Robust Approach of Address Block Localization in Business Mail by Graph Coloring

Djamel Gaceb, Véronique Eglin, Frank Lebourgeois, and Hubert Emptoz
University of Lyon, France

Abstract: *An efficient mail sorting system is mainly based on an accurate optical recognition of the addresses on the envelopes. However, the localizing of the address block should be done before the OCR recognition process. The location step is very crucial as it has a great impact on the global performance of the system. Actually, a good localizing step leads to a better recognition rate. The limit of current methods depends on modular linear architectures used for address block localization. Their performances depend on each independent module performance. We are presenting in this paper a new approach for ABL based on the hierarchical graph coloring and on the pyramidal data organization. This new approach presents the advantage to guarantee a good coherence between different modules and that reduces both the computation time and the rejection rate. The proposed method gives a very satisfying rate of 98% of good locations on a set of 750 envelope images.*

Keywords: *Text localization, physical segmentation, real time processing, business documents processing, graph coloring.*

Received August 17, 2007; accepted December 13, 2007

1. Introduction

Automatic mail sorting machines of most recent systems process about 17 mail pieces per second that requires a fast and precise block address OCR based recognition. This recognition is mainly conditioned by a correct address line organization. The Address Block Localization (ABL) is a non trivial operation due to the very large variability of characteristics of this image region and to the significant number of parasitic informative blocks. Once the envelope image has been acquired by a linear CCD camera, three principal modules contribute to the task of the ABL, as shown in Figure 1:

- Envelope image segmentation.
- Envelope layout analysis.
- Blocks interpretation.

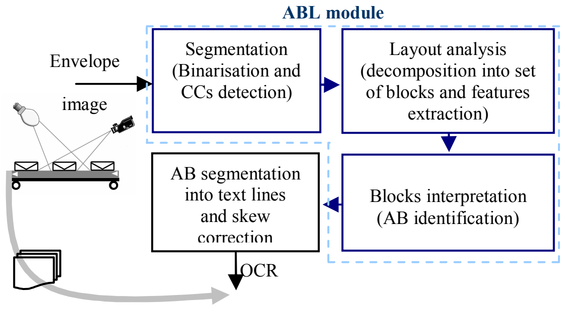

ABL module
Envelope
image
Segmentation (Binarisation and CCs detection)
Layout analysis (decomposition into set of blocks and features extraction)
Blocks interpretation (AB identification)
AB segmentation into text lines and skew correction
OCR

Figure 1. Principal ABL modules.

Following a binarizing step of the envelope image, a first module detects the Connected Components (CCs).

The second module carries out the hierarchical analysis of the layout of these CCs to recompose the blocks and establish their descriptions. Lastly, a decisional phase inspects the whole of the data obtained to recognize the address block. Practically, a dysfunction of one of these modules reduces the performances of the others, and consequently leads to a bad localization of Address Block (AB) and so to a false optical character recognition of their contents. It is obvious that AB represents the zone of interest containing necessary information to recognize the destination. Consequently, any badly localized address (i.e., badly recognized) leads to the immediate rejection of the mail piece. It should be finally noted, that the destination address is not systematically written on the bottom left corner (as it is the case for most of the mails of our study): some page-settings do not respect this strict layout as shown in Figure 2 and must be taken into account.

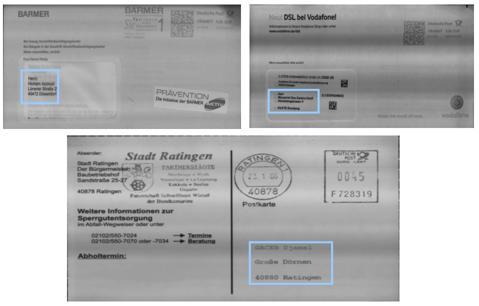


Figure 2. Presence of parasitic information near the address block.

Moreover, the presence of stamps, post office marks, printed logos, various advertisements and other parasitic information on the mail makes the task of localization more difficult, as shown in Figure 2. Other constraints particularly related to our industrial application can also be pointed out, namely:

- Very large mail variety (quality, colour and different paper textures).
- Real time constraints (limited processing time).
- Result's obligation (in a very competitive market the system must be the most powerful possible to avoid expensive manual interventions).
- High spatial resolutions of the images (200~300 dpi).

Although the performances of ABL systems do not stop growing, some limits are always presented and in almost of time they depend on a linear modular architecture limited by the above mentioned constraints. This is the global framework in which our study situates.

Taking into account all of these limits, we propose in this paper an original and robust ABL architecture resulting from multi-resolution representations and based on classification modules related to graph theory. The high level stages are based on the Hierarchical Graphs Colouring (HGC), allowing managing through a pyramidal data organization, the compound rules governing the interpretation of the decomposition into connected components of interest zones. To date, no other work in this field has made the use of the powerfulness of this tool. In opposite to traditional methods that use linear architectures, our strategy consists in increasing the performances of each module and its coherence with the other ones in order to reduce mail rejection and time processing to the maximum. The remainder of this paper is organized as follows, the various existing ABL methods are quoted in section 2 in which the previous works and the set limits are presented. In the third section, the formal aspects of graphs coloring are detailed. The fourth section describes the application of the coloring to the ABL problem. The results of ABL are then commented and discussed.

2. Existing AB Localization Systems

A significant number of works have been devoted in the recent years to the improvement of the ABL. All suggested methods can be divided into two great categories: the methods that select the AB among several candidate blocks and those that directly extract the AB starting from the image of the envelope. The methods of the first category use various segmentation techniques to let emerge perceptible blocks of the envelope image. The principle consists in extracting descriptors for each block, in order to identify the one which explicitly contains the destination address. The methods of the second category are limited to a direct AB localization without segmenting the envelope image into several blocks. In spite of a processing speed that is slightly higher than that of the first category methods, the rate of bad localization remains higher. This is why we were interested, in our study, by the first category methods.

In 1988, an expert system was proposed by Wang and al [20] to sort mails automatically. The authors use a blackboard to preserve and exploit the geometrical features of the blocks obtained during the data processing stage of various types of envelope images. A few years after, Viard-Gaudin and Barbara proposed in [19] a new approach of ABL based on a pyramidal data structure construction, in which, a downward analysis was used to extract the spatial relationships between the different segmented blocks with their features on each level of the pyramid. Yu [21] adapted an almost similar principle to complex mails. The approach suggested by Jeong [15] is based on the grouping of the connected components resulting from the binary image, where each group is assigned to one of the nine classes: the destination address block is given by selecting only some classes. Eiterer [5], recently and without segmenting the image, proposed a new track through an approach based on fractal dimensions. A classification by the K-Means method is used to label the pixels in grey-levels as background, noise or semantic objects which constitute the basic classes defining the stamps, the postmarks, and the address blocks.

We present hereafter the various existing techniques used at each stage of the ABL: binarisation, CCs detection and physical layout segmentation.

2.1. Binarisation and CCs Detection

The binarisation (or thresholding) is applied in the first stage of the ABL process and has a very strong impact on the performances of the sorting system. The thresholding methods are in general divided into two categories: Global (e.g., Otsu's method [11]), and local (e.g., Sauvola's method [16]). Obviously global techniques can not produce satisfactory results when the grey-levels input image has non-uniform shading or multi modal histogram. Local algorithms usually involve more computation and therefore are slower when running on a single-processor computer. For more details, a comparison of several binarisation techniques is presented in [9]. After the binarisation stage, an analysis of CCs is carried out to extract various vital information for the incoming stages. Formally, a connected component is a set of foreground pixels immediately adjacent to each other. Typically, in a machine printed ABL under ideal digitizing conditions, each alphanumerical character is a separate CC. In order to reduce the processing time necessary to the CCs detection, several methods were developed. A good state of the art is presented in [14].

In our study we concerned ourselves with the work of Pavlidis [12] which modelled the problem of CCs detection by a Line Adjacency Graph (LAG). This method is based on run-length representation and is quite efficient when implemented in software form. To increase its speed further, the algorithm was modified to generate components directly from the bit-packed image.

2.2. Physical Layout Segmentation

Generally, the physical layout segmentation of the envelope image indicates its decomposition into constitutive elements containing homogeneous data. These elements are often spaced and form elementary geometrical blocks, based on a rectangle in the large majority of the cases. To obtain this segmentation, one proceeds either by recursive splitting starting from spaces, by recursive merging of the objects progressively, or still by the combination of both [10]. The CCs merging segmentation methods (e.g., progressive regrouping of CCs, RLSA, segmentation by scaling method of cumulated gradients) are more used by the bottom-up strategies [4, 3] whereas the methods of segmentation by splitting (e.g., profile projection, segmentation by spaces analysis, Hough's transform) are adapted to the top-down strategies [18]. Other methods, known as hybrid, benefit from the two strategies at the same time [18].

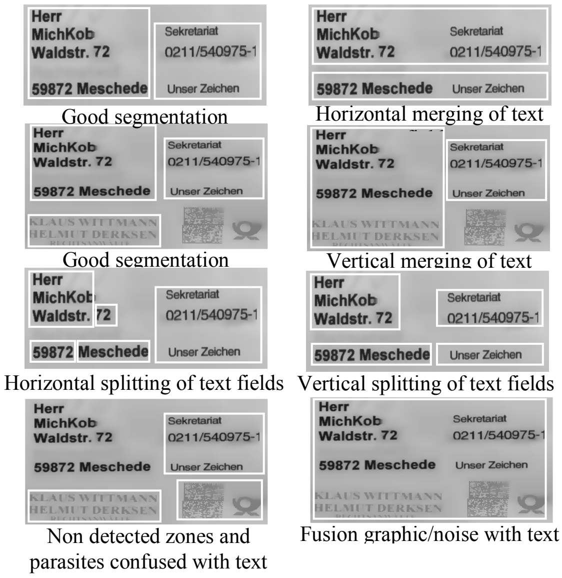
Déforges and Barba [3] presented a bottom-up generic method based on a multi-resolution description of the document image used for ABL. An almost similar structure was used by Wang [17] to distinguish the text blocks from graphic blocks, and to represent them in a structural model. Shi and Govindaraju [18] proposed an algorithm based on the application of fuzzy directional run-length.

2.3. Various Errors of Physical Segmentation

It is evident that the blocks resulting from the traditional segmentation methods can contain parasitic elements. Generally, this segmentation encounters several types of errors as shown in [1] and in Figure 3 strongly related to the bad application of the segmentation techniques in front of the presence of parasitic objects near the address block (noise, small table's fragments, logos, publicity text, or other markings or graphics). These errors can be summarized in the following points:

- Horizontal or vertical merging of text blocks or lines.
- Horizontal or vertical splitting of text blocks or lines.
- Text fusion or confusion with graphics or noise.
- Bad detection of textual blocks or lines.

The physical segmentation and ABL methods mentioned above, all use complex data structures. The management of the criteria and knowledge becomes more difficult to control, given the great variability on the envelopes to be sorted. For more robustness, we focused our work on the pyramidal data representation by introducing graph coloring.


Good segmentation

Horizontal merging of text

Good segmentation

Vertical merging of text

Horizontal splitting of text fields

Vertical splitting of text fields

Non detected zones and parasites confused with text

Fusion graphic/noise with text

Figure 3. Various errors of the physical layout segmentation.

3. Formal Aspects of the Graph Coloring Problem

Various practical classification problems can be modelled by the graph coloring. The general form of these applications requires the formation of a graph by the nodes (vertex) which represent the objects of interest and the edges (arcs) which define the relations between these objects.

One wants for example to break up a set of items into several homogeneous classes without knowing their a priori number. In doing so, it is sufficient to represent each item *i* by a node *Vi* and to add an edge $E(v_i, v_j)$ between each pair of sufficiently different individuals. The finite graph $G= (V, E)$ is defined by the finite set $V= \{v_1,v_2...,v_n\}$ $(|V|=n)$ whose elements are called nodes, and by the finite set $E=\{e_1,e_2...,e_m\}$ $(|E|=m)$ whose elements are called edges.

The coloring of the nodes of the graph $G(V,E)$ consists in assigning to all nodes a color so that two adjacent nodes do not carry the same color. These colors will correspond to the various classes of items. A coloring with *K* colors is thus a partition of the set of nodes in *k* firms. The number of colors used to colour the graph *G* of *n* nodes is called chromatic number $\chi(G) \leq n$ which represents the smallest integer *K* for which there is a partition of *V* into *K* firm subsets.

On the graph *G* of order $|V|=8$ of Figure 4, whose set of nodes is $V=\{1...,8\}$, three colors were needed to color the nodes so that two adjacent nodes can not have the same color. This number is the minimal chromatic number.

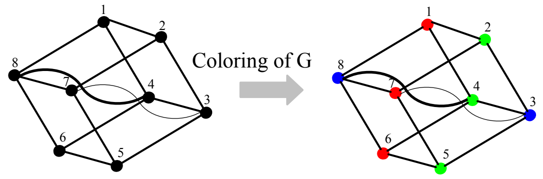


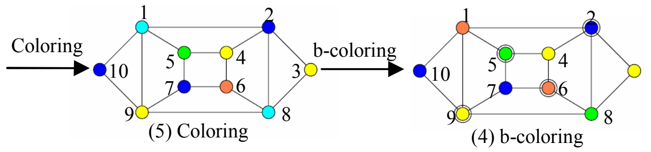
Figure 4. Coloring of graph G.

The adjacency matrix *Ma* of *G* is the symmetrical square matrix 8×8 defined as follows matrix, and a complete consultation of *Ma* takes a time:

$$\Gamma = O\left[0.5 \times \left(|v| \times |v|\right) - |v|\right]$$

$$Ma = \left|\begin{array}{ccccccccc} & x1 & x2 & x3 & x4 & x5 & x6 & x7 & x8 \\ x1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ x2 & & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ x3 & & & 0 & 1 & 1 & 0 & 1 & 0 \\ x4 & & & & 0 & 0 & 1 & 0 & 1 \\ x5 & & & & & 0 & 1 & 1 & 0 \\ x6 & & & & & & 0 & 0 & 1 \\ x7 & & & & & & & 0 & 1 \\ x8 & & & & & & & & 0 \end{array}\right|$$

The coloring is called b-coloring, if for each color C_i, there exists at least a colored V_i node C_i whose neighbourhood is colored by all the other colors. The node V_i is known as a dominating node for color C_i. The example of Figure 5 presents the possibility of b-coloring of the nodes of a colour class using the other colors.



(5) Coloring

(4) b-coloring

Figure 5. b-Coloring example, the nodes 2, 5, 6 and 9 are the dominating nodes.

The b-chromatics number of a graph *G*, defined by *b(G)*, is the maximum integer number of colors *K* so that *G* can have a b-coloring by the *K* colors. It can be easily noticed that: $\chi(G) \leq b(G) \leq \Delta(G)+1$ where $\Delta(G)$ is the maximum degree of *G*, called the degree of node *Vi*, and its number of incidental edges is noted *d(vi)*. The majority of the evaluations of $\chi(G)$ and *b(G)* come from coloring algorithms. There exist many of them, and so that not to drown ourselves with this question, we will limit ourselves to quoting the fastest and most recent ones.

New graph coloring and b-coloring algorithms have been proposed by Effantin and Kheddouci [6, 8]. More details on the approximation of the b-chromatic number were proposed by Corteel [2]. All of these algorithms were efficiently introduced into Elghazel's works [7] who proposed a new unsupervised classification method of medical data based on graph b-coloring where the number of the classes is not known a priori. On the same database the comparison between the accuracy of this method and that of the method of agglomerative technique offers a true representation of the classes by the dominant individuals and guarantees a better interclass disparity.

4. Application of Graph Coloring to Our Problem

The ABL strongly depends on the parasitic object's density near the address block. The knowledge delivered by the features of the blocks description, resulting from the physical segmentation stage, is not efficient to discriminate heterogeneous blocks (containing parasitic elements). To efficiently locate AB on envelopes known as difficult, it has been necessary to choose an even more advanced tool of progressive regrouping of CCs and of AB identification. Therefore, we have taken the powerfulness of the graph coloring into account to automatically separate the elements into homogeneous groups as shown in Figure 6 and in that sense to considerably improve the ABL system.

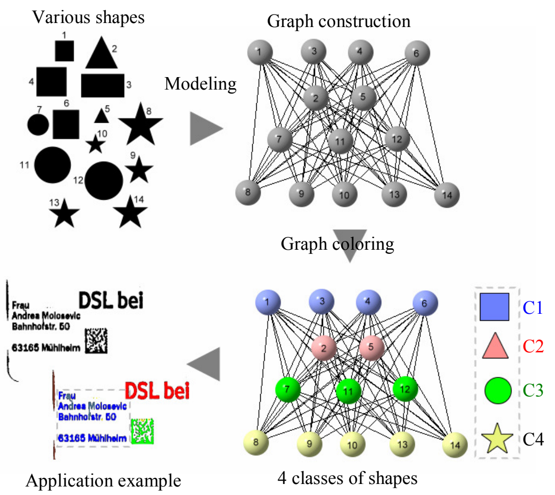


Figure 6. Example of the graph coloring application.

The hierarchical graph coloring is introduced, to correct the over (and/or under) segmentation of the envelope into blocks, and b-coloring is used, to train a classifier to identify block-addresses among several candidates.

We present in this section the different stages of our ABL approach. The diagram of Figure 7 represents our pyramidal architecture that allows obtaining the best possible coherence between the various modules. This architecture is based on three essential modules:

- The envelope image segmentation: coupling of the binarisation with the localization of the layout zones and detection of CCs.
- The physical layout analysis of the envelope based on the graph coloring: progressive regrouping and hierarchical analysis of CCs to recompose the blocks and establish their descriptions.
- The blocks interpretation: training of classifier by b-coloring and identification of the address block among several candidates.

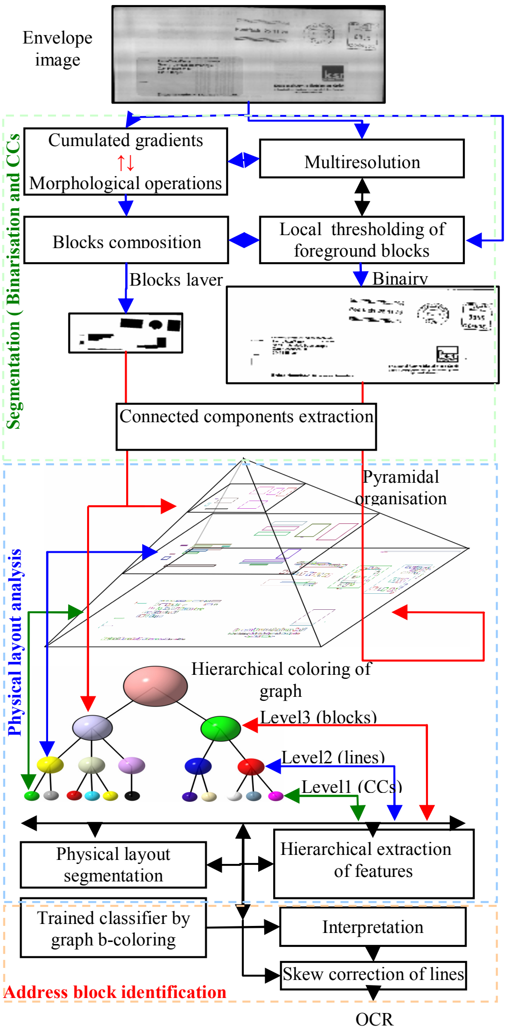

Envelope image
Segmentation (Binarisation and CCs
Cumulated gradients
Morphological operations
Multiresolution
Blocks composition
Local thresholding of foreground blocks
Blocks layer
Binairy
Connected components extraction
Physical layout analysis
Pyramidal organisation
Hierarchical coloring of graph
Level3 (blocks)
Level2 (lines)
Level1 (CCs)
Physical layout segmentation
Hierarchical extraction of features
Trained classifier by graph b-coloring
Interpretation
Skew correction of lines
Address block identification
OCR

Figure 7. Functional diagram of the proposed approach.

4.1. Thresholding and CCs Detection Methods Used

The separation between the image thresholding and foreground localization stages considerably increase the computation time and lead to an over-segmentation of the noise and of the paper texture on empty zones of the image. Indeed, none of the traditional methods (whether global or local) efficiently meets all the required conditions, in particular, a certain efficiency on all the images during a limited computing time. We have managed to optimize this stage by applying a local threshold only near the text zones as shown in Figure 8 that can be located by the cumulated gradients method with the multiresolution and mathematical morphology [9]. After the binarisation step, we detect CCs of foreground of the binary and block layers. The method used is inspired from Pavlidis's studies [12] based on the LAG (Line Adjacency Graph) structure, a structure particularly adapted to the line by line image scan. It consists in connecting the black pixels runs of two consecutive lines of a binary image. Each CC is represented by the coordinates of its bounding box with: $CC(i)=(x_i^1,y_i^1,x_i^2,y_i^2)$. Let $V(L_1)$ (or $V(L_3)$) be the set of the CCs of the binary layer (or of the blocks-layer) which represents the finest (or the coarsest) level L_1*(or* L_3) of the pyramid. With $V(L_k) =\{CC_k(i)_{/i=1...Nk}\}$, N_k it is the number of CCs in the layer $K/_{k=1,2 \text{ and } 3}$ The physical layout segmentation is then based on a hierarchical analysis on each pyramid level of the bounding boxes. Each level contains different features. This CCs, constitutes a significant information source, very often used during the description process.

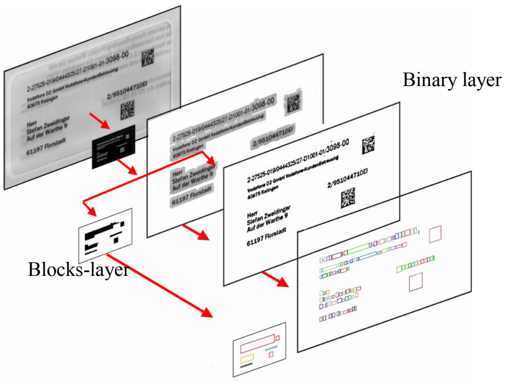


Figure 8. Our hybrid approach of binarisation (text localization/thresholding) and CCs detection.

4.2. Hierarchical Analysis and Block Description Strategy

Each block can be described by a set of the features resulting from the hierarchical analysis of the three levels of the data pyramid. To manage knowledges that are associated to them, each group of objects has features and a strategy of classification. At the bottom of the hierarchy there are the binary layer CCs. The progression in the hierarchy makes it possible, at each level, to acquire more precise knowledge on the image contents. Each set of features can be visible at various levels of perception. For example, the alignment of the text lines is not perceived on the same level as the character spacing, nor that of the blocks position on the envelope (see figure9 and table1).

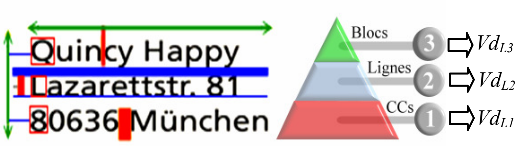


Figure 9. Hierarchical extraction of features.

Our idea consists in making the blocks description phase cooperate with the physical segmentation phase and allowing, at any level of the hierarchy, the use of all information expressed in the other levels as shown in Figure 9. In that way, the description can take advantage of two phases. Let Vd_{Lk} (i) be the descriptor-vector of block i on the level k=1, 2 and 3. The complete description of this block is given by the combination of the descriptions of three levels L_1, L_2, and L_3 as shown in equation 1.

$$Vd_{Total}(i) = Vd_{L1} \cup Vd_{L2} \cup Vd_{L3} \tag{1}$$

Table 1. Features perception at the various pyramid levels.

Features on the CCs layer Vd_{L1}	**Features on the lines-layer Vd_{L2}**	**Features on the blocks-layer Vd_{L3}**
Position CCs height CCs width Inter-character space	Position Line height Line width Inter-line space Alignment Eccentricity Overlapping degree	Position Block width Number of lines Eccentricity Spatial relations Density Uniformity

Using the Principal Component Analysis (PCA), we could select a minimal number of features while ensuring a perfect disparity between the objects of different natures.

4.3. Application of the Graph Colouring to the Physical Layout Segmentation

The segmentation techniques can not systematically produce uniform and well located blocks in complex environments. One speaks about over-segmentation when the constitutive components are fragmented and about under-segmentation when several constitutive components can not be isolated. Consequently, the presence of parasites or incomplete information in an address block can introduce errors into its description and can lead to a bad interpretation. Segmentation methods by merging and splitting can all have both advantages and disadvantages [10]. Hybrid segmentation approaches (or mixed approaches) gather both strategies in the same time (top-down and bottom-up) and can benefit from the advantages of one strategy to fill the disadvantages of the other. Our concept pf physical layout segmentation is based on the same principle of a hybrid strategy. High stages of our approach base on the Hierarchical Graph Coloring (HGC) that largely makes use of all the levels of the pyramidal structure and the coloring powerfulness, so as to extract, to characterize and precisely to group the objects of same nature as shown in Figure 10. Therefore, our idea consists in forming at each level of the hierarchy of the groups of components that must be as homogeneous as possible in order to lead to a more precise description.

Let G be a non orientated graph at three levels independent but coherent defined by the following relationship:

$$G(V, E) = \bigcup_{k=1}^{3} G_k(vd_{Lk}, E_{Lk>Sk}) \qquad (2)$$

where, $vd_{Lk} = \{vd_{Lk}\,(\mathrm{i})\}_{\mathrm{i}=1\ldots\mathrm{N_k}}$ the finite set of represented nodes starting from the descriptors of the set $V(L_k)_{k=1,2 \text{ and } 3}$ of N_k constitutive elements of the data pyramid at level k (see Figure 7), and $E_{Lk>Sk}$, the finite set of the edges represented by the pairs of adjacent nodes. Taking into account the fact that each node is represented by a features vector, two nodes are then considered as adjacent if and only if their dissemblance $d_{i,j}$ (distance between their two features vectors) is strictly greater than the threshold $S_{k.}$ (the optimization mechanism of the thresholds S_k is detailed in section 4.4.). This definition is given by the following relationship:

$$E_{Lk>Sk}\left[vd_{Lk}(i), vd_{Lk}(j)\right] = \begin{cases} 1 & \text{if} \quad d_{i,j} > s_k \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

The hierarchical coloring of graph G is used here to split-up the set of nodes of each level K into homogeneous subsets. It focuses on the dissemblance of the objects (represented by nodes) of the same level in the data pyramid to make their resemblances emerging. The colouring process uses a hybrid strategy of progression into the hierarchy of the graph: the colors of a level take part in the formation and the description of the next level nodes as shown in Figure 10.

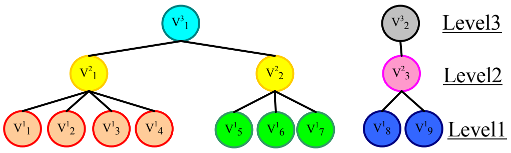


Figure 10. Hierarchical graph coloring.

The coloring steps of the graph *Gk* are given by the following algorithm [8]:

Algorithm 1: Graph_coloring (Graph)

```
Begin
 If $c(i) \neq \varnothing$ then
   Let $M := Nc(i) \cup \{c(i)\};\ q := 0;$
    For every node $j \in N(i)$ such that $c(j) := \varnothing$ do
       $q := \min\{k|k > q, k \notin M$ and $k \notin c(j)\};$
       If $q \leq \Delta + 1$ then
          $c(j) := q;$
       Else
          $c(j) := \min\{k|k \notin Nc(j)\};$
          Endif
    End do;
 End if;
End
```

where *c(i)* is the color of node *i*, *N(i)* is the set of nodes adjacent to node *i*. *Nc(i)* is the set of node colors of *N(i)*, *d(i)*=|*N(i)*| is its degree, and:

$$\Delta = max\{d(i) \mid i \in V\} \qquad (4)$$

Our method of physical layout segmentation is based at the beginning on the first stage graph construction (which is noted *G1*) which models the CCs layer V*(L1)* of the first data pyramid level. The graph *G1(Vd$_{L1}$,E$_{>S1}$)* is then colored by the algorithm1 to form *G3(Vd$_{L3}$,E$_{>S3}$)*. Then these colors are superimposed on the layer *V(L$_3$)* of the blocks to subdivide each block that contains several colors and also each color that contains several blocks. By exploiting this new knowledge, we apply once again a second coloring of graph *G3 (Vd$_{L3}$,E$_{>S3}$)* formed by

Vd_{L3} set of fragments descriptors that we merge to form a uniform blocks layer noted $V^*(L_3)$. Finally, the total description of each block is improved by a new set of features extracted from the second layer resulting from a coloring process of the graph $G2\ (Vd_{L2},E_{>S2})$ with Vd_{L2} the set of descriptors defined by the analysis of $V(L_1)$ and $V^*(L_3)$. The following algorithm summarizes all the segmentation stages:

Algorithm 2*:* Physical_segmentation()

```
Begin
  Level 1: regrouping of similar CCs
    For every CC1 (i)/i:=1…N1
      Extract VdL1(i)
    Endfor
    V (L1):= {VdL1(i)/ i=1…N1)
    G1:= G(V(L1),E(L1)>S1);
    Execute: Graph_coloring(G1)
  Level 2: homogenization of the blocks layer
    Let M: = G1∩V (L3)
    For every item i of M
      Extract VdL3 (i)
    Endfor
    G3:= G (M, E(M)>S3) ;
    Execute: Graph_coloring(G3)
  Level 3: emergence of the text lines
    Let M: = V (L1) ∩ G3
    For every item i of
      M Extract VdL2 (i)
    Endfor
    G2:= G(M, E(M)>S2);
    Execute: Graph_coloring(G2)
    V(L2)=G2, V(L3)=G3;
    Extract VdL1 := V(L1)/G2 and G3;
    Extract VdL2 := G2/ V(L1) and G3;
    Extract VdL3 := G3/ V(L1) and G2;
    For every block find
      $Vd_{Total}(i)=\{\ Vd_{L1\in VdL3}\ \}\cup\{\ Vd_{L2\in VdL3}\ \}\cup Vd_{L3}$ ;
   Endfor.
End
```

4.4. Optimisation of Dissemblance Thresholds

In a preparatory phase of auto-parameter setting, our system uses the Levine and Nazif [13] combination of intraclass and interclass disparities to automatically adjust all the threshold values necessary for the coloring process. The principle consists in choosing the thresholds that maximize the following cost function:

$$F = M_{Inter\_groups} + M_{Intra\_groups} \tag{5}$$

Where, $M_{Inter\_groupes}$ represents the sum of dissimilarities between the groups (colors) pondered by their areas:

$$M_{Inter\_groups} = \frac{\sum_{Gi} AiCi}{\sum_{Gi} Ai} \quad \text{where } Ci = \sum_{Gi} \frac{lij}{li} \frac{|mi - mj|}{|mi + mj|} \tag{6}$$

mi is the average of group G_i, l_{ij} is the length of the common border between G_i and G_j and L_i is the perimeter of group G_i. The criterion M_{Intra_groups} computes the sum of the standardized variances of the groups (cut down to 1).

4.5. Training Based on the B-Coloring

In order to prepare a representative training data set, 400 blocks of several categories (AB, stamp, logos…), resulting from the physical segmentation of a large variety envelope images have been selected. The training graph (noted $G_{Training}$) is constructed just after the description of each block of this set by a discriminating features vector (section 4.2). After the $G_{Training}$ coloring process done by the algorithme1, some resulting colors do not have any dominating node. We use then the algorithm3 [8] for the b-coloring of the non-dominating colors of $G_{Training}$.

Algorithm 3: b-coloring_Graph()

```
BEGIN
  Repeat,
    q: = max{k| k∈ NDm}; L := L\{q}; NDm:= L\Dm;
     For each vertex vi such that c (vi):=q do
       K: = {k| k∈L and k∉ Nc(vi)};
       c(vj):= {c|dist(vi,c):=mink∈K(dist(vi,k))};
     Enddo;
     For each vertex vj such that c(vj)∈NDm  do
       Update(Nc(vj));
       If Nc(vj) := L\{c(vj)} then
         Add(c(vj),Dm);
       EndIf;
     Enddo;

  Until (NDm := ϕ );
End
```

where ND_m, is the set of the non-dominating colors, D_m is the set of the dominating colors, and *C(vi)* is the colors of node *i*. Whereas a supervised classification technique requires the introduction of the number of the classes by a supervisor (knowing that an imprecision of this number can easily force the classifier to make classification errors), a non supervised technique does not present this kind of disadvantage. The b-coloring process perfectly decomposes the set of blocks into uniform subsets (colors), without knowing a priori their optimal number. Moreover, this process offers a good representation of the classes by the dominating nodes (representatives of the blocks) that ensure a great interclass disparity as shown in Figure11. These representative nodes will thus be used in the ABL phase to identify in real time the block address among several candidates.

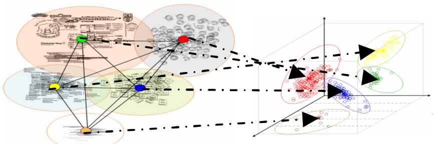


Figure 11. Features separability and training block classification by b-coloring and detections of the representatives of the classes.

4.6. Real-Time Identification of the Address Block Based on the Graph B-Coloring

To select the AB in a set of the candidate blocks $S=\{V_{i=1\dots N}\}$ we compare in the features space the description of each one with that of all representative blocks (dominating nodes) $S^*\{V^*_{j=1\dots M}\}$ resulting from the training phase. The matching algorithm determines in real time for each block of S the same designation of its adjacent nodes of S^*. This interpretation provides new knowledge on the spatial relationships between the envelope zones that are necessary to take a final decision. The dissimilarity between V_i and V^*_j is given by the generalized Minkowski distance of order α (α = 1).

$$di,j = (\sum_{k=1}^{Nf} g_k(v_i^k, v_j^{k^*})^a)\frac{1}{a} \qquad (7)$$

If $\alpha=1$ → *di,j* is the City Block distance, α=2 → *di,j* is Euclidean distance and the more α increases, the more *di,j* tends to the Chebyshev distance. *Nf* is the length of the feature vectors. g_k is the dissemblance function that compares each pair of features.

5. Experimentation

The evaluation of the performances of our approach has been achieved on a corpus of 750 envelope images (considered as difficult and noisy) and more than 98 % of good localizations have been obtained. Others tests were carried out, as well, on a set of 100 images rejected by the currently most powerful system of localization. All these rejections are being due to the failure of one of the three localization phases (Sauvola's, RLSA and KMeans methods). The following curve shows that 18 envelopes were rejected due to a bad binarisation, 53 due to a bad physical segmentation and 29 due to a bad identification of the address block. 93% of good localization was obtained using our approach.

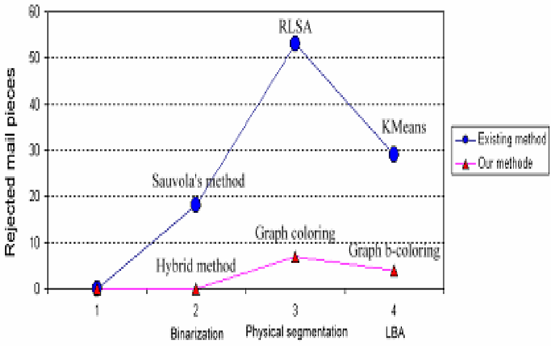


Figure 12. Performances of our method.

Thanks to its robustness to the parasitic elements, the HGC method is definitely more efficient for noisy images segmentation by comparison to classical RLSA approach as shown in Figure 13.

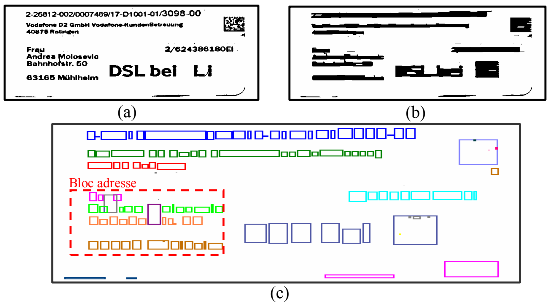


Figure 13. Noised block addresses (a), emergence of the text lines by the RLSA approach (b), perception of the lines by coloring (c).

6. Conclusion

We have presented in this paper a new approach of address-block localization based on a hierarchical graph coloring and a pyramidal data organization. We have shown that the hierarchical graph coloring give a great robustness with parasitic objects considered as principal factors of physical segmentation error, and that the b-coloring leads to a noticeable improvement of the interpretation step of candidate blocks. Moreover, we also intend to increase coherences between the various modules so as to reduce the computing times and the rejection rates.

References

[1] Agne S. and Rogger M., "Benchmarking of Document Page Segmentation," *in Proceedings of the IS&T/SPIE Conference on Document Recognition and Retrieval VII*, California, pp. 165-171, 2000.

[2] Corteel S., Valencia-Pabon M., and Vera C., "On Approximating the b-Chromatic Number," *Computer Journal of Discrete Applied Mathematics Archive*, vol. 146, no. 1, pp. 106-110, 2005.

[3] Déforges O. and Barba D., "A Fast Multiresolution Text-Line and Non Text Line Structures Extraction and Discrimination Scheme for Document Image Analysis," *in Proceedings of the International Conference on Pattern Recognition (ICPR),* pp. 134-138, 1994.

[4] Drivas D. and Amin A., "Page Segmentation and Classification Utilizing a Bottom up Approach," *in Proceedings of Document Analysis and Recognition (ICDAR)*, USA, pp. 610-614, 1995.

[5] Eiterer F., Facon J., and Menoti D., "Postal Envelope Address Block Location by Fractal Based Approach," *in Proceedings of Computer Graphics and Image Processing 17th Brazilian Symposium*, Brazil pp. 90-97, 2004.

[6] Effantin B. and Kheddouci H., "The b-Chromatic Number of Power Graphs," *Computer Journal of DMTCS'03*, vol. 6, no. 4, pp. 45-54, 2003.

[7] Elghazel H., "A New Clustering Approach for Symbolic Data: Algorithms and Application to Healthcare Data," *Computer Journal of Bases de Données Avancées (BDA)*, vol. 1, no. 4, pp. 467-510, 2006.

[8] Effantin B. and Kheddouci H., "A Distributed Algorithm for a b-Coloring of a Graph," *in Proceedings of International Symposium on Signal Processing and its Applications (ISPA'06)*, Italy, pp. 312-313, 2006.

[9] Gaceb D., "Contribution to the Automatic Recognition of Business Documents," *International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, France, 2006.

[10] Mullot R., *Les documents écrits de la Numérisation à l'indexation par le Contenu*, Hermes Science Publication, France, 2006

[11] Otsu N., "A Threshold Selection Method from Grey Level Histogram," *Computer Journal of IEEE Transactions System*, vol. 9, no. 1, pp. 62-66, 1979.

[12] Pavlidis Z. and Zhou J., "A Page Segmentation and Classification," *Computer Journal of CVGIP'92*, vol. 54, no. 6, pp. 484-496, 1992.

[13] Philipp S., Evaluation de la Segmentation, *Technique Rapport*, Louis Pasteur, 2001.

[14] Regentova E., "An Algorithm with Reduced Operations for Connected Components Detection in ITU-T Group 3/4 Coded Images," *Computer Journal of IEEE Transactions System*, vol. 24, no. 8, pp. 1039-1047, 2002.

[15] Seon J., Seung J., and Yun N., "Locating Destination Address Block in Korean Mail Images," *in Proceedings of International Conference on Pattern Recognition (ICPR*), USA pp. 387-390, 2004.

[16] Sauvola J., Seppänen T., Haapakoski S., and Pietikäinen M., "Adaptive Document Binarization," *in Proceedings of Document Analysis and Recognition (ICDAR)*, Scotland, pp. 147-152, 1997.

[17] Shin W. and Yagasaki T., "Block Selection: A Method for Segmenting a Page Image of Various Editing Styles," *in Proceedings of Document Analysis and Recognition (ICDAR)*, Canada, pp. 128-133, 1995. .

[18] Shi Z. and Govindaraju V., "Line Separation for Complex Document Images Using Fuzzy Runlength," *in Proceedings of DIAL04 First International Workshop*, California, pp. 306-312, 2004.

[19] Viard C. and Barba D., "A Multi Resolution Approach to Extract the Address Block on Flat Mail Pieces," *in Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP-91)*, pp. 2701-2704, 1991.

[20] Wang C., Palumbo W., and Srihari N., "Object Recognition in Visually Complex Environments: An Architecture for Locating Address Blocks on Mail Pieces," *in Proceedings of the 9th International Conference on Pattern Recognition*, USA, pp. 365-367, 1988.

[21] Yu B., Jain K., and Mohiuddin M., "Address Block Location on Complex Mail Pieces," *Document Analysis and Recognition 4th International Conference, pp. 897-901, 1997.*

[22] Malkawi M., Al-Haj Hassan M., and Al-Haj Hassan O., "A New Exam Scheduling Algorithm Using Graph Coloring," *in International Arab Journal of Information Technology*, vol. 5, no. 1, pp. 80-86, 2008.

Djamel Gaceb received the engineer degree in electronics from BLIDA University, Algeria, in 2002, and the Master degree in computer science from University Claude Bernard of Lyon France, in 2005. He is currently working on his PhD thesis on the topic of postal mail sorting and of document recognition, retrieval and analysis at the National Institute of Applied Sciences (INSA), in Lyon, and more precisely in the LIRIS laboratory.

Véronique Eglin is graduated from the INSA of Lyon in 1995 and holder in 1998 of the PhD in computing science on the document structure analysis. She is working since September 2000 as associate professor in the INSA of Lyon and is working since 2003 in the LIRIS UMR 5205 laboratory.

Frank Lebourgeois is graduated in 1987 from University of Lyon I with a Master of science in mathematics, then he gets a PhD in 1992 at INSA de Lyon in computer sciences. He is currently an assistant professor in the LIRIS laboratory and works on documents images restoration and analysis.

Hubert Emptoz is professor in the Institute INSA of Lyon since 1990 and has been director of the laboratory Reconnaissance de formes et Vision from 1996 to 2003. Since 2003, he is attached to the LIRIS UMR 5205 laboratory at the same Institute. He is currently coordinator of different projects of digitization of ancient documents corpus and digital libraries dedicated to forensic collections. He supervises four PhD in that domain.