown documents and the error rate is 0%.

3.5.2 ROC curve

In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the true positive rate (or sensitivity) versus the false positive rate (or $1-$specificity). In this context it represents the trade-off of a system between incorrectly rejecting a document when it was from a known class, or incorrectly accepting a document from an unknown class (or misclassifying it into another known class).

In the results shown in section 4.3, the area under the ROC curve (AUC) is given. This value summarizes the ROC curve in a single number going from 0.0 to 1.0. The AUC is also equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

In figure 3.5 an example of a ROC curve is given.

3.5.3 KL-divergence

The Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions, P and Q. It is normally defined as:

$$D_{KL}(PQ) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3.6}$$

As noted in [Arlandis 09], a good behaviour for these distance-based methods can be expected only if the distribution of the dissimilarities (or similarities) of a document to its own class does not overlap with the distribution relating to different classes (either other classes from the corpus or the *unknown* class). The KL-divergence measure can thus be a good indicator of the degree of overlap between distance distributions.

As the KL-divergence as defined in equation 3.6 is asymmetric, a compromise can be found by using the minimum distance beteween distributions, as in:

$$mD_{KL}(PQ) = \min(\sum_i P(i) \log \frac{P(i)}{Q(i)}, \sum_i Q(i) \log \frac{Q(i)}{P(i)}) \tag{3.7}$$

In section 4, two KL-divergence values were computed for each experiment:

- One between the distribution P of similarities of a document to its own class versus the distribution Q of similarities to other classes.
- The other between the distribution P of similarities of a document to its own class versus the distribution Q of similarities to unknown documents.

The individual features' correlation values were used for the computations.

Because the correlation values of the features can be zero a significant number of times (by convention, they are nullified when a feature for a class falls out of bounds for a given test document), the KL-divergence, as computed in equation 3.7 has no upper bound, due to being undefined when $\log(0)$ appears. Because of this, the results given in section 4.3 are relative to the maximum KL-divergence attainable with the number of features in that particular case, computed with this *relative* KL-divergence function:

$$rD_{KL} = \frac{mD_{KL}(PQ)}{\max D_{KL}} \tag{3.8}$$

where $\max D_{KL}$ is the KL-divergence obtained by a distribution where the correlation between the features of a given class with documents of its own class is maximum and with other classes is minimum, having the same number of features as $mD_{KL}(PQ)$.

This is a relative measure, so it has no intrinsic value. In the experiments it was mainly used to compare the performance of the different texture filters.

Chapter 4

Experiments

4.1 Corpus description

In order to make the experiments, it was necessary to build a database with documents useful for the task at hand. Initially the first intention was to use a standard corpus with which the results could be compared, but none was found that was fully adapted to the needs of this project. The only one available in the literature is the NIST corpus SD6 [Dimmick 92], but it only contains 20 types of documents, so it was decided to fill it with documents from other sources. The corpus that was finally used comprises:

- 20 document classes from the SD6 NIST standard database. They are all federal tax forms from the United States, with synthesized simulated handwritten information in them. The handwriting synthesis method and other details of this corpus are explained in [Dimmick 92].
- 47 document classes scanned for the author's research group from personal invoices, bank receipts, etc. All of them with variable printed content, different sizes and aspect ratios

In all cases, the documents were mechanically scanned from printed originals, thus all of them have rotation defects. The corpus contains documents with different formats, sizes and types, so it was necessary a preprocessing phase in order to convert them to a single format. The chosen final resolution was 300 dpi.

Table 4.1 shows a detailed description of the document classes used in the experiments, with the labels assigned to each class, a thumbnail image of an example from that class, the approximate size, image type (file format and pixel depth), document type (form, invoice, delivery note, receipt), origin (from the aforementioned types) and number of items available in that class for experiments.

In order to complete the experiments by way of considering a reject option, 200 forms obtained from different sites in Internet, all different between them and also different to the abovementioned classes. These formed the class of unknown documents.

In the experiments, a validation/test partition was not performed because the number of known classes available for the experiments was very limited fo rthe problem tackled in this work.

Table 4.1: Corpus used in the project.

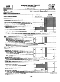
Label	Thumbnail	Size	Type	Document	Source	Num. items
1040		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
1041		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
2106		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
2107		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
2441		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
4562		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
4563		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
6251		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
AGAG		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**11**
ALDA		1648x719	TIFF 1-bit B/N	Invoice	ITI	**13**
AME_		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
ARCO		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
BCD_		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
BNCJ		2362x1181	TIFF 8-bit Gris	Invoice	[Castelló-Fos 11]	**13**
CAM_		1656x2339	PNM 1-bit B/N	Invoice	ITI	**11**
CENS		3508x2480	TIFF 8-bit Gris	Form	ITI	**13**
CIDA		2480x2362	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
CIDF		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**

Table 4.1 – Continued

Label	Thumbnail	Sizeï¿½o	Type	Document	Source	Num. items
COLT		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**9**
COMU		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**10**
CORR		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
CRDT		2480x1181	TIFF 8-bit Gris	Invoice	[Castelló-Fos 11]	**13**
DEDA		2480x1771	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
DEDF		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
EDU1		2410x3549	PNM 8-bit Gris	Form	ITI	**13**
EDU4		2438x3537	PNM 8-bit Gris	Form	ITI	**13**
EDU7		2413x3513	PNM 8-bit Gris	Form	ITI	**13**
EHLS		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
EMVI		2480x3508	TIFF 1-bit B/N	Form	ITI	**13**
FARN		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
FOXN		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
IBDL		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
IBER		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
INCO		2480x2362	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**12**
LEDE		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**

Table 4.1 – Continued

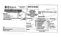
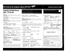
Label	Thumbnail	Sizeï¿½o	Type	Document	Source	Num. items
LOTE		1530x1600	TIFF 1-bit B/N	Form	ITI	**13**
MARA		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**12**
MONX		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
MULT		794x1382	TIFF 1-bit B/N	Form	ITI	**13**
MURC		2480x1175	TIFF 1-bit B/N	Invoice	ITI	**13**
NACX		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**11**
NIST		2560x3300	PNM 1-bit B/N	Form	ITI	**13**
ONO_		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
RICO		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**13**
RIEL		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
ROLS		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
SARA		3406x2517	TIFF 1-bit B/N	Form	ITI	**13**
SCC1		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCC2		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCD1		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCD2		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCE1		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCE2		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**

Table 4.1 – Continued

Label	**Thumbnail**	**Sizeï¿½o**	**Type**	**Document**	**Source**	**Num. items**
SCF1		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCF2		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCHA		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SCHB		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SHEL		2542x3551	TIFF 1-bit B/N	Form	ITI	**13**
SONI		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
SSE1		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
SSE2		2560x3300	PNM 1-bit B/N	Form	SD6 NIST	**13**
TIMB		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
UNOE		2480x1181	TIFF 8-bit Gris	Invoice	[Castelló-Fos 11]	**13**
VICE		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
WATR		1656x2339	PNM 1-bit B/N	Invoice/Receipt	ITI	**9**
YOI2		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**
YOIG		2480x3507	TIFF 8-bit Gris	Invoice/Receipt	[Castelló-Fos 11]	**13**

4.2 Performance indices

The results obtained in the experiments have been presented using for different indices: error rate, recall at 100% precision, area under the ROC curve and KL-divergence. *Error rate* is a straightforward metric, which does not use the unknown documents class. *Recall at 100% precision*, *area under the ROC curve* and *KL-divergence* have been used to evaluate a rejection option using the set of unknown documents class as explained in section 3.5.

4.3 Parameter optimization and experiment results

The experiments were designed in four phases:

- Tentative experiments to set δ-landmark and search window sizes.
- Experiments using the variance as a filter for candidate feature selection, in order to determine the suitable scale factor to be applied to the document images as a function of the system performance, including the processing time.
- Test of all the texture filters in order to compare their performance indices.
- Test different document scale factors using the texture that gives the best performance.

4.3.1 δ-landmark size and search window size

In [Arlandis 09] it was established that the best result for this kind of techniques are obtained using δ-landmark widths from 48 to 80 pixels, which correspond roughly to the average width of a bunch of characters in a 300-dpi scanned page, with a height of the line of characters in the neighbourhood of 36 pixels. Based on this, tentative experiments were carried out and finally a δ-landmark size of 80x36 pixels was chosen.

The search window size used in the experiments was 200 pixels, to compensate the maximum displacement between two forms of the same class in the corpus available for these experiments.

4.3.2 Scale tests using the variance filter

Figures 4.1, 4.2, 4.3 and 4.4 show that even when using only the variance as parameter for feature candidate preselection (as noted in section 3.4.2), the performance of the system can be extremely good. It is of note that performance actually is better when the scale is 1:2 than when the scale is 1:1. This can be due to the fact that moderate values of scaling can smooth the images enough for features that have been scanned with some differences to match with images of the same class; while preserving the differences between features that are really from different classes. The results seem consistent with this idea, with the best results obtainable with scales between 1:2 and 1:6.

That idea is also consistent with figures 4.9 and 4.10, which show the time taken to train all the 67 classes used, and to indentify a document on average, using different scale values on a six-core AMD Opteron with 2.2 GHz. This shows that scales above 1:4 could tend to be somewhat unusable in real time processing with many document classes. As a conclusion, 1:4 seems a good intermediate point from which to enhance the results, as it keeps the processing times low and was the scale factor chosen for the next phase.

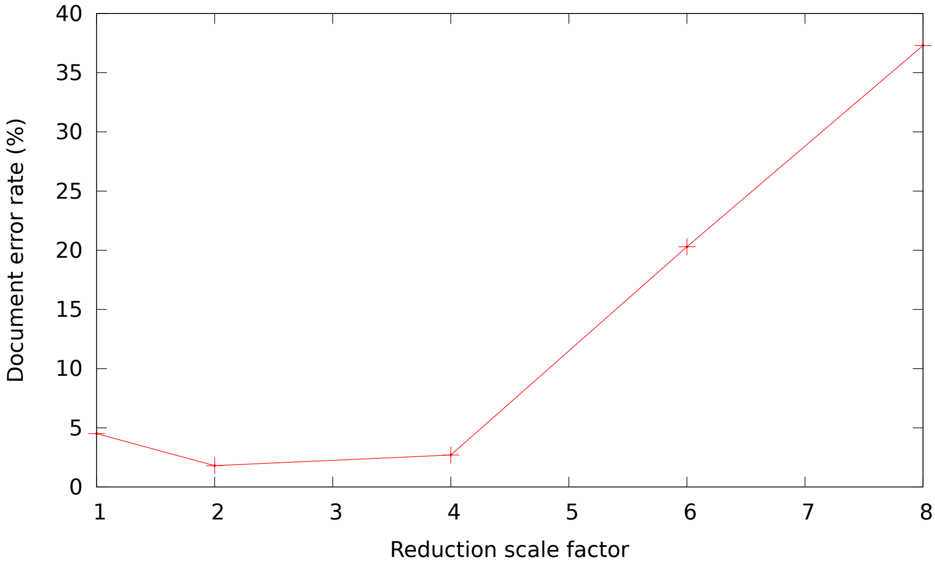


Figure 4.1: Error rate depending on the reduction scale of the document images using variance as candidate δ-landmark feature selection criterion.

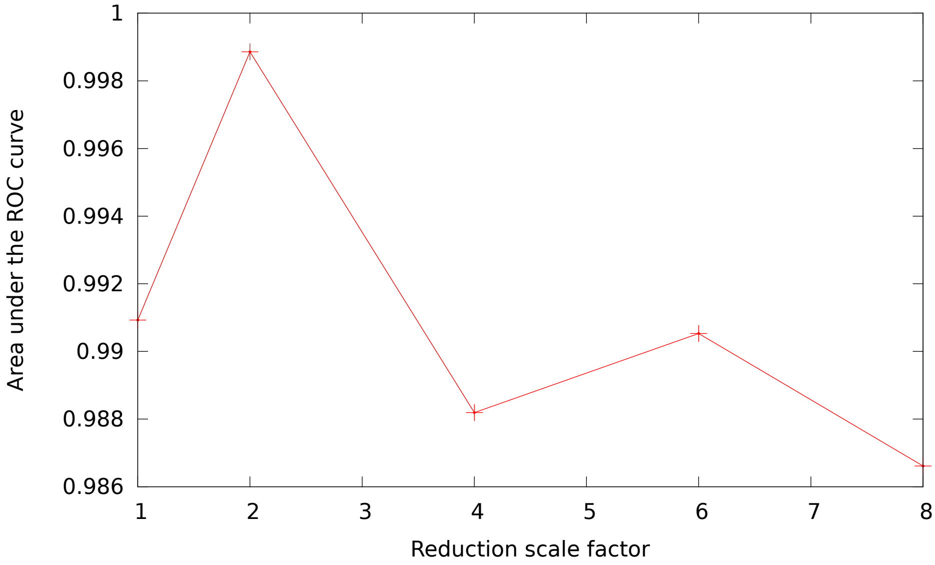


Figure 4.2: Area under the ROC curve depending on the reduction scale of the document images using variance as candidate δ-landmark feature selection criterion.

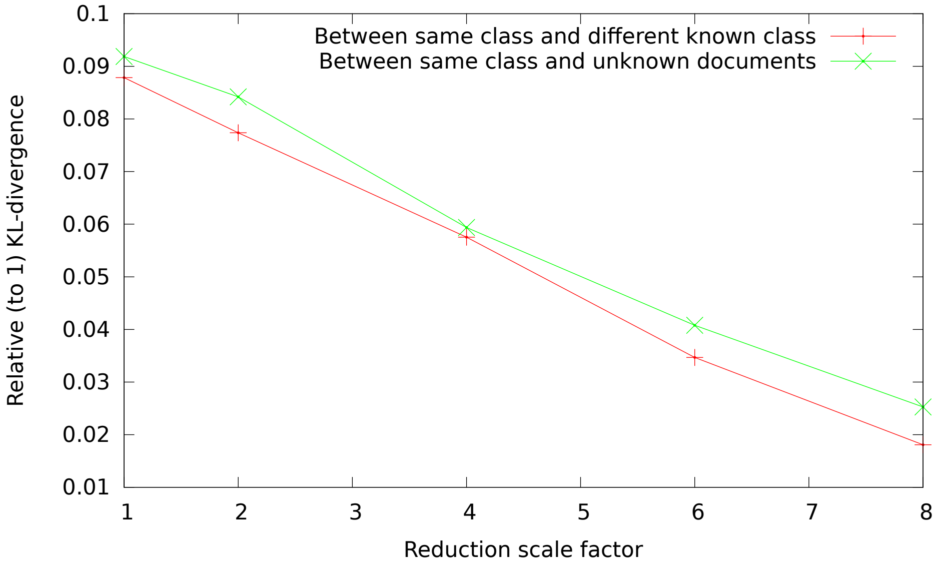


Figure 4.3: (relative-to-one) KL-distance values depending on the reduction scale of the document images using variance as candidate δ-landmark feature selection criterion.

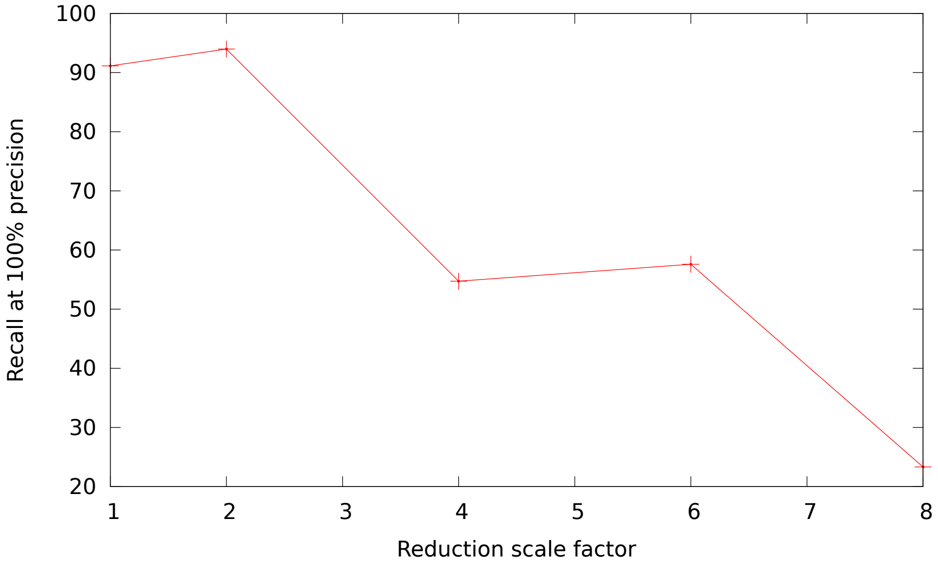


Figure 4.4: Recall at 100% precision depending on the reduction scale of the document images using variance as candidate δ-landmark feature selection criterion.

4.3.3 Test of texture filters

The texture filters explained in section 3.4.2 were tried, always with a fixed scaling of 1:4. All 14 texture measures were tried, with varying displacements between 1 and 5 pixels, and both trying with the minimum value obtained from the 4 possible angles and the average value.

Table 4.2: Best possible error rate, area under ROC, KL-divergence against other classes, KL-divergence against unknown documents and recall at 100% precision for each texture type.

Texture type	Error Rate (%)	Area under ROC	KL-d.known	KL-d. unknown	Recall at 100% prec. (%)
f_1	84.96	0.98085	0.020329	0.015586	6.44
f_2	0.30	0.99996	0.088838	0.101268	99.38
f_3	4.60	0.98938	0.054937	0.057100	65.64
f_4	0.00	0.99921	0.077776	0.076093	97.54
f_5	79.44	0.9779	0.027242	0.025529	6.44
f_6	0.00	0.99989	0.052870	0.059370	99.69
f_7	0.00	0.9995	0.063585	0.065660	98.46
f_8	0.30	0.99826	0.052374	0.062216	68.09
f_9	1.22	0.99663	0.090771	0.119281	93.25
f_{10}	44.47	0.98470	0.032569	0.045577	42.02
f_{11}	0.00	0.99996	0.076772	0.089868	99.69
f_{12}	29.44	0.99101	0.034232	0.045535	53.68
f_{13}	38.65	0.99330	0.044686	0.049929	48.46
f_{14}	19.63	0.98739	0.029846	0.038085	47.23

Table 4.2, shows the best results achievable with each texture type, with 1:4 scaling. With some texture types a 0% rate of error can be achieved. Particularly, a 0% error rate can be obtained with texture types f_4 (variance), f_6 (sum average), f_7 (sum variance) and f_{11} (difference entropy). Also, recall at 100% precision and the ROC curve indices were very good when testing unknown documents. For a given texture type, the differences in performance from using different displacement values and angles were non-significant.

4.3.4 Retrying best parameters found with different scale levels

Of all the combinations tried, the best one in all areas except on relative KL-divergence was: Difference entropy (f_{11}) with a displacement of 2 pixels, used in decreasing order and using the minimum result obtained from the four angles. This combination had 0% error rate, 99.69% recall rate at 100% precision, and an area below the ROC curve of 0.99996.

The test set was rerun with those parameters but with different scalings, to see if a different scale could improve or at least repeat the results. Figures 4.5, 4.6 and 4.8 show that a scaling of 1:4 is the best with those parameters.

Curiously, figures 4.3 and 4.7 show that the KL-divergence measure seems to depend on the scale value. This means that the KL-divergence measure only correlates with the error rate when scaling is constant, as in table 4.2.

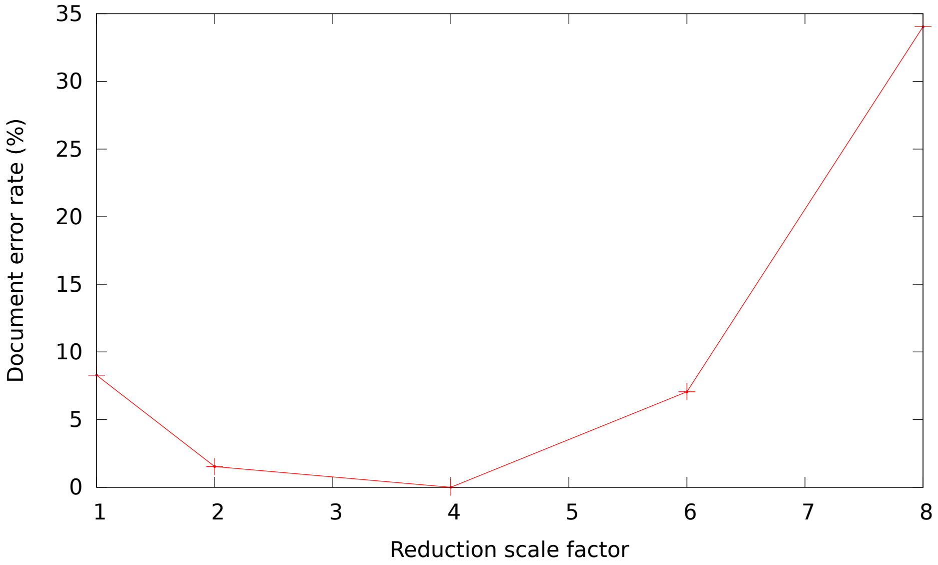


Figure 4.5: Error rate depending on the reduction scale of the document images using difference entropy as candidate δ-landmark feature selection criterion.

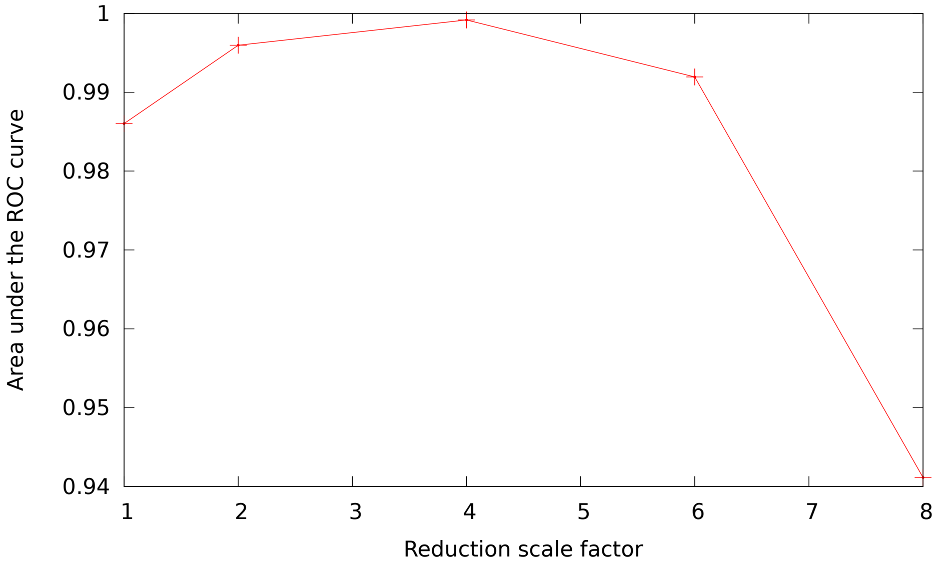


Figure 4.6: Area under the ROC curve depending on the reduction scale of the document images using difference entropy as candidate δ-landmark feature selection criterion.

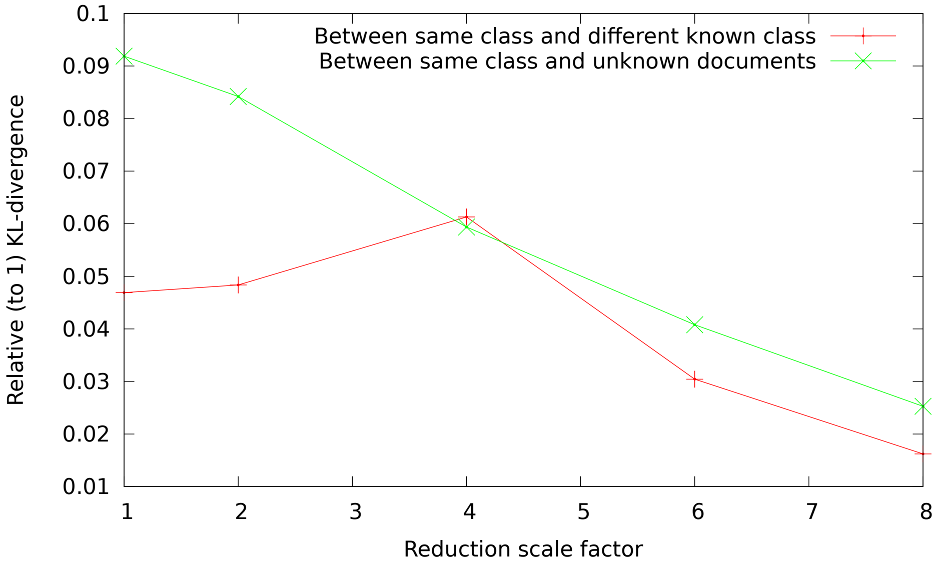


Figure 4.7: (relative-to-one) KL-distance values depending on the reduction scale of the document images using difference entropy as candidate δ-landmark feature selection criterion.

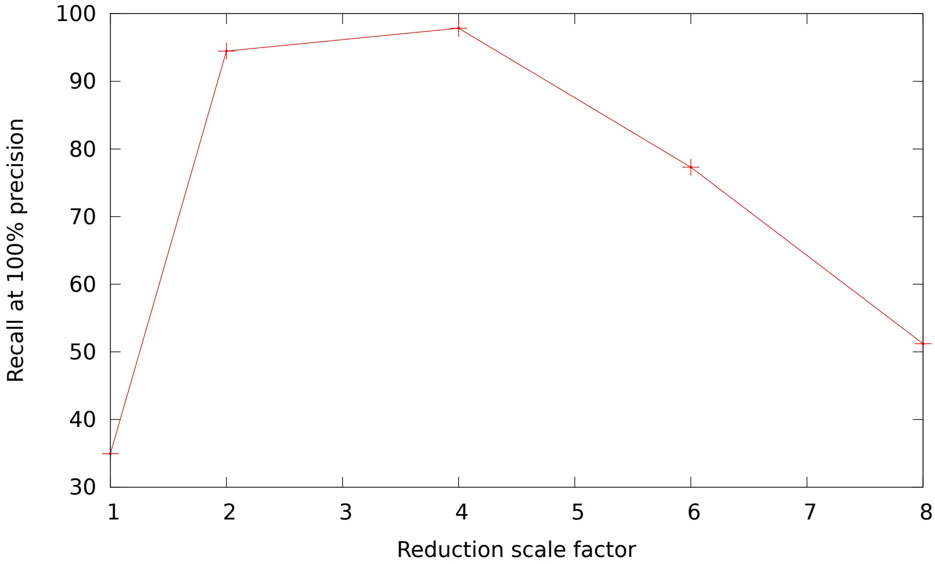


Figure 4.8: Recall at 100% precision depending on the reduction scale of the document images using difference entropy as candidate δ-landmark feature selection criterion.

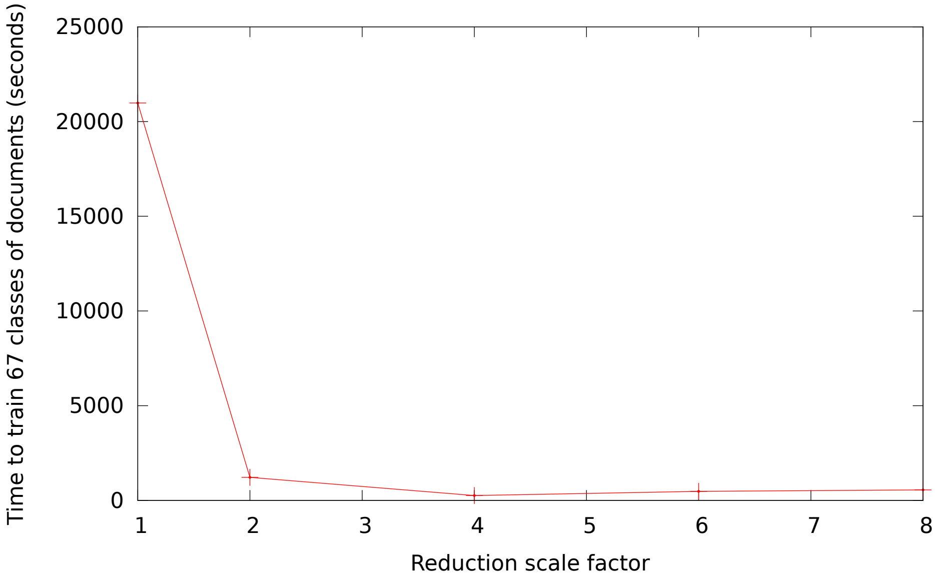


Figure 4.9: Total time of training the 67 classes depending on the reduction scale of the document images.

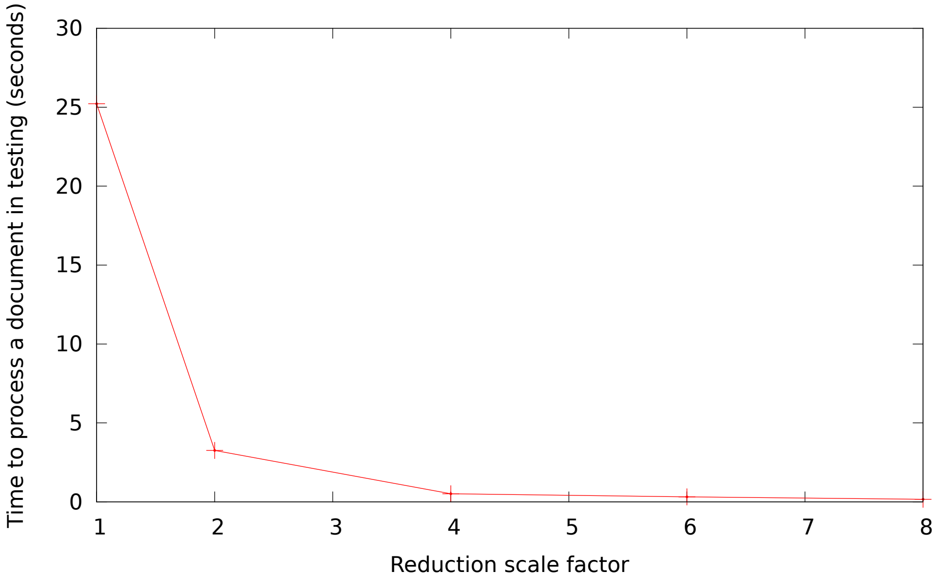


Figure 4.10: Time for identifying a single document depending on the reduction scale of the document images.

Chapter 5

Conclusions

In this work a technique is proposed to classify documents with total flexibility of designs, layouts, sizes, and amount of filled-in contents in an efficient way. The technique also allows the identification of images not belonging to the set of expected documents via a reject option.

The method consists of an off-line phase for automatic location of the most discriminant regions of a set of forms, which we called δ-landmarks, that are then used as features for a distance-based classifier.

An extensive database with document images has been collected for conducting the experiments, and it could be useful for future work related to identification of filled-in documents.

The experimentation identified the optimal values for the relevant parameters, achieving a 0% error rate in document classification and very good results in unknown document rejection.

The processing times were low enough to use a system based on this technique in a large-scale document identification application.

Bibliography

[Andreu-Cerezo 10] Luis Andreu-Cerezo. Detección y modelización de casillas de campos de formularios. Master's thesis, Universitat Politècnica de València, 2010.

[Arlandis 03] Joaquim Arlandis. *La transformació contínua de la distància. Estudi i aplicació a un sistema OCR*. PhD thesis, Universitat Politècnica de València, 2003.

[Arlandis 09] Joaquim Arlandis, Juan-Carlos Perez-Cortes & Emilio Ungria. *Identification of very similar filled-in forms with a reject option*. In ICDAR, pages 246–250, 2009.

[Arlandis 11] Joaquim Arlandis, Vicent Castello-Fos & Juan-Carlos Perez-Cortes. *Filled-in document identification using local features and a direct voting scheme*. In IbPRIA, 2011.

[Carrion-Robles 11] Diego Carrion-Robles, Vicent Castello-Fos, Juan-Carlos Perez-Cortes & Joaquim Arlandis. *Two methods for filled-in document image identification using local features*. In ICDAR, 2011.

[Castelló-Fos 11] Vicente Castelló-Fos. Un sistema automático de identificacion de tipos de documentos. Master's thesis, Universitat Politècnica de València, 2011.

[Crow 84] F. C. Crow. *Summed-area tables for texture mapping*. Computer Graphics - Proceedings of SIGGRAPH'84, vol. 18(3), pages 207–212, 1984.

[Dimmick 92] D. L. Dimmick & M. D. Garris. *Structured Forms Database 2, NIST Special Database 6*. Rapport technique, National Institute of Standards and Technology, 1992.

[Doermann 98] D. Doermann. *The indexing and retrieval of document images: A survey*. Computer Vision and Image Understanding, vol. 70(3), pages 287–298, 1998.

[Fan 01] K.-C. Fan, M.-L. Chang & Y.-K. Wang. *Form document identification using line structure based features*. ICDAR, pages 704–708, 2001.

[Haralick 73] Robert M. Haralick, K. Shanmugan & Its'Hak Dinstein. *Textural Features for Image Classification*. IEEE Transactions on Systems, Man, and Cybernetics, vol. 3, pages 610–621, 1973.

[Heroux 98] P. Heroux, S. Diana, A. Ribert & E. Trupin. *Classification method study for automatic form class identification*. Proc. 14th Int. Conf. on Pattern Recognition, ICPR'98, pages 926–928, 1998.

[Mandal 05] S. Mandal, S. P. Chowdhury, A. K. Das & B. Chanda. *A hierarchical method for automated identification and segmentation of forms*. Document Analysis and Recognition, International Conference on, vol. 0, pages 705–709, 2005.

[Nagasaki 06] T. Nagasaki, K. Marukawa, T. Kagehiro & H. Sako. *A coupon classification method based on adaptive image vector matching*. 18th International Conference on Pattern Recognition, pages 280–283, 2006.

[Ogata 03] H. Ogata, S. Watanabe, A. Imaizumi, T. Yasue, N. Furukawa, H. Sako & H. Fujisawa. *Form-type identification for banking applications and its implementation issues*. DRR, pages 208–218, 2003.

[Ohtera 04] R. Ohtera & T. Horiuchi. *Faxed form identification using histogram of the hough-space*. Pattern Recognition, International Conference on, vol. 2, pages 566–569, 2004.

[Parker 10] Charles Parker. *Anchor point selection by KL-divergence*. Image Processing Workshop (WNYIPW), pages 42–45, 2010.

[Sako 03] H. Sako, M. Seki, N. Furukawa, H. Ikeda & A. Imaizumi. *Form Reading based on Form-type Identification and Form-data Recognition*. 7th International Conference on Document Analysis and Recognition, pages 926–930, 2003.

[Sarkar 06] Prateek Sarkar. *Image classification: Classifying distributions of visual features*. International Conference on Pattern Recognition, pages 472–475, 2006.

[Sarkar 10] Prateek Sarkar. *Learning Image Anchor Templates for Document Classification and Data Extraction*. International Conference on Pattern Recognition, pages 3428–3431, 2010.

[Shafait 08] F. Shafait, D. Keysers & T. M. Breuel. *Efficient implementation of local adaptive thresholding techniques using integral images*. Document Recognition and Retrieval, vol. XV (San Jose, USA), 2008.

[Ting 96] A. Ting & M. Leung. *Business form classification using strings*. ICPR96, vol. II, pages 690–694, 1996.

[Tuceryan 98] Mihran Tuceryan & Anil K. Jain. *Texture Analysis*. The Handbook of Pattern Recognition and Computer Vision (2nd Edition), pages 207–248, 1998.

[Viola 04] P. Viola & M. J. Jones. *Robust real-time face detection*. Int. Journal of Computer Vision, vol. 57(2), pages 137–154, 2004.