A Statistical Method for an Automatic Detection of Form Types

Saddok Kebairi, Bruno Taconet, Abderrazak Zahour, and Said Ramdane

Laboratoire d'Informatique du Havre, Université du Havre
Place Robert Schuman, 76610, Le Havre, France
{kebairi,taconet,zahour,ramdane}@iut.univ-lehavre.fr

Abstract. In this paper, we present a method to classify forms by a statistical approach ; the physical structure may vary from one writer to another. An automatic form segmentation is performed to extract the physical structure which is described by the main rectangular block set. During the form learning phase, a block matching is made inside each class ; the number of occurrences of each block is counted, and statistical block attributes are computed. During the phase of identification, we solve the block instability by introducing a block penalty coefficient, which modifies the classical expression of Mahalanobis distance. A block penalty coefficient depends on the block occurrence probability. Experimental results, using the different form types, are given.

1   Introduction

An important problem in an automatic form reading system is the form type identification. As Doermann et al.[1] said, the form identification lies essentially on an appropriate choice of the information primitives extracted by the document segmentation. Several methods for matching forms structures have been reported in literature: for the construction of the model, Mao et al[2] use the attributes of the horizontal and vertical lines, of the junction points and the regions. However, there must be at least two horizontal lines in every form for reliable recognition and registration. The authors[3] used three types of line segment to represent a form. A fuzzy matching is used for the form recognition. Ishitani[4] presented a method that operates in a hierarchical way. At first, a line matching is constructed, which gives a compatibility graph. On this graph, the first clique is searched to identify the homogenous regions. The used compatibility criterion takes in account the similarities based on the intersection number in the lines. The author explains his choice by the robustness, insensibility to scale variation and noise distortion. Another approach[5] proposed a model based on detecting lines as basic items. A matching graph is proposed for an automatic localization and the extraction of the form field-in data.[6] Another model[7] is based on attributed relational graphs and the system performs form

registration and location fields using algorithms based on the hypothesize-and-verify paradigm. The authors[8] described a system that locates form structures by matching horizontal and vertical lines between two forms. The approach is based on the A^*-searh-algorithm. The system proposed by Dubiel et al.[9] runs independently from any restriction on form layout and requires neither an identification number nor any prespecified   line structure. The classification is done by searching counterparts of characteristic blocks of text lines of each reference pattern in filled form.   The authors[10] described three classifiers which focus on form identification. For the first two, the information based on a pyramidal image decomposition is used by the k-Nearest-Neighbor and the Multilayer Perceptron. The third uses the information extracted from the form content as a tree structure.

     This paper describes a system which allows to identify automatically different types of forms **without any sign of reference**. A method for the extraction of the main representative rectangular blocks of the physical structure document is presented. The required thresholds for this extraction are determined automatically and are adapted to the form type. In the phase of training, models are described by a vector of features that integrates elements of statistical nature (average, standard deviation). The difficulty lies in the fact that, for several samples of a given model, the obtained blocks are not necessarily stable. They can group together (phenomenon of the merging block) or divide in several blocks (phenomenon of the fragmentation block). Therefore, every model is going to appear according to several possible configurations of blocks. During the form learning phase, the probability of occurrence of every block is counted. During the phase of identification, we watch the blocks' instability while introducing a coefficient of penalty based on this probability of occurrence for every block. The decision of affecting a form to a class is taken by calculating a distance between the unknown form and the profile of matching of every model, if one exists. Recently,[11] we used the Mahalanobis distance which could not be used in the general case, because the vector of feature has a variable dimension. For this reason, we propose a modified expression of this distance. This modified distance is enriched by a balancing of the penalty affected to every block. The more rarely the concerned block appears, the more this penalty increases the distance. Thus the smallest distance calculated in this way, permits to affect to the unknown form the nearest class.

     This paper is organized as follows: In section 2, we describe the general organization of our system. Section 3 presents the automatic form segmentation. Section 4 describes the learning mechanism. Section 5 describes the form type identification procedure. Experimental results and conclusion are presented in section 6.

2   General Structure of Our System

     The general structure of our system is illustrated in figure 1. During the phase of learning, first of all we, define the filled items:[12] the professor localizes the rectangular zones of insertion of handwritten data, and enters the attributes of the support of the handwritten data (number and type of support: rectangular boxes,

continuous or dotted reference lines, etc.). For this use, we have developed a graphical user interface (Figure 2). Secondly, a vectorial statistical model of every class is constructed automatically. This model takes into account the attributes of the main rectangular blocks which define the physical structure of the document.
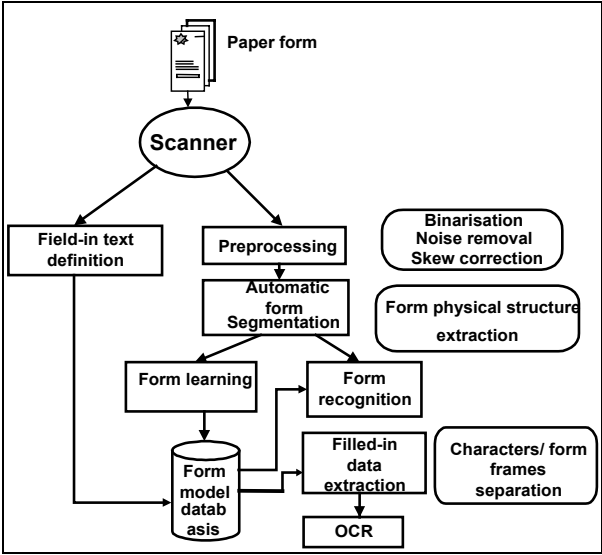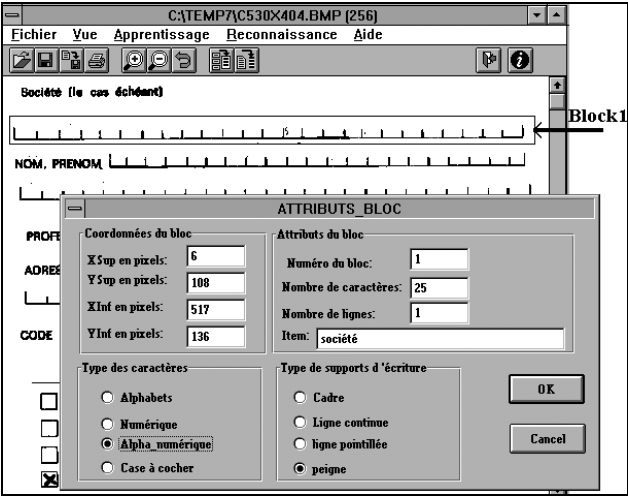

Fig. 1. General structure of our system

Fig. 2. Learning of the first block of a form

3   Form Automatic Segmentation

The process of the automatic segmentation into main rectangular blocks of text is described as follows: we begin by extracting all the white inscribed maximal rectangles by using the algorithm of rectangulation.[13,14] We construct the distribution of these white rectangles versus two variables ; length of horizontal side, length of vertical side (Figure 3). We are going to exploit this distribution in order to determine the two threshold rectangles: one is vertical, the other one is horizontal. A threshold rectangle is the minimum boundary of the text block. So, these threshold rectangles are adapted automatically to the type of the form. The exploitation of the distribution is made as following: we note that in the region around the origin (small x, small y) the peaks are very elevated and the density is very strong ; these rectangles represent the inter-character separators and inter- word separators. The horizontally lengthened rectangles (big x, small y) are few and represent the horizontal inter-block spaces. The vertically lengthened rectangles (small x, big y) are also few. They represent the inter-blocks and inter-columns. The aim is to retain only inter-block and inter-column separator rectangles. For this, we construct a separator line in the plane of the distribution, in two steps: first of all, we calculate the length average of the horizontal side (resp. vertical) m_h (resp. $m_{v)}$ and the corresponding standard deviation s_h (resp. s_v), for all white horizontal rectangles (resp. vertical). The broken line whose sides are parallel to the axes and of abscissa $m_h+s_h/2$ (resp. ordinate $m_v+s_v/2$) forms the first separating line (Figure 5). Afterwards, we only keep the rectangles situated beyond this separating line. In the second step, we calculate the average m_{ch} (resp. m_{cv}) of the length (resp. of the width) of the remaining rectangles. The real separator line is the broken line, for which the sides are parallel to the axes and of abscissa m_{ch}(rep. of ordinate m_{cv}). The horizontal threshold rectangle (resp. vertical) is chosen as the nearest rectangle to the separating line in the authorized corresponding zone. Therefore, the set of the text separators is given by all white rectangles whose sizes are superior or equal to a threshold rectangle. Then, we construct the complement of the image of all the text separators. The minimal bounding rectangular blocks are then searched in the text blocks.
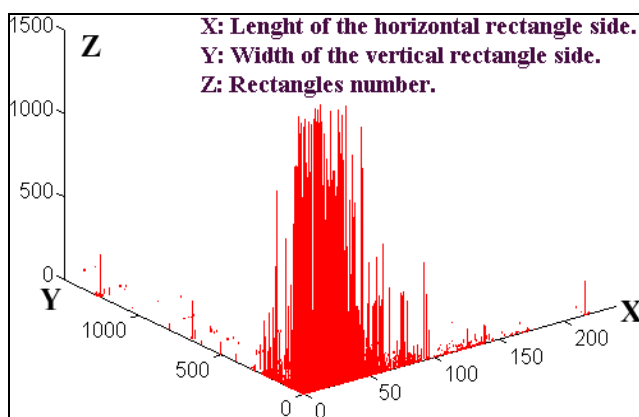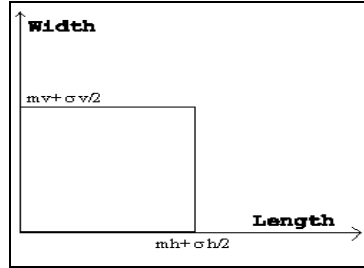

Fig. 3. White rectangles distribution

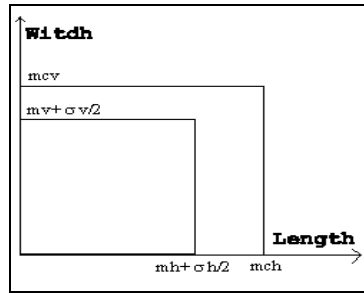**Fig. 4.** First separator

Fig. 5. Second separator

4   Form Type Learning

4.1   Detection of the Phenomenon of Merging and Fragmentation

In the general case, the disposition of blocks and the separator spaces are variable owing to the position or the textual information size variations introduced by the different writers. It is manifest when the handwritten writing overflows the data fields. This variability generates two phenomena that result in several configurations for one same type of form (Figure 6). These two phenomena are given by either the merging and the fragmentation of blocks or change of the size of one or several blocks independently of others. For the clarity of exposing our approach, we limit the discussion to the case of having two forms of the same class. Be C a class of a type form and $F_1, F_2 \in C$; two filled-in forms of which physical structure is represented respectively by rectangular blocks set: E1={$b_{11},b_{12},...,b_{1n}$}, E2={$b_{21},b_{22},...,b_{2m}$} where b_{ik} represents a block of k label belonging to the F_i form , n and m designate respectively the finished block numbers of the F_1 and F_2 physical structure form. Every block $b_{ik} \in E_i$ is characterized by a vector V of attributes: V={$x_{ik},y_{ik},l_{ik},h_{ik}$} where x_{ik}, y_{ik} represent the rectangle center coordinates defining the block k $\in E_i$; l_{ik}, h_{ik} designate respectively the length and height of the block b_{ik}. The matching of blocks of F_1 with those of the F_2 form is made according to a criterion of Euclidean distance between their center position attributes:

$$d(b_{1k}, b_{2l}) = \left[(x_{1k} - x_{2l})^2 + (y_{1k} - y_{2l})^2 \right]^{1/2} \tag{1}$$

where k=1,...n and l=1,...m. To a block $b_{1k} \in F_1$ corresponds a block $b_{2k'} \in F_2$ if the following condition is satisfied:

$$d(b_{1k}, b_{2k'}) = \min_l \left[(x_{1k} - x_{2l})^2 + (y_{1k} - y_{2l})^2 \right]^{1/2} \tag{2}$$

An erroneous matching between blocks, could occur during a phenomenon of merging or fragmentation. Indeed a big block could be matched to several small corresponding blocks. In this case, there is a conflict, therefore, no bijective matching can be done. However, we must establish a correspondence between the big block of the picture of one sample and the small blocks of another one.

We are going to define the 8-neighborhood of a block, then we will associate to every block a vector of attributes that characterizes the relation with the 8 neighbors. We got inspiration from Allen's[15] and Walischewski's[16] works. Any block possesses 8 neighbors, the picture sides being sometimes considered as a neighbor's border, if needs be. A neighboring block can be in one of the eight directions given in Figure 7. Figure 8 shows the 13 positions of the horizontal side of the superior block.[13] Figure 9 and Figure 10 show the restriction to 9 and 4 positions that we made respectively. In the same way, the respective positions of the 7 other neighbors can be easily deducted.
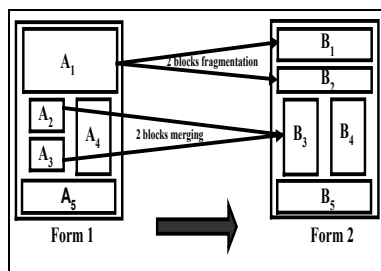
Fig. 6. Merging and fragmentation phenomena
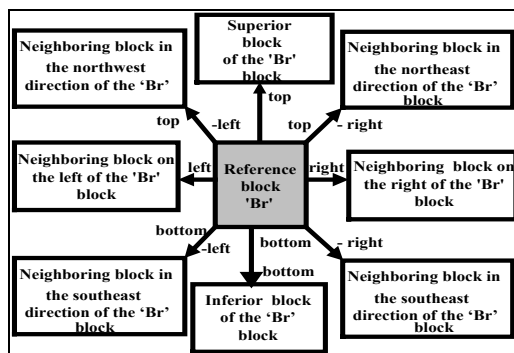
Fig. 7. 8-neighborhood of a block
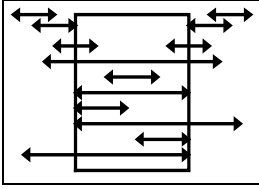
Fig. 8. The 13 relative positions the superior block[13]
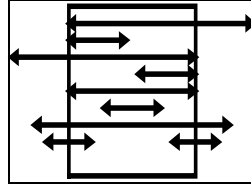
Fig. 9. The 9 relative positions of the inferior bloc
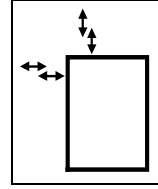
Fig. 10. The 4 relative positions of the Northwest block

The attributes of the vector are divided in two types (Figure 11): the first is the separator distances between a block and its 8 neighbors, and the second type is the heights respectively (lengths) of the neighboring blocks according to the vertical direction respectively (horizontal). Both types number twelve each. Let's investigate, the block merging of the block b_1 and its right neighbor b_2. After merging, b_1 and b_2 will be matched with the B block (Figure 12). Figure 13 illustrates attributes of a block in a F_i form (F_1 or F_2). The principal cases of a block elongation and of its neighbor in the right direction situation, noted by the experience, are represented by Figure 14. The investigation of the previous cases showed in Figure 14 enabled us to propose the general formula in merging East case:

$$l_2 \geq l_1 + dD_1 + lD_1/2 \tag{3}$$

In the slanting direction, the merging of the b_1 block will take place with the b_2 block (Figure 15). While taking block b_3 as a reference, we get the two conditions of merging respectively in the two directions East and North :

$$l_2 + dGB_2 > l_1 + dGB_1 + dDH_1 \quad \text{and} \quad l_2 \geq l_1 + dDH_1 + lDH_1/2 \tag{4}$$

and

$$h_2 + dBG_2 > h_1 + dBG_1 + dHD_1 \quad \text{and} \quad (h_2 \geq h_1 + dHD_1 + hHD_1/2$$

where the coefficient 1/2 comes from the fact that a block can not have a size change superior to 50% from a form to another. With the same reasoning, we can deduct the rules of merging of the b_1 block in the other directions. The determination of the fragmentation of a block b_1 of the form F_1 in several blocks in the form F_2, comes back to the detection of a merging in the opposite sense. Therefore the processing of the phenomenon «merging-fragmentation» comes back to applying the method of merging detection described before successively in both directions (of F_1 toward F_2 then, of F_2 toward F_1).
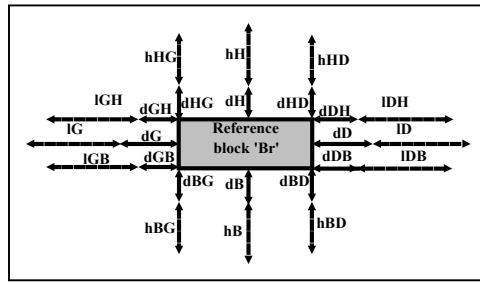
Fig. 11. Vector attributes
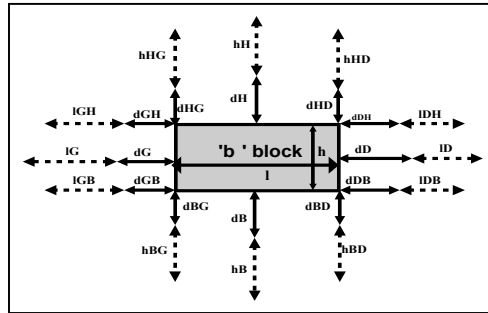
Fig. 12. Block attributes in F_i Form
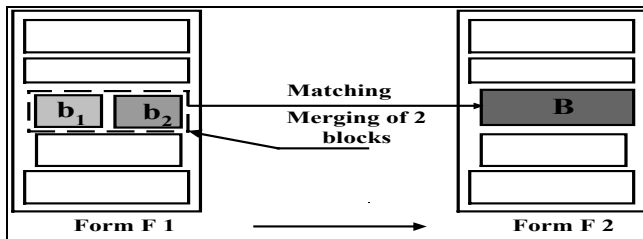
Fig. 13. Merging on the right
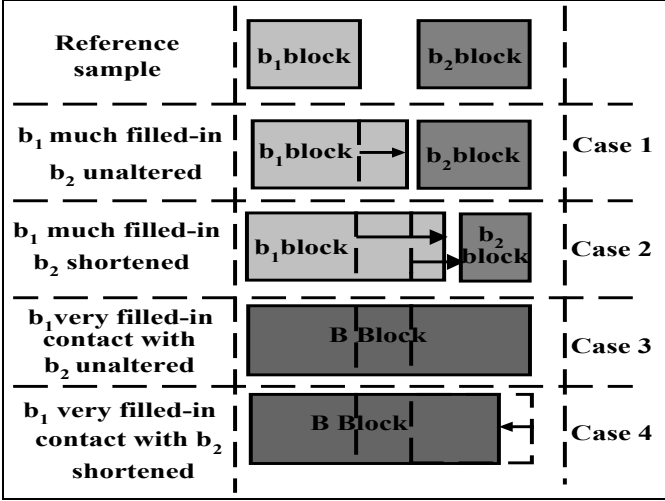
Fig. 14. Principal elongation's of the block

4.2   Form Models Construction

From N forms belonging to the same class, each filled-in by a different writer without constraint, we construct a statistical model. Each of these forms is described by a set of descending blocks from the automatic form segmentation. The number of these blocks is not necessarily identical from a form to another because of the above mentioned problems concerning merging and fragmentation of blocks. One can notice that this model is not the reunion of all the configurations, but every block having appeared in a learning sample, at least, appears in the model. A block of the model will be characterized by a stability coefficient and a vector constructed by using statistical geometric quantities (average and standard deviation). Figure 16, summarizes the learning phase applied to a reduced set of 4 samples in order to form the model of one class.
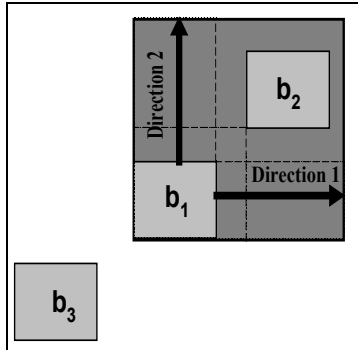
Fig. 15. Merging of the b_1 block with the b_2 to the Northeast
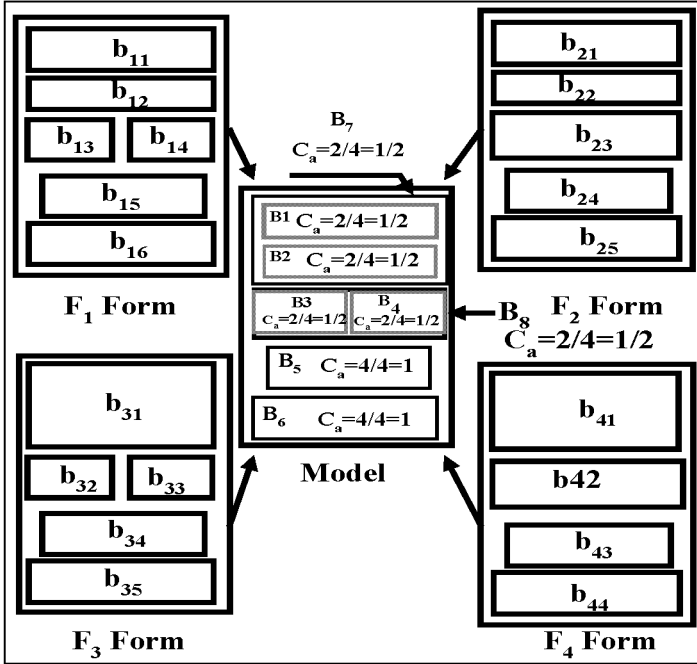
Fig.16. Example of the training done on 4 forms

The correspondence between blocks of different forms of a given class is established according to the following criteria: the Euclidean distance between centers of two matched blocks must be minimal and have the same behavior with its neighbors during merging or fragmentation, if needed. A coefficient of apparition of every block for one same class is calculated:

$$C_a = \frac{N_a}{N_t} \tag{5}$$

where N_a is the number of samples in the class containing the block and N_t is the total number of samples of the class under test. To each block that participates to form the model, we associate a feature vector V_c : $V_c = \left\{ x_m, y_m, l_m, h_m, \sigma_x, \sigma_y, \sigma_l, \sigma_h \right\}$

where $x_m = \dfrac{1}{N_a} \sum_{i=1}^{N_a} x_i$, $\sigma_x^2 = \dfrac{1}{N_a} \sum_{i=1}^{N_a} (x_i - x_m)^2$, the expressions of the other elements can easily deduced, x_m and y_m (resp. l_m and h_m)represent the mean value of the coordinates of the gravity center (resp. the lengths of horizontal and vertical sides) of blocks matched, σ_x and σ_y (resp. σ_l and σ_h) are the standard deviations of the coordinates of the gravity center (resp. the lengths of horizontal and vertical sides) of blocks matched.

5   Identification

To be able to calculate the distance, one must only keep models that have a block number superior or equal to the one of the unknown form.

5.1   Blocks Matching

Each block of the unknown form must be matched to a block of the selected model, by a bijective way. This matching is done while minimizing Euclidean distance between the geometric size vector (position and size) of the unknown form and the average geometric sizes of the model.

$$d\left(b_i, b_j^k\right) = \min_j \left[\left(x_i - x_{mj}\right)^2 + \left(y_i - y_{mj}\right)^2 + \left(l_i - l_{mj}\right)^2 + \left(h_i - h_{mj}\right)^2\right] \qquad (6)$$

where b_i is the block of i-th label of the unknown form ; b_j^k is the block of label j of k-th model.

5.2 *Mahalanobis Distance*

When the model presents a single configuration, the description vector always has the same dimension N. The probability to getting jointly the unknown form F and the M_k model is:

$$\Pr ob(F, M_k) = \Pr ob(M_k) \cdot \Pr ob(F / M_k) \qquad (7)$$

While supposing that feature follow Gaussian law, the expression of the probability of getting the unknown form F knowing the M_k model is given as:

$$\Pr ob(F / M_k) = \frac{1}{[2\pi]^{N/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}\right)\left[\left(X - X_k\right)^t \sum_k^{-1} \left(X - X_k\right)\right] \qquad (8)$$

where X is the features vector of the unknown form F ; X_k is average feature vector of the model k and Σ_k is covariance matrices of the k model (N*N dimension). The distance of Mahalanobis only keeps the part of the formula that joins the unknown shape to the model :

$$dist(F, M_k) = \left[\left(X - X_k\right)^t \sum_k^{-1} \left(X - X_k\right)\right]^{1/2} \qquad (9)$$

When there is statistical independence of the variables intra-block, the covariance matrix Σ_i becomes diagonal because of the statistical independence of variables and

the expression of the distance becomes:

$$d(F, C) = \left[\sum_{i=1}^{N} \left(\frac{(x_i - x_{mi})^2}{\sigma_{xi}^2} + \frac{(y_i - y_{mi})^2}{\sigma_{yi}^2} + \frac{(l_i - l_{mi})^2}{\sigma_{li}^2} + \frac{(h_i - h_{mi})^2}{\sigma_{hi}^2}\right)\right]^{1/2} \qquad (10)$$

5.3   Weighted Statistical Distance

The expression of the Mahalanobis distance contains a sum of the relative terms of each block, due to the inter-block independence. Using the probability to get a configuration C, knowing the model M_k, increases the distance. Since blocks are treated simultaneously on independent way, in the expression of the distance and in the construction of models, it appears more natural to ponder every term of the distance by a coefficient that varies in inverse sense of the stability coefficient (i.e. of the probability of apparition of the block). The expression of the weighted statistical distance becomes:

5.4   *Decision: Affectation to a Class or Rejection*

The decision of affectation to a class is taken according to a double criteria: i) the distance to the model representing the class must be as small as possible. ii) this distance must be small enough to avoid rejection.

6   Experimental Results and Conclusion

The learning basis is made up of 50 classes. Each contains twenty forms which are filled by different writers. The recognition is tested using 4 new elements for each class. In addition, 4 unlearned class samples are used. All the elements of the 4 unlearned classes were rejected. Figure 17 illustrates samples for 4 different classes. A recognition rate of 97% was obtained.

References

1. D. Doermann, A. Rosenfeld, E, Rivlin : The Function of documents, *Proc. of ICDAR* '97, Ulm, Germany (1997) 1077-1081.
2.   J. Mao, M. Abayan, K. Mohiuddin: A Model-Based Form Processing Sub-System, *Proc. of ICPR '96*, Vienna, Austria (1996) 691-695.
3. L. Y .Tseng, R.C. Chen: The Recognition of Form Documents Based on Three Types of Line Segments, *Proc. of ICDAR* '97, Ulm, Germany (1997) 71-75.
4. Y. Ishitani, "Model Matching Based on Association Graph for Form Image Understanding, Proc. o*f ICDAR '95*, Montreal, Canada (1995) 287-292.
5. C.D. Yan, Y.Y Tang, C.Y. suen: Form Understanding System Based on Form Description Language. *Proc. of ICDAR '91*,Saint Malo, France (1991) 283-293
6. J. Yuan, Y. Y. Tang, C. Y. Suen: Four Directional Adjacency Graphs (FDAG) and Their Application in Locating Field in Forms. *Proc o*f *ICDAR'95*, Montreal, Canada (1995) 752-755
7. F. Cesarini, M. Gori, S. Marinai, G. Soda: A System for Data Extraction from Forms of Known Class. Proc. of ICDAR'95, Montreal, Canada (1995) 1136-1140

8. U. Bohnacker, J. Schacht, T. Yücel: Matching form lines Based on a Heuristic Search ", *Proc. of ICDAR* '97, Ulm, Germany, (1997) 86-90.
9. F. Dubiel, A. Dengel.:FormClass-A System For OCR Free identification Of Forms. DAS'96, USA (1996) 189-208
10. P. Héroux, S. Diana, A. Ribert, E. Trupin:Etude de Méthodes de Classification pour l'Identification Automatique de Classes de Formulaires. *Proc. of CIFED'98*, Quebec, Canada (1998) 463-472
11. S. Kebairi, B. Taconet, A. Zahour, P. Mercy: Détection Automatique du Type de Formulaire Parmi un Ensemble Appris et Extraction des Données Utiles. CIFED'98, Quebec, Canada (1998) 255-264
12. S. Kebairi, B. Taconet: A System of Automatic Reading of Forms: *Int. Conf. of Pattern Recognition and Information Analysis, PRIP'97*, Minsk Belarus , (1997) 264-270.
13. L. Boukined, B. Taconet, A. Zahour: Recherche de la Structure Physique d'un Document Imprimé par Rectangulation., *Proc. RFIA 91*, France (1991) 1027-1031
14. S.Kebairi, A. Zahour, B. Taconet, L. Boukined: Segmentation of Composite Documents Into Homogenous Blocks. *Proc. IGS'98*, Genova Italy (1997) 111-112
15. J.F. Allen: Maintaing Knowledge About Temporel Intervals. Communication of the ACM, 26 (11), (1983) 832-843
16. H. Walischewski: Automatic Knowledge Acquisition for Spatial Document Interpretation. *Proc. of ICDAR'97*, Ulm, Germany (1997) 243-247

a)

b)

c)

d)

Fig. 17. Four segmented samples for 4 different classes. The c) sample is very close to the d) sample