IJDAR (2008) 11:67–80
DOI 10.1007/s10032-008-0070-8

ORIGINAL PAPER

Improvement of postal mail sorting system

Djamel Gaceb · Véronique Eglin · Frank Lebourgeois · Hubert Emptoz

Received: 25 February 2008 / Revised: 23 June 2008 / Accepted: 19 August 2008 / Published online: 9 September 2008

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 (ABL) 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. Consequently a good localizing step leads to a better recognition rate. The limits of current methods are mainly caused by modular linear architectures used for ABL and the lack of cooperation between modules: their performances greatly depend on each independent module performance. We are presenting in this paper a new approach for ABL based on a pyramidal data organization and on a hierarchical graph coloring for classification process. This new approach presents the advantage to guarantee a good coherence between different modules and it also 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 location · Physical segmentation · Real time processing · Business documents processing · Graph coloring

D. Gaceb (✉) · V. Eglin · F. Lebourgeois · H. Emptoz
LIRIS INSA de Lyon, 20 av. Albert Einstein,
69621 Villeurbanne Cedex, France
e-mail: djamel.gaceb1@insa-lyon.fr

V. Eglin
e-mail: veronique.eglin@insa-lyon.fr

F. Lebourgeois
e-mail: flebourg@insa-lyon.fr

H. Emptoz
e-mail: hubert.emptoz@insa-lyon.fr

1 Introduction

Automatic mail sorting machines of most recent systems process about 17 mail pieces per second that requires a fast and precise address block OCR based recognition. This recognition is mainly conditioned by a correct address line organization. The address block localization (noted ABL) is a nontrivial 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 (see Fig. 1): envelope image segmentation, envelope layout analysis, blocks interpretation and recognition.

After the binarizing step of the envelope image, a first module detects the connected components (CCs) including all textual and graphical objects of the page. 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 all extracted components 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 location 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 (see Fig. 2) and must be taken into account.

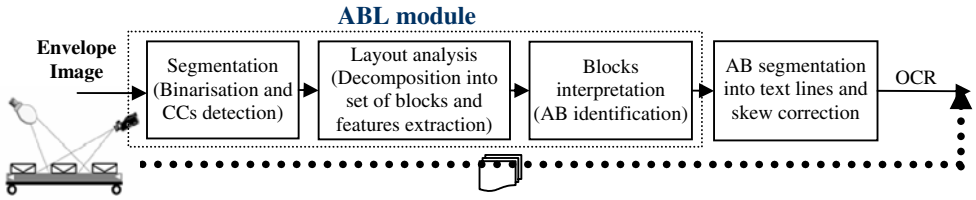

Fig. 1 Principal ABL modules

Fig. 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 (Fig. 2). Other constraints particularly related to our industrial application can also be pointed out, namely:

- A very large mail variety (size, quality, color and different paper textures),
- A result's obligation (in a very competitive market, the system must be the most powerful as possible to avoid expensive manual interventions),
- Real time constraints (limited processing time) and high spatial resolution of the images (200–300 dpi).

Although the performances of ABL systems do not stop growing, some limits are always presented, depending most of time on a linear modular architecture limited by the above mentioned constraints. This is the global framework of the purpose of this paper.

By taking into account all of these limits, we propose in this paper an original and robust ABL architecture that results from multi-resolution representations and that is based on classification modules related to the Graph theory. The high level stages are based on the hierarchical graphs coloring (which are noted HGC), allowing to manage, through a pyramidal data organization, the compound rules governing the interpretation of the decomposition into connected components of interest zones.

This manuscript describes a new method for region-of-interest-finding for mailpieces using a hierarchic multi-resolution images decomposition and an adapted hierarchical graph coloring method. This overall proposed methodology is used for improving the connected components extraction and the data organization description. The main contribution based on graph coloring is dedicated to the classification of homogeneous regions like address blocks, logos, stamps, etc. It considerably improves the segmentation steps by reducing time consuming and by locating with a great precision all objects of mailpieces address blocks. We also demonstrate by different experiments the effectiveness of the proposed hierarchical multi-resolution structure in combination with the applied graph-coloring method in regard to other standard approaches. 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 date, no other work in this field has made the use of the powerfulness of this tool.

The paper is organized as follows: various existing ABL methods are quoted in Sect. 2. We point out all previous works of the domain and their limits. 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. Experiments and results of ABL are then commented and discussed. We also demonstrate in this last part the effectiveness of our proposition with a detailed description and evaluation

of the performance of each individual module of our pyramidal architecture. In that way, other well-known standard approaches are used to quantify the real performances of our system.

2 Existing address-block 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 informative blocks of the envelope image. The principle consists in extracting descriptors for each block, in order to identify the one that 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 et al. [1] to sort mail 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 Barba proposed in [2] 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 [3] adapted an almost similar principle to complex mails. The approach suggested by Jeong [4] 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. Recently, Eiterer [5] proposed a new track through an approach based on fractal dimensions, without segmenting the image. 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 here the various existing techniques used at each stage of the ABL: binarization, connected components (CCs) detection and physical layout segmentation.

2.1 Binarization and connected components detection

The binarization (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 [6]), and local (e.g., Sauvola's method [7]). 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 binarization techniques is presented in [8].

After the binarization 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 [9]. In our study we has been interested by Pavlidis's work in [10] who has modeled 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, the algorithm was modified to generate components directly from the bit-packed image.

2.2 Physical layout segmentation

The physical layout segmentation of the envelope image is mostly based on a decomposition into constitutive elements containing homogeneous data. These elements are often spaced and form elementary geometrical blocks, based on rectangle bounding boxes. To obtain this segmentation, one proceeds either by a recursive splitting starting from spaces, by a progressive recursive merging of objects, or still by the combination of both [11]. The CCs merging segmentation methods (progressive regrouping of CCs, RLSA, segmentation by scaling method of cumulated gradients) are more used by the bottom-up strategies [12,13], whereas the methods of segmentation by splitting (profile projection, segmentation by spaces analysis, Hough's transform) are adapted to the top-down strategies [14]. Other methods, known as hybrid take advantage of the two strategies at the same time [14].

Déforges and Barba [12] 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 [15] to distinguish the text blocks from graphic blocks, and to represent them in a structural model. Shi and Govindaraju [14] proposed an algorithm based on the application of "fuzzy directional run-length."

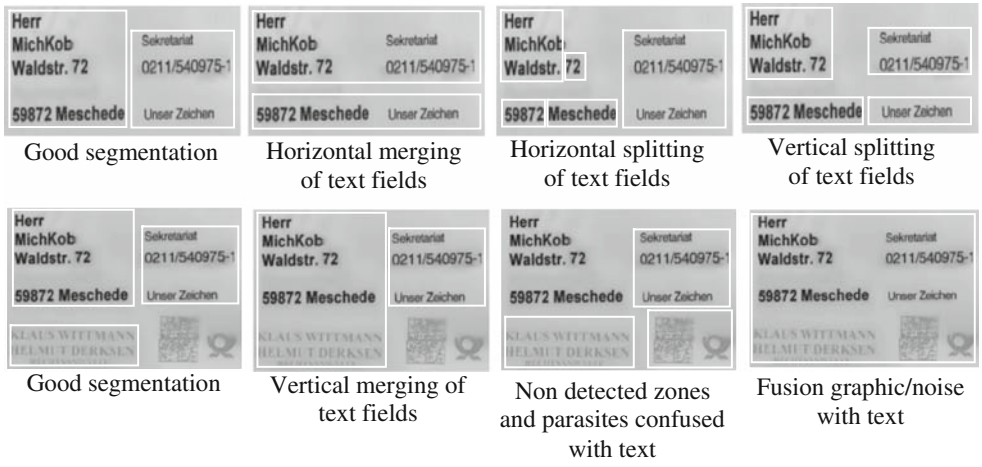

Fig. 3 Various errors of the physical layout segmentation

Those works present a great interest in term of genericity and adaptability to a great diversity of document image. In our work, we have endeavored to work with the same spirit.

2.3 Various errors of the physical segmentation

It is obvious that the blocks resulting from the traditional segmentation methods can contain parasitic elements. Generally, this segmentation encounters several types of errors (described in [16] and in Fig. 3) strongly related to the bad application of the segmentation techniques near the address block (noise, small tables 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 mostly use complex data structures. So the management of the criteria and knowledge becomes more difficult to control the great variability of the envelopes to be sorted. For more robustness, we have focused our work on a pyramidal data representation by introducing graph coloring.

3 Formal aspects of the graph coloring

Various practical classification problems can be modeled 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. To do that, it is sufficient to represent each item i by a node v_i and to add an edge $E(v_i, v_j)$ between each pair of different individuals. The finite graph $G = (V, E)$ is defined by the finite set $V = \{v_1, v_2, \ldots, v_n\}$, $(|V| = n)$ whose elements are called nodes, and by the finite set $E = \{e_1, e_2, \ldots, e_m\}(|E| = m)$ whose elements are called edges.

3.1 Graph coloring

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 V in k homogeneous subsets. The number of colors used to color the graph G of n nodes is called chromatic number χ which represents the smallest integer k for which there is a partition of V into k homogeneous subsets.

On the graph G of order $|V| = 8$ of Fig. 4, whose set of nodes is $V = \{1, \ldots, 8\}$, three colors were needed to color the nodes so that two adjacent nodes can not have the same color. $\chi(G) = 3$ is the minimal chromatic number.

3.2 Graph b-coloring

The adjacency matrix M_a of G is the symmetrical square matrix 8×8 defined as follows:

$$M_a = \begin{array}{c|cccccccc} & x1 & x2 & x3 & x4 & x5 & x6 & x7 & x8 \\ x1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ x2 & & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ x3 & & & 0 & 1 & 1 & 0 & 1 & 0 \\ x4 & & & & 0 & 0 & 1 & 0 & 1 \\ x5 & & & & & 0 & 1 & 1 & 0 \\ x6 & & & & & & 0 & 0 & 1 \\ x7 & & & & & & & 0 & 1 \\ x8 & & & & & & & & 0 \end{array}$$

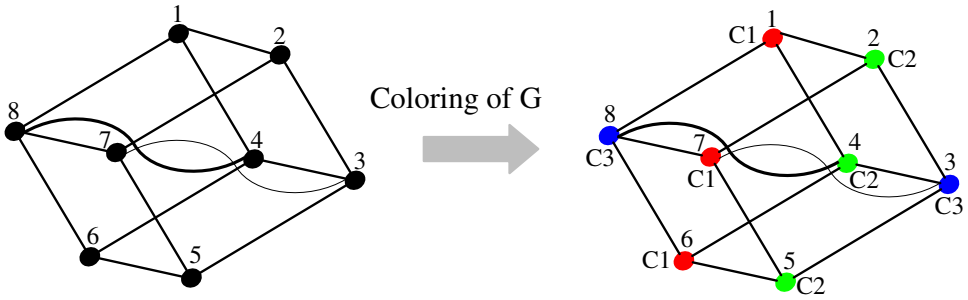

Fig. 4 Coloring of graph G in three colors

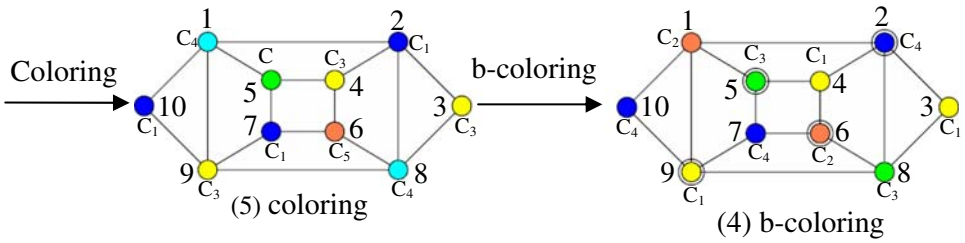

Fig. 5 b-Coloring example, the nodes 2, 5, 6 and 9 are the dominating nodes

A complete consultation of M_a takes a time: $\Gamma = O[0.5 \times n\,(n-1)\,]$

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

The b-chromatics number of a graph G, defined by $b(G)$, is the maximum integer number of colors k_b so that G can have a b-coloring by the k_b 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 v_i, and its number of incidental edges is noted $d(v_i)$.

3.3 Adopted algorithms

The majority of the evaluations of $\chi(G)$ and $b(G)$ come from coloring algorithms. There exist many of them. So, we have chosen to limit the choice of the fastest and most recent ones.

New graph coloring and b-coloring algorithms have been proposed by Effantin and Kheddouci [17,18]. More details on the approximation of the b-chromatic number were proposed by Corteel [19] and a good state-of-art is presented in [20]. All of these algorithms were efficiently introduced into Elghazel's works [21] who proposed a new unsupervised classification method of medical data based on graph b-coloring where the number of classes is not a priori known. On the same database, the comparison between this method and different approaches like the agglomerative hierarchical classification, the approach of Hansen and the classification of DRG, shows that this 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 blocks features, resulting from the physical segmentation stage, is not efficient to discriminate heterogeneous blocks (containing parasitic elements). So as to locate AB on difficult envelopes, more efficiently, it has been necessary to choose an even more advanced tool of progressive grouping of CCs and of AB identification. So, we have taken the powerfulness of the graph coloring into account to automatically separate the elements into homogeneous groups (Fig. 6) and in that sense to considerably improve the ABL system.

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

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

- The envelope image segmentation: coupling of the binarization 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 block interpretation: training of classifier by b-coloring and identification of the address block among several candidates.

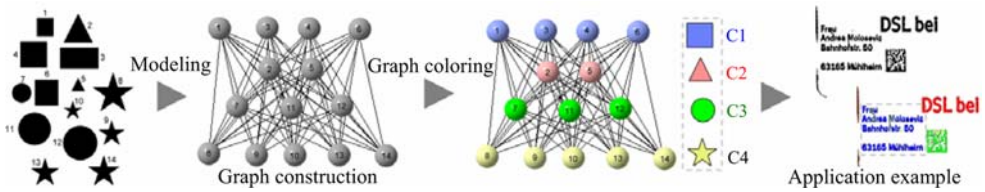

Fig. 6 Example of graph coloring application

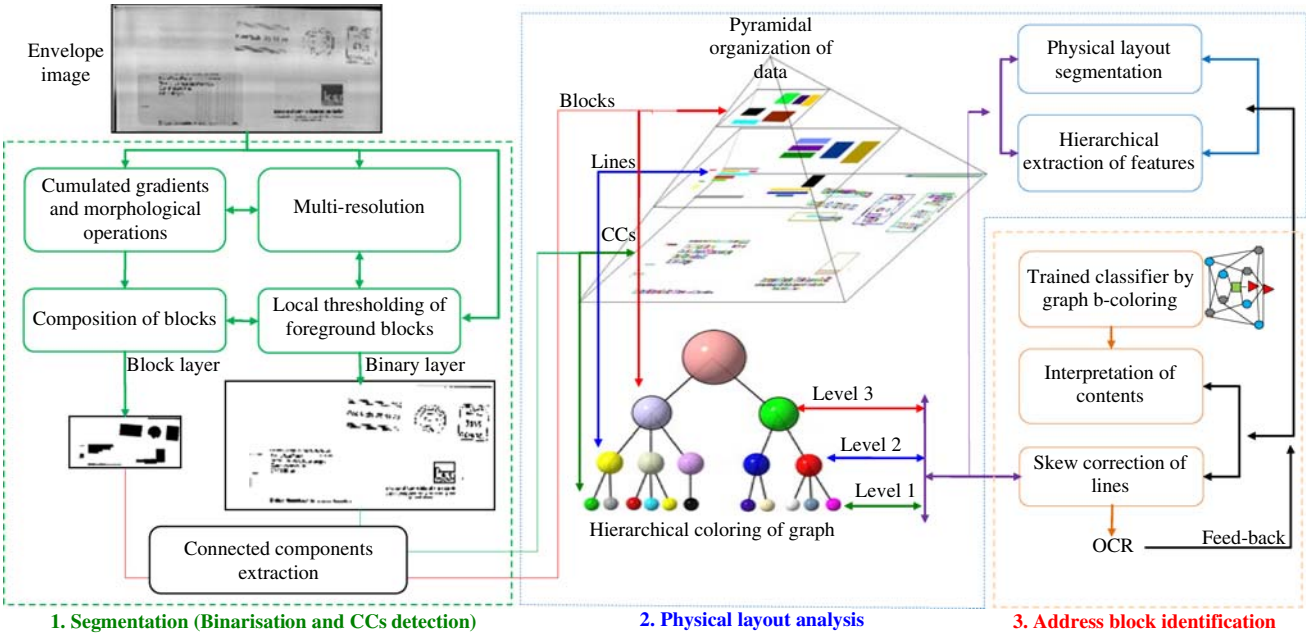

Fig. 7 Functional diagram of the proposed approach

4.1 Proposition of a new thresholding and connected components detection method

The separation between the text zones and the location 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 combines all the required conditions, especially a low time consuming. We have managed to optimize this stage by applying a local threshold only near the text zones (Fig. 8) that can be located by the cumulated gradients method with the multi-resolution and mathematical morphology [8].

After the binarization step, we detect CCs of foreground of the binary and block layers (Fig. 9). The method used is inspired from Pavlidis's studies [10] based on the LAG (line adjacency graph) structure, a structure particularly adapted to a 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: $cc(i) = (x_i^d, y_i^d, x_i^f, y_i^f)$. Let $V(L_1)$ (or $V(L_3)$) be the set of the CCs of the binary layer (or of the block layer) which represents the finest (or the coarsest) level L_1 (or L_3) of the pyramid. With $V(L_k) = \{CC_k(i) | i = 1, \ldots, n_k\}, n_k$ is the number of CCs in the layer $k_{/k=1,2,3}$. 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.

4.2 Hierarchical analysis and block description strategy

Each block can be described by a set of features resulting from the hierarchical analysis of the three levels of the data pyramid. To manage knowledge that are associated to them, each group of objects has features and a strategy of classification. At the bottom of the hierarchy (corresponding to the high resolution image) we find the binary layer CCs. The progression in the hierarchy makes it possible, at each

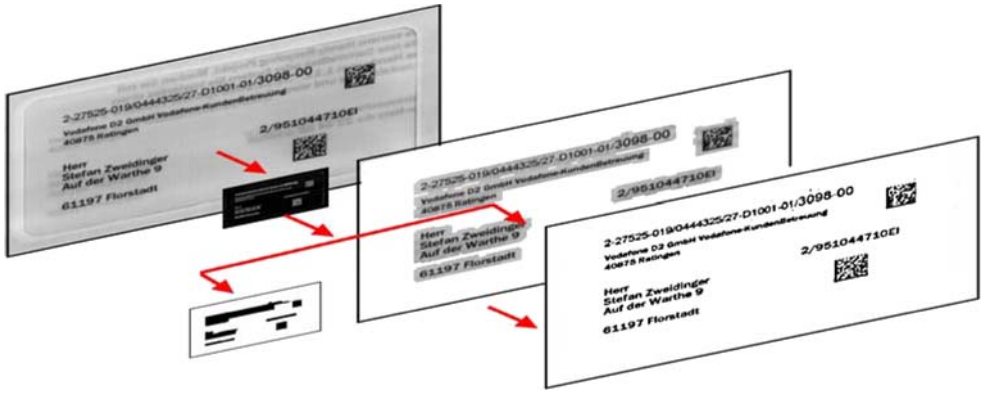

Fig. 8 Our hybrid approach of binarization (text localization/thresholding)

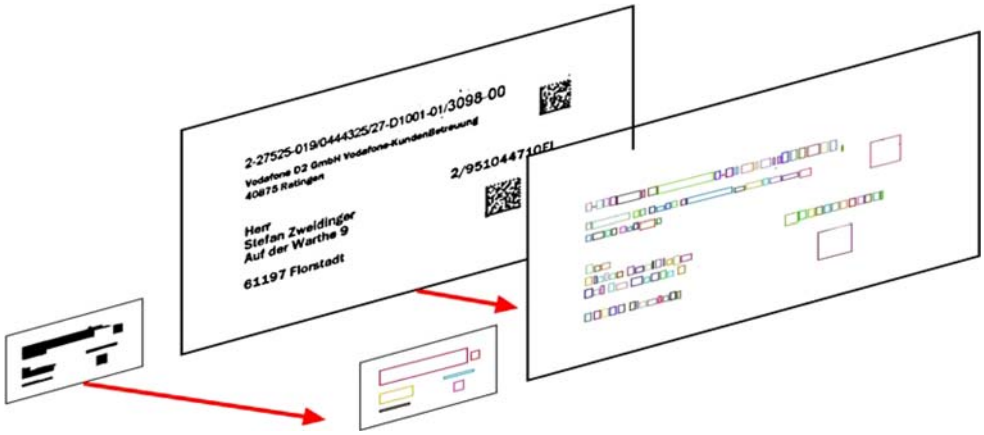

Fig. 9 Bounding boxes of connected components of a binarized envelope image

level, to acquire more precise knowledge on the image content. 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 than the character spacings, or than the blocks position on the envelope (see Fig. 10, Table 1).

Our idea consists in making the block description phase cooperate with the physical segmentation phase. It also allows, at any level of the hierarchy, the use of all information expressed in the other levels (Fig. 10). In that way, the description can take advantage of the two phases. Let $Vd_{Lk}(i)$ be the descriptor-vector of block i at the levels $k = 1, 2$ 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 (see Formula 1):

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

Using the PCA, it is possible to select a minimal number of features while ensuring a perfect disparity between the objects of different natures.

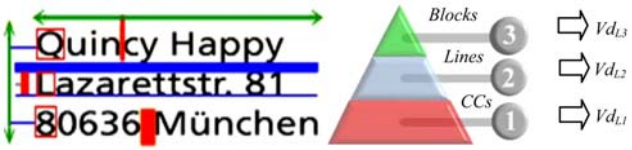

Fig. 10 Hierarchical extraction of features

4.3 Application of the graph coloring to the physical layout

The segmentation techniques can not systematically produce uniform and well located blocks in complex environments. One speaks about oversegmentation when 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 [11].

Table 1 Features perception at the various pyramid levels

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

Table 2 Improvement of the recognition by our thresholding method (location/segmentation)

Mail classes	Increase in OCR-ratio (%)	Mail classes	Increase in OCR-ratio (%)
CCH	+2	FMR	+23
NPAI	+26	PLN	+20
CB	+13	HIM	+76
LA3	+11	TIM	+16
LA4	+11		

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 of physical layout segmentation is based on the same principle of a hybrid strategy. High stages of our approach are based on the Hierarchical Graph Coloring (HGC), that largely makes use of all the levels of the pyramidal structure and the coloring effectiveness, so as to extract, to characterize and precisely to group objects of same nature (see Fig. 11). Therefore, our idea consists in forming, at each level of the hierarchy, groups of components that must be as homogeneous as possible in order to lead to a more precise description. Let G be a non oriented graph at three independent levels defined by the following relationship:

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

with $vd_{Lk} = \{vd_{Lk}(\mathrm{i})\}_{\mathrm{i}=1\ldots\mathrm{n_k}}$ is 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 (Fig. 7), and $E_{Lk>Sk}$ is the finite set of 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 Sect. 4.4.). This definition is given by the following relationship:

$$E_{Lk>Sk}\,[vd_{Lk}(i), vd_{Lk}(j)] = \begin{cases} 1 \text{ si } & d_{i,j} > s_k \\ 0 \text{ si } & \text{non} \end{cases} \tag{3}$$

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 coloring process uses a hybrid strategy of progression into the hierarchy of the graph: the

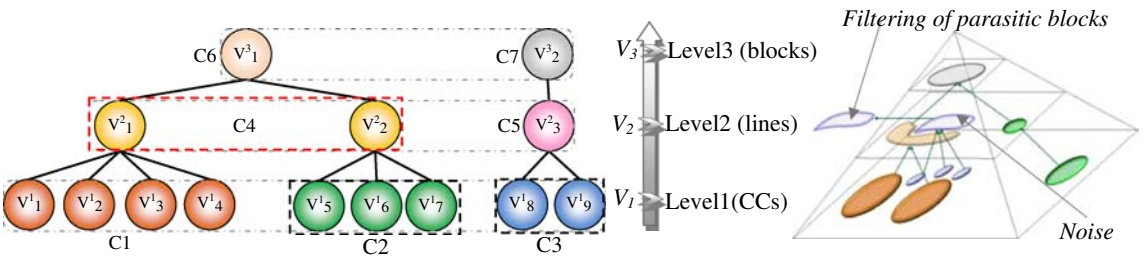

Fig. 11 Hierarchical graph coloring (c_i are the colors of nodes v_j^k)

colors of a level take part in the formation and the description of the nodes of the next level (see Fig. 11).

The steps of the graph G_k coloring are given by the following algorithm [18]:

```
Algorithm 1: Graph_coloring (G_k)
Begin
 If c_i ≠ ∅ then
  Let M := Nc(i) ∪ {c_i};
    q := 0;
  For every node j ∈ N(i) such that c_j := ∅ do
    q := min{k|k > q, k ∉ M and k ∉ c_j};
    If q ≤ Δ + 1 then c_j := q;
    Else c_j := min{k|k ∉ Nc(j)};
    Endif
  Enddo; Endif; 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 $N(i)$, $d(i) = |N(i)|$ its degree, and $\Delta = \text{Max}\{d(i)|i \in V\}$.

Our method of physical layout segmentation is based at the beginning on the first stage graph construction (which is noted G_1) which models the CCs layer $V(L_1)$ of the first data pyramid level. The graph $G_1(Vd_{L1}, E_{L1>S1})$ is then colored by the algorithm1 to form $G_3(Vd_{L3}, E_{L3>S3})$. 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 those new knowledge, we apply once again a second coloring of graph $G_3(Vd_{L3}, E_{L3>S3})$ 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 the coloring process of the graph $G_2(Vd_{L2}, E_{L2>S2})$ with Vd_{L2} the set of descriptors defined by the analysis of $V(L_1)$ and $V^*(L_3)$ (see the stages in Fig. 12). The following algorithm summarizes all the segmentation stages:

```
Algorithm 2 : Physical_segmentation()
Begin
    Level1: regrouping of similar CCs
    For every CC1(i)/i := 1...n1 Extract Vd_L1(i)
    Endfor
    V(L1) := {Vd_L1(i)/i = 1...n1); G1 := G(Vd_L1, E_L1>S1);
Execute: Graph_coloring (G1);
    Level2: homogenization of the blocks layer
      Let M := G1 ∩ Vd_L3
      For every item i of M Extract Vd_L3(i) endfor
      G3 := G(M, E_M>S3); Execute: Graph_coloring (G3);
    Level3: emergence of the text lines
      Let M := Vd_L1 ∩ G3
      For every item i of M Extract Vd_L2(i) endfor
      G2 := G(M, E_M>S2); Execute: Graph_coloring (G2);
Vd_L2 = G2; Vd_L3 = G3;
      Extract Vd_L1 := Vd_L1/G2 and G3;
  Extract Vd_L2 := G2/VdL1 and G3;
Vd_total(i) = {Vd_L1∈VdL3} ∪ {Vd_L2∈VdL3}
Extract Vd_L3 := G3/VdL1 and G2;
      For every block find ;
  End
```

The various stages of segmentation are thus illustrated by the following example:

4.4 Optimization of dissemblance thresholds

In a preparatory phase of auto-parameter setting, our system uses the Levine and Nazif [22] combination of intraclass and interclass disparities to automatically adjust all the threshold values that are necessary for the coloring process. The principle consists in choosing the thresholds that maximize the following cost function:

$$F = M_{\text{Inter\_groups}} + M_{\text{Intra\_groups}} \tag{4}$$

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

$$M_{\text{Inter\_groups}} = \frac{\sum_{Gri} A_i \cdot CT_i}{\sum_{Gri} A_i} \quad \text{where}$$

$$CT_i = \sum_{Gri} \frac{l_{ij}}{l_i} \frac{|m_i - m_j|}{|m_i + m_j|} \tag{5}$$

A_i is the area of nodes of the ith group (color) Gr_i and m_i its average, l_{ij} is the length of the common border between Gr_i and Gr_j, and l_i is the perimeter of group Gr_i. The criterion $M_{\text{Intra\_groupes}}$ 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 basis, 400 blocks of several categories (AB, stamp, logos…), resulting from

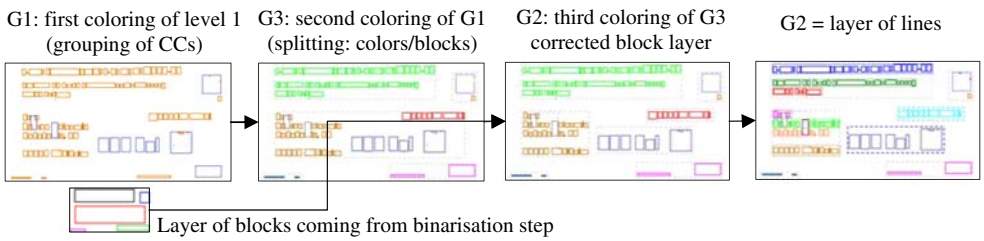

Fig. 12 Example of various stages of physical segmentation

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 basis by a discriminating features vector (Sect. 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 [18] 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 ci := q do
     K := {k|k ∈ L and k ∉ Nc(vi)};
     cj := {c|dist(vi, c) := min_k ∈K (dist(vi, k))};
Enddo;
 For each vertex vj such that cj ∈ NDm do
     Update (Nc(vj));
     If Nc(vj) := L\{cj} then Add(cj, 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_i is the colors of node i.

whereas a supervised classification technique requires the introduction of the number of 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 dominant nodes (representative of the blocks) that ensure a great interclass disparity (Fig. 13). These representative nodes will thus be used in the ABL phase to identify in real time the block address among several candidates.

4.6 Real-time identification of the address-block based on the graph b-coloring

To select the AB in a list of candidate blocks $S = \{v_{i=1,\ldots,n}\}$, we compare the description of each blocks in the space of features with the description of all representative blocks (dominating nodes) $S^*\{V^*_{j=1,\ldots,M}\}$ resulting from the training phase. The matching algorithm determines in real time for each block in S the designation of its adjacent nodes in S^* (Fig. 14). 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 $\alpha(\alpha = 1)$.

$$d_{i,j} = \left(\sum_{k=1}^{Nf} g_k(v_i^k, v_j^{k*})^\alpha \right)^{\frac{1}{\alpha}} \tag{6}$$

If $\alpha = 1 \rightarrow d_{i,j}$ is the City Block distance, $\alpha = 2 \rightarrow d_{i,j}$ is the Euclidean distance and the more α increases, the more $d_{i,j}$ tends to the Chebyshev distance. Nf is the features vector length. g_k is the dissemblance function that compares each pair of feature k (Fig. 14).

5 Experimentation

For the first step of the ABL evaluation, we have produced comparative curves that show that the run-times of our hybrid method of binarization are approximately similar to those of global methods and very inferior than those of local methods. These run-times are calculated on a set of 29,225 envelope images divided in nine classes (Figs. 15, 16).

In order to evaluate the quality of binarized characters, we compare the OCR-rates on images binarized by our hybrid method with those binarized by Sauvola's method. The results of the Table 2 present the increased ratio of OCR performance with our hybrid method of binarization.

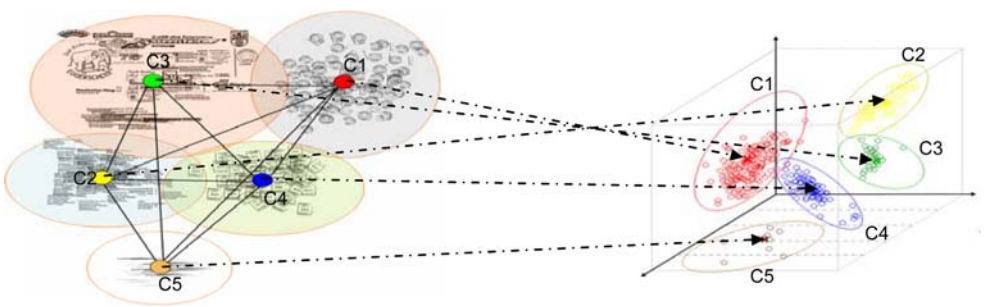

Fig. 13 Features separability and training: block classification by b-coloring and detection of the representative centers of classes

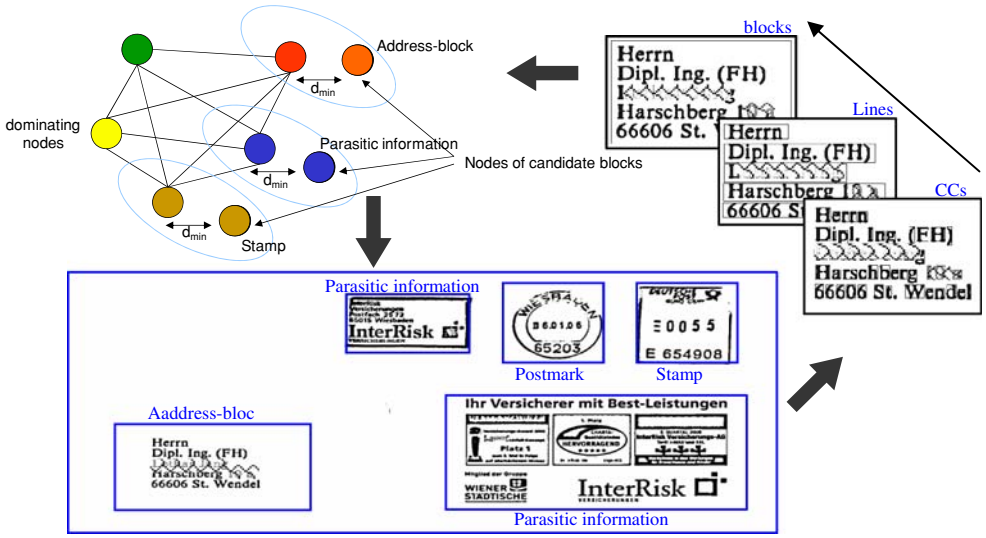

Fig. 14 Real-time identification of the address-block based on the graph b-coloring

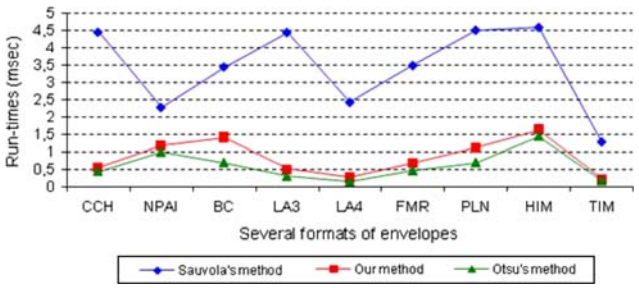

Fig. 15 Run-times comparison of various thresholding methods

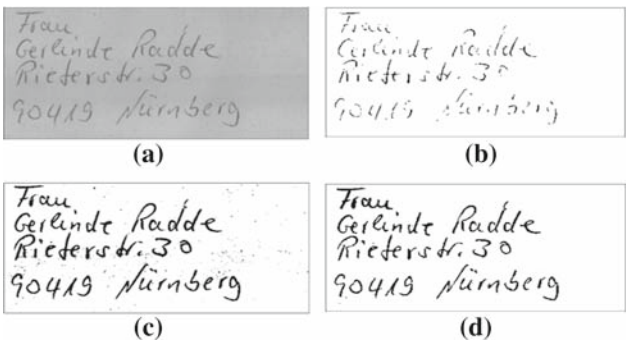

Fig. 16 Comparative performance of thresholding methods, **a** *gray*-levels address-block, thresholding result by **b** Otsu's global method, **c** Sauvola's local method and **d** our hybrid method

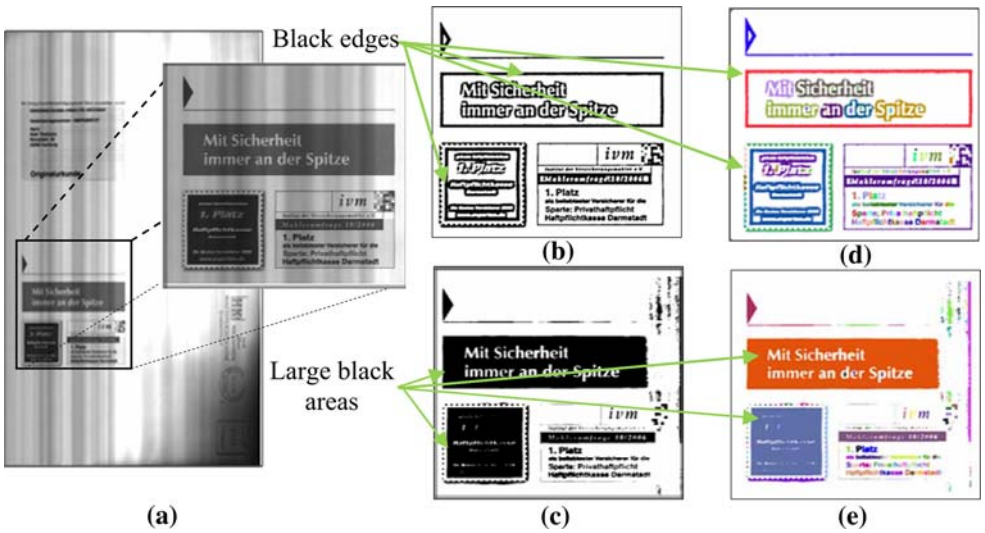

Fig. 17 **a** A *gray*-levels image of envelope, **b** thresholding result by our hybrid algorithm, **c** thresholding result by a classical algorithm, **d**, **e** the connected component labeling of binary images

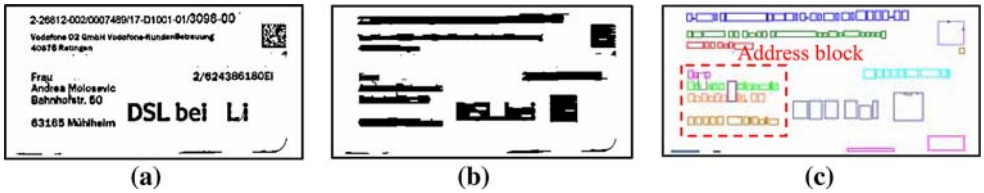

Fig. 18 Noisy address-block (**a**), emergence of the lines by the RLSA method (**b**), segmentation in lines by our coloring graph approach (**c**)

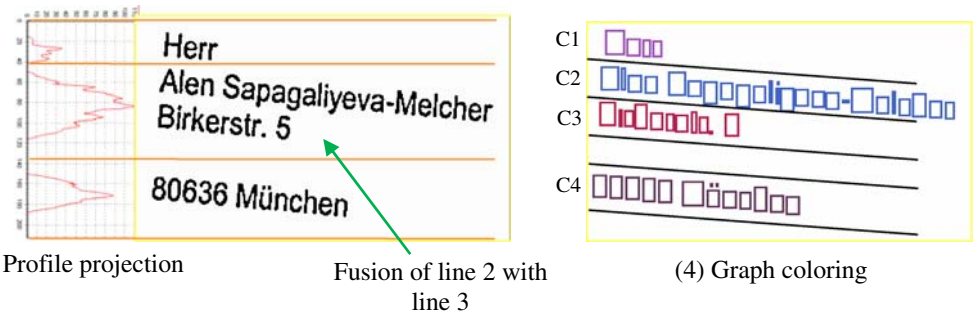

Fig. 19 Skewed text line segmentation: fusion of line 2 with line 3 by the profile projection method (*left*), correct segmentation of the text lines by our HGC method (*right*)

Our hybrid thresholding method has also a second advantage: it speeds up the stage of connected components extraction by the reduction of black pixels in all large black areas that are located as black edges with white centers (see Fig. 17).

Concerning the second step of ABL, the effectiveness of our physical segmentation method has been tested on a set of 10,000 envelope images that were rejected by the old sorting system because of their complexity (Fig. 20). More than 95% of address-blocks are correctly segmented by our method, in opposition to 60% by RLSA method and 30% by profile projection method. The analysis of these results shows that several errors of over or under segmentation introduced by RLSA and profile projection methods can be considerably reduced by using our graph coloring based method (Figs. 18, 19). These performances can be justified by the effectiveness of our method in extracting and separating the textual components (characters, lines and blocs of text) and by its ability to reject most of parasitic components. Thanks to this robustness, the HGC method is definitely more efficient for noisy images segmentation by comparison to classical approaches.

The increase of coherence between the different segmentation phases of our proposition led to a considerable reduction of processing time. To justify this assertion, we show in Fig. 20 the run-time comparison of our method and two standards: the RLSA and the profile projection.

As concerned the third step of ABL, the evaluation of the performances of our approach has been achieved on a corpus of 750 envelope images (considered as difficult and noisy, see Fig. 22). We have obtained more than 98% of good localization. Others tests were carried out, as well, on a basis of 100 images that are rejected by the currently most

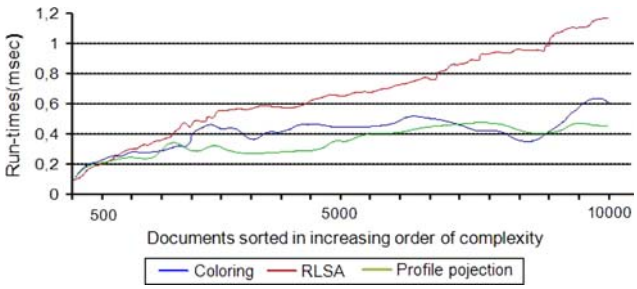

Fig. 20 Run-time comparison of three methods of physical layout segmentation (ours in named Coloring)

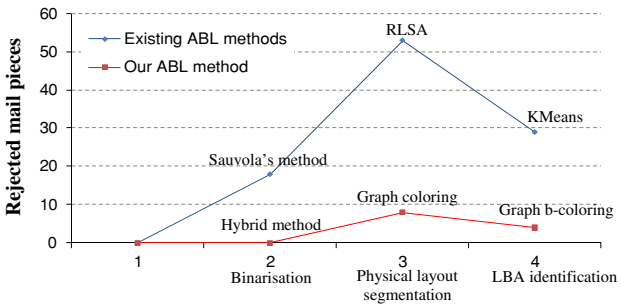

Fig. 21 Performances of our method of ABL

powerful systems of localization. All these rejections are due to the failure of one of the localization phases. The following curve shows that 18 envelopes were rejected due to a bad binarization, 53 due to a bad physical layout segmentation and 29 due to a bad identification of the address block. By using our approach, we obtain an average of 93% of good localization (Fig. 21).

Fig. 22 Some examples of ABL on an envelope images considered as difficult and noisy

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 causes of physical segmentation errors, and that the b-coloring leads to a noticeable improvement of the interpretation step of candidate blocks. Moreover, our proposition has also shown a real increase of coherence between the different modules of segmentation while decreasing the rejection rate and reducing the computing times. This work is granted by CESA Company (http://www.cesa.fr).

References

1. Ching-Huei, W., Palumbo, P.W., Srihari, S.N.: Object recognition in visually complex environments: an architecture for locating address blocks on mail pieces. Pattern Recognition, 9th International Conference, IEEE, vol. 1, pp. 365–367 (1988)
2. Viard-Gaudin, C., Barba, D.: A multi-resolution approach to extract the address block on flat mail pieces, ICASSP-91. International Conference, vol. 4, pp. 2701–2704 (1991)
3. Yu, B., Jain, A.K., Mohiuddin, M.: Address block location on complex mail pieces. Document Analysis and Recognition. Fourth International Conference, IEEE, vol. 2, pp. 897–901 (1997)
4. Jeong, S.H., Jang, S.I., Nam, Y.-S.: Locating destination address block in Korean mail images. ICPR 2004, IEEE, 17th International Conference, vol. 2, pp. 387–390 (2004)
5. Eiterer, L.F., Facon, J., Menoti, D.: Postal envelope address block location by fractal-based approach. Computer Graphics and Image Processing. 17th Brazilian Symposium, IEEE, pp. 90–97 (2004)
6. Otsu, N.: A threshold selection method from grey-level histogram. IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979)
7. Sauvola, J. et al: Adaptive Document Binarization, ICDAR'97, vol. 1, pp. 147–152 (1997)
8. Gaceb, D., Lebourgeois, F., Eglin, V., Emptoz, H.: Contribution to the automatic recognition of business documents, 6p, IWFHR, La Baule, France (2006)
9. Regentova, E., Latifi, S., Deng, S., Yao, D.: An algorithm with reduced operations for connected components detection in ITU-T group 3/4 coded images. Pattern Anal. Mach. Intell. IEEE **24**, 1039–1047 (2002)
10. Pavlidis, Z., Zhou, J.: A page segmentation and classification. CVGIP92, vol. 54, no. 6, pp. 484–496 (1997)
11. Mullot, R.: book: Les documents écrits de la numérisation à l'indexation par le contenu, Editeur: Hermes science Publication, ISBN-10: 2746211432, p. 365 (2006)
12. Déforges O., Barba D.: A fast multiresolution text-line and non text line structures extraction and discrimination scheme for document image analysis. ICPR 94, pp. 134–138 (1994)
13. Drivas, D., Amin, A.: Page segmentation and classification utilising a bottom-up approach. Document Analysis and Recognition. ICDAR. Proceedings of the Third International Conference, vol. 2, pp. 610–614 (1995)
14. Shi, Z., Govindaraju, V.: Line separation for complex document images using fuzzy runlength, Document Image Analysis for Libraries, DIAL 2004. Proceedings, First International Workshop, pp. 306–312 (2004)
15. Wang S.-Y., Yagasaki T.: Block selection: a method for segmenting a page image of various editing styles, ICDAR. Proceedings of the Third International Conference on vol 1, pp. 128–133 (1995)
16. Agne, S., Rogger, M.: Benchmarking of Document Page Segmentation, Part of the IS&T/SPIE Conference on Document Recognition and Retrieval VII. San Jose, California, pp. 165–171 (2000)
17. Effantin, B., Kheddouci, H.: The b-chromatic number of power graphs. Discrete Math. Theor. Comput. Sci. (DMTCS) **6**, 45–54 (2003)
18. Effantin, B., Kheddouci, H.: a distributed algorithm for a b-coloring of a graph, the fourth international symposium on parallel and distributed processing and applications (ISPA'2006), serrento, Italy (2006)
19. Corteel, S., Valencia-Pabon, M., Vera, J.-C.: On approximating the b-chromatic number , Discrete Applied Mathematics archive, vol 146, pp. 106–110. ISSN:0166-218X (2005)
20. Paschos, V.: Book: Optimisation combinatoire 5: problèmes paradigmatiques et nouvelles problématiques, p. 270. Lavoisier, France (2007)

21. Elghazel, H., Hacid, M.-S., Khedddouci1 H., Dussauchoy, A.: A new clustering approach for symbolic data: algorithms and application to healthcare data. 22 éme journées bases de données avancées, BDA 2006, Lille. Actes On formal proceeding (2006)

22. Philipp-Foliguet, S.: Evaluation de la segmentation. Rapp. Tech. (2001)