Interest of perceptive vision for document structure analysis

Aurélie Lemaitre[a] and Jean Camillerapp [b] and Bertrand Coüasnon [b]

[a]Université de Rennes 2, Campus de Villejean, 35043 Rennes, France;
[b]INSA, avenue des Buttes de Coësmes, 35043 Rennes, France

UMR IRISA, Campus de Beaulieu, 35042 Rennes, France
Université Européenne de Bretagne, France

ABSTRACT

This work addresses the problem of document image analysis, and more particularly the topic of document structure recognition in old, damaged and handwritten document. The goal of this paper is to present the interest of the human perceptive vision for document analysis. We focus on two aspects of the model of perceptive vision: the perceptive cycle and the visual attention. We present the key elements of the perceptive vision that can be used for document analysis.

Thus, we introduce the perceptive vision in an existing method for document structure recognition, which enable both to show how we used the properties of the perceptive vision and to compare the results obtained with and without perceptive vision. We apply our method for the analysis of several kinds of documents (archive registers, old newspapers, incoming mails ...) and show that the perceptive vision significantly improves their recognition. Moreover, the use of the perceptive vision simplifies the description of complex documents. At last, the running time is often reduced.

Keywords: Perceptive vision, perceptive cycle, visual attention, document analysis

1. INTRODUCTION

During the last years, numerous digitization campaigns have been led in order to preserve historical documents. Thus, big amounts of document images are now available and it is necessary to recognize parts of their content in order to index them.

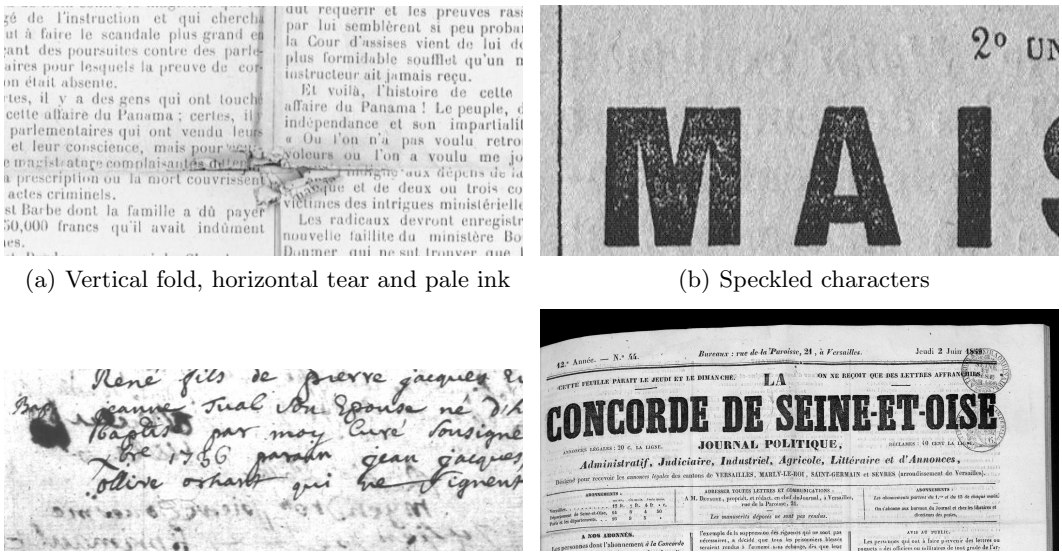
In this paper, we focus on the problem of document structure recognition in old, damaged and handwritten document, like archive documents, old newspapers, handwritten registers... These old documents present some specific difficulties for their recognition,[1] such as:

- damages of the paper: tear (figure 1(a)), fold (figure 1(a)),
- damages of the content and of the ink: speckled characters (figure 1(b)), blots (figure 1(c)), pale ink (figure 1(a)), bleed trough ink (figure 1(c));
- damages due to the digitization step: curvature (figure 1(d)).

These artifacts are called "noise" in documents.

Many approaches have been proposed in the field of document analysis, but Mao *et al.*[2] point out their limits concerning the recognition of archive documents. Firstly, the methods are usually dedicated to specific goals such as segmentation of text blocks, titles. They are not generic enough for being applied to historical documents which structure may vary from one collection to another. Secondly, many works aims at labeling the logical structure and suppose that the physical segmentation has been properly realized, which is not obvious in archive documents. Indeed, we are faced with Sayre paradox: the physical segmentation sometimes requires knowledge about the logical contents. Thirdly, most of the approaches are not able to properly deal with noise that is

Further author information: (Send correspondence to A. Lemaitre)
A. Lemaitre: E-mail: aurelie.lemaitre@irisa.fr



(a) Vertical fold, horizontal tear and pale ink

(b) Speckled characters

(c) Blot and bleed trough ink

(d) Curvature due to digitization step

Figure 1. Difficulties for the recognition of old documents

present in old documents. Thus, the problem of the recognition of old, damaged or handwritten documents is still open.[1]

In order to solve a few problems and make easier the creation of recognition systems, we propose to exploit the properties of the perceptive vision for document structure recognition. We call perceptive vision the faculty of the human brain to combine several points of view for the interpretation of a scene.

In this paper, we first present the interesting components of the perceptive vision that we used in our work. Then, we present how this elements have been used in the literature, in the field of document structure recognition. In part 4 and 5, we present the way we introduce the perceptive vision in an existing system of document structure recognition. At last, we validate the interest of the perceptive vision for the recognition of four kinds of documents and more than 131,000 pages.

2. INTERESTING ELEMENTS OF PERCEPTIVE VISION

In the models of the perceptive vision proposed in the field of neuropsychology, we retained two main components: the *perceptive cycle* guided by the *visual attention* (figure 2). We present the models of these mechanisms that we applied for document image analysis.

2.1 Perceptive cycle

The perceptive cycle[3,4] is a physiological model of the vision. We consider three steps.

The first step is the acquisition of an image on the retina ("Image" box of figure 2). In the second step, the human is able to automatically extract preattentive primitives in the image such as colors or shapes[5] ("Visual primitives" box of figure 2). In the third step, the brain is able to organize these primitives into coherent objects thanks to the activation of specific models stored in the memory ("Knowledge" box of figure 2). This process is cyclic: the total recognition of a model may require a new acquisition of an image, with a new point of view.[6] A point of view is composed of the choice of the level of perception, and of the area of interest. The cooperation between primitives extracted with different points of view is guided by the knowledge about the models.

Intuitively, the perceptive cycle is interesting for document image recognition as it enables the cooperation between several points of view, being guided by the knowledge.

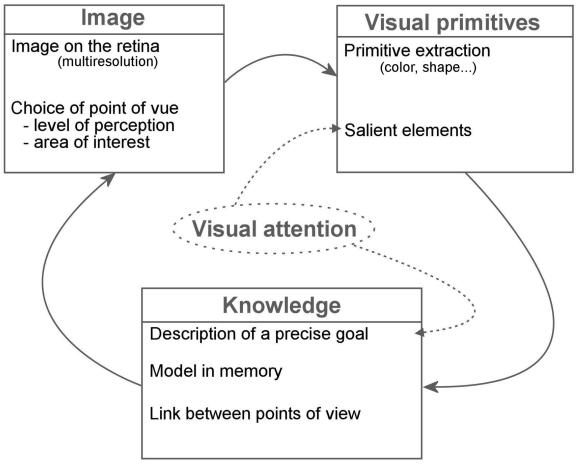


Figure 2. Model of human vision : perceptive cycle guided by visual attention

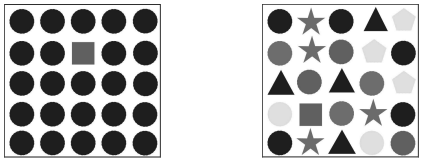
2.2 Visual attention

The visual attention[7] is a psychological aspect that guides the perceptive cycle (figure 2). It takes two forms.

The first kind of attention is guided by the vision of primitives that are particularly salient. It is the "pop-out" effect (figure 3(a)). This form of attention is automatic. It can be used when their is no *a priori* knowledge about what we can see.[8] This kind of attention is an entry point of the perceptive cycle (figure 2) as it enables to detect visual primitives that are particularly salient.

The second form of visual attention is guided by a precise goal, that is to say the activation of a specific model in the memory. It requires an effort (figure 3(b)). In that case, the visual attention enables to recognize complex modules, even in unfavorable conditions.[9] This kind of attention is another entry point of the perceptive cycle (figure 2) as it enables to start a recognition process led by the description of a precise goal.

In human vision, both mechanisms operate in parallel.



(a) The square is salient

(b) Looking for the square requires an effort

Figure 3. The two forms of visual attention

Intuitively, the visual attention is interesting for document image recognition as it enables to select easily the relevant information in all the available data.

2.3 Perceptive vision

As a conclusion, we consider that the following points may contribute to set up a system based on perceptive vision for document structure recognition:

- the perceptive cycle and its three components:
  - image,

  - visual primitives,
  - knowledge,

- the two forms of visual attention:
  - attention guided by salient elements,
  - attention guided by a precise goal,

3. RELATED WORK ON PERCEPTIVE VISION

We now present how the perceptive vision is used in the field of document recognition, and its main interests.

The perceptive vision has been used in several fields of computer vision but we concentrate on related works in the field of document analysis. We propose to classify the approaches of the literature according to the kind of attention they are inspired.

On the one hand, several perceptive works are proposed around the attention guided by salient elements.[10] For example, Likforman *et al.* propose to use the saliency of text lines for their recognition.[11] The main interest of these data-guided approaches is the ability to analyze basic components (text lines, line segments), without specific knowledge dedicated to this kind of document.

On the other hand, some methods are inspired by the attention guided by a goal. Thus, several works use the perceptive vision to describe complex models.[12–14] In these approaches, the interest is the ability to recognize complex patterns in noisy environments.

In short, the two kinds of attention have shown an interest for document analysis. However, to our knowledge, the methods of the literature always use only one kind of visual attention, whereas human vision needs to combine the two kinds of attention. Moreover, in the literature, each work proposes a dedicated mechanism of cooperation that can be applied for only one kind of problem.

In this paper, we propose a method which novelty consists in two aspects:

- we combine the two forms of visual attention, which enable both to extract salient elements and to describe models being guided by a precise goal,
- we provide a generic framework for the creation of perceptive cooperation mechanisms that may be adapted to each kind of applications.

4. CONTEXT OF OUR WORK

We work in the context of an existing method for document structure recognition, called DMOS[15](Description and MOdification of the Segmentation). This method is made of a bidimensionnal grammatical language, EPF (Enhanced Position Formalism), which enables a physical and logical description of the structure of documents.

For each kind of document to study, the user builds a grammatical description that consist in explaining the relative position of each structural element in the image. The terminals of the grammatical analysis are the connected components and the line segments extracted from the image. When a description has been realized in EPF, the associated parser is automatically produced by a compilation step.

Thanks to this EPF formalism, the knowledge is separated to the system, and the digital level is entirely guided by the symbolical description. Thus, this method is generic and can be applied on any kind of document. It has been validated on various kinds of documents: musical scores, mathematical formulas, military forms and at a large scale on more than 500,000 document pages.[16]

As DMOS method can be adapted for the recognition of many kinds of document, we propose to work in this context in order to validate the interest of the perceptive vision on various problems. Thus, in this work, we propose to improve the existing method DMOS in order to build a system based on the perceptive vision. We call this improved method DMOS-P. Using this approach, we can then compare the two methods (original DMOS and perceptive DMOS-P) in order to show:

- what is required to set up a system based on elements of the perceptive vision presented in section 2.3,
- how much the perceptive vision can improve an existing system.

5. SETTING UP A SYSTEM BASED ON PERCEPTIVE VISION

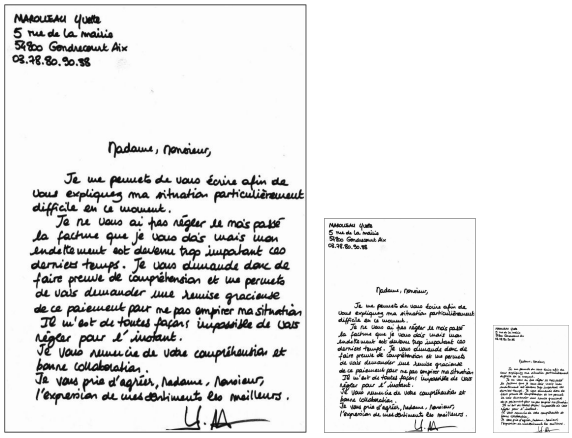
We identified in section 2.3 all the elements that are required to set up a system based on perceptive vision, that are presented on figure 2.

As an example of implementation, we describe in this section how we implement each item of the figure 2 in the existing DMOS method, to produce the DMOS-P method.

5.1 Image

This part concern the implementation of the "Image" box of figure 2. Indeed, we have shown that we need to imitate the perception if multiresolution images on the retina, and choose a specific point of view, that is composed of a level of perception and an area of interest.

Multiresolution images on the retina In order to imitate the acquisition of images on the retina (focusing with fovea), we based our approach on a pyramid of multiresolution images (figure 4), built by a recursive low filtering and sub-sampling from the initial image.



(a) (b) (c)

Figure 4. Pyramid of multiresolution images

Choice of point of view At each moment of the analysis, we must be able to choose the point of view, that is to say

- the level of perception: which resolution is used,
- the area of interest: which part of the image is studied.

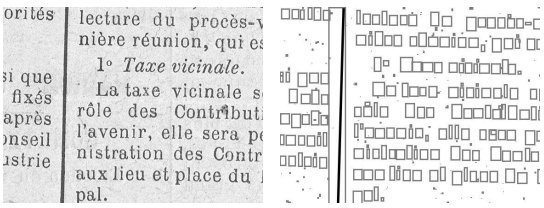
For this purpose, we propose to introduce two operators in the grammatical description. In each step of the analysis, the operator `USE_LAYER` makes it possible to choose the resolution that is analyzed, whereas the operator `AT` makes it possible to choose the current area of interest in the image. Thus, each step of the analysis is realized in a specific point of view with a choice of the level of perception (resolution) and area of interest.

5.2 Visual primitives

This part concern the implementation of the "Visual primitives" box of figure 2. We present the visual primitives that we used and their organization.

Primitive extraction Our method is based on two kinds of primitives: line segments and connected components. Thus, we apply the corresponding extractors on each image of the pyramid, and obtain several sets of primitives. All of these primitives can be used as terminals for the grammar.

Data organization However, we have to organize the primitives. So, we propose a new formalism, the *direct perceptive layer*. It is defined as: set of primitive extracted in an image at a given resolution (example on figure 5).



(a) Image

(b) Perceptive layer containing connected components and line segments extracted in the image

Figure 5. Example of perceptive layer

The key point of this formalism is the ability to organize the primitives, keeping a link with their original level of perception. Each primitive can then be used as a terminal for the grammar (figure 6), whatever its resolution of origin.

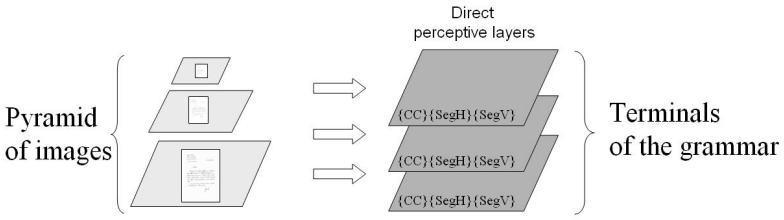


Figure 6. Primitives organized in perceptive layers, used as terminals for the grammar

5.3 Knowledge

This section explain how we implemented the "Knowledge" box of figure 2.

Model in memory Thanks to EPF language, it is possible to describe specific multiresolution models. This fits with the concept of model in memory. Thus, each model is expressed by one grammatical description, which relies on the terminal contained in the perceptive layers associated to each level of perception.

Link between points of view The key point for the cooperation between resolutions is to set up a correspondence between elements provided by the various points of view. Our main solution consists in using coherent coordinates in every layer. Thus, it is possible to directly combine elements obtained from several points of view.

However, we have been faced with two problems that are the quantization error and the changing nature of primitives. We then proposed the concept of *abstract line*. This concept is detailed in.[17] It is a polygonal approximation of the segments that can be used independently of the resolution and which position can be adapted depending on the presence of pixels. This is important for the cooperation between points of view.

5.4 Cycling aspect

We have presented each element of the perceptive cycle: image, visual primitives, knowledge (resumed on figure 2). They are organized into a cycle inherent in DMOS method. Indeed, the analysis is totally led by the symbolical description defined in EPF. Particularly, the description may ask a change of point of view which involve a new run of the cycle. Moreover, the parser of DMOS method is able to deal with backtracking, which also involve a cycling analysis.

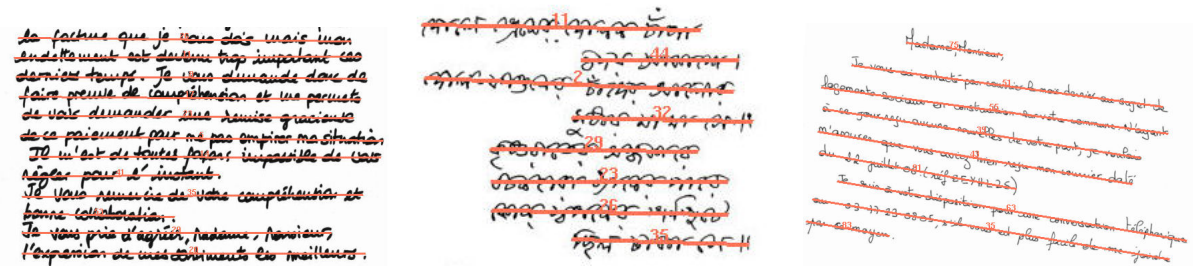
5.5 Visual attention

We now present how our method, based on the perceptive cycle, also fits with the two kinds of attention.

Description of a precise goal The attention guided by a goal is realized by the grammatical description of a precise objective in EPF language. Thus, we can express the description of complex structure that will be detected. In the applications (section 6), we will present some examples of recognition of complex structures, being guided by a given model of the kind of document.

Salient elements In the section 3, we have shown the interest to use the attention guided by salient elements as it enables to recognize basic structural elements without knowledge dedicated to the kind of document.

In our work, we propose to detect two kinds of elements that are particularly salient in document images: text lines (figure 5.5) and rulings. For these two kinds of elements, the use of perceptive vision simplify their perception. Thus, we proposed two grammatical descriptions, based on the use of perceptive vision and that makes it possible to extract elements independently of the kind of document.



(a) (b) (c)

Figure 7. Example of salient text lines

The salient elements that are built below are some specific primitives in the perceptive cycle. Thus, they must be used as terminals for the grammar, like any of the other primitives. So, we propose an extension of our formalism of layer: the induced perceptive layer, composed of the salient elements that are built by combination of data contained in other layers. As the other layers, the induced layers contain primitives that can be used as terminals of the grammar (figure 8. Thus, the text lines and the rulings that are salient are used as new terminals of the grammar.

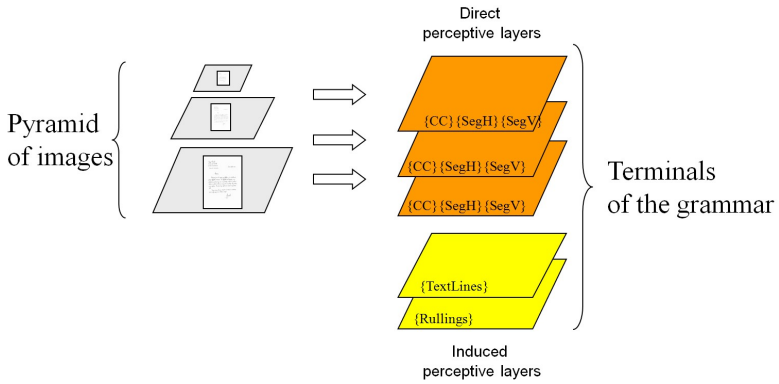


Figure 8. Induced layers containing salient elements: text lines and rulings

6. INTEREST OF PERCEPTIVE VISION

Thanks to the DMOS-P method, we propose to demonstrate the interest of the perceptive vision (by comparison with DMOS).

6.1 Validation on several kinds of documents

We propose to study four kinds of documents and problems:

- localization of fields on archive registers[18] (figure 9(a)),
- localization of text lines and text blocs in incoming mails[19] (figure 9(b)),
- structure extraction in old newspapers[20] (figure 9(c)),
- baseline positioning on bangla handwritten pages[21] (figure 9(d)).

Thus, we treated various kinds of documents: handwritten, printed, old, recent, damaged. For each kind of problem, we easily created a specific system of perceptive cooperation. Indeed, the creation of a new system requires only two steps:

1. the choice of the resolutions and perception levels, which is entirely modular thanks to the formalism of perceptive vision;
2. the symbolical grammatical description of the kind of document, which guides the cooperation between perceptive layers.

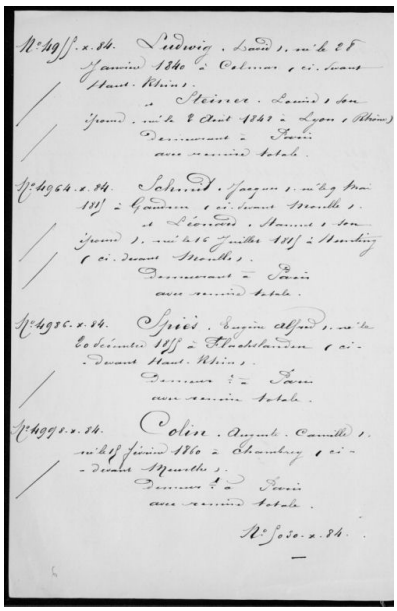
Then, according to the principle of DMOS method, the associated parser is automatically produced by compilation.

6.2 Numerical results

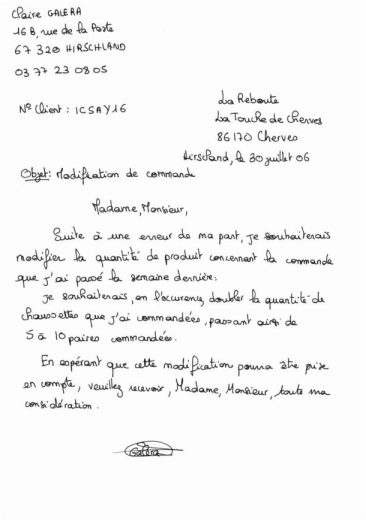
For each of the for studied kinds of documents, we compare two recognition systems: one created with the initial DMOS method (without perceptive vision) and the other with the perceptive DMOS-P method. We sum-up the obtained recognition rates in table 1.

First, the results show that the perceptive vision significantly improve the recognition rates in all the studied kinds of documents. For example, in the case of old newspapers, the recognition of rulings goes from 69.1% to 94.4% of correct recognition.

Moreover, we noticed in these applications that the perceptive vision simplifies the grammatical description. For example, the initial DMOS method was inapt to extract baseline in Bangla handwriting, whereas DMOS-P is very suitable thanks to the perceptive vision.



(a) Archive registers



Claire GALERA
16 B, rue de la Poste
67 320 HIRSCHLAND
03 77 23 08 05

N° Client : ICSAY16

Ets Rebouté
Zae Touche de Cherves
86 170 Cherves

Hirschland, le 30 juillet 06

Objet: Modification de commande

Madame, Monsieur,

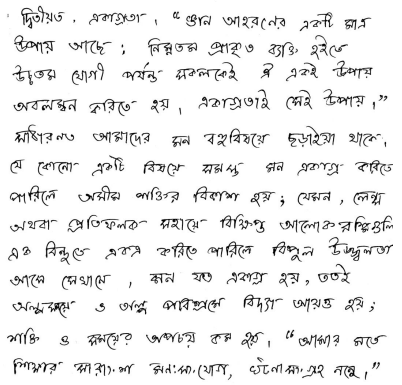
Suite à une erreur de ma part, je souhaiterais modifier la quantité de produit concernant la commande que j'ai passé la semaine dernière.

Je souhaiterais, en l'occurrence, doubler la quantité de chaussettes que j'ai commandées, passant ainsi de 5 à 10 paires commandées.

En espérant que cette modification pourra être prise en compte, veuillez recevoir, Madame, Monsieur, toute ma considération.

(b) Incoming mails

(c) Old newspapers



(d) Bangla handwritten pages

Figure 9. Examples of treated documents

		DMOS (Initial method)		DMOS-P (Perceptive vision)	
Collection	Size	Recognition rate	Running time	Recognition rate	Running time
Archive registers	3,186 acts	92.7%	6.4 sec	98.3%	1.2 sec
Incoming mails	1,250 pages	81.3%	51.4 sec	92.0%	4.85 sec
Old newspapers	4,967 rulings	69.1%	0.34 sec	94.4%	0.48 sec
Bangla handwriting	2,886 words	Inapt	Inapt	97.9%	0.18 sec

Table 1. Results on four kinds of documents for a total of 1,793 pages

At last, the perceptive vision simplifies the combinatory, and often decreases the running time. For example, the running time is divided by 10 for the localization of text blocs in incoming mails.

Generally speaking, we point out a mechanism of prediction/verification, linked to the perceptive vision: at low resolution, hypotheses on the contents are proposed, that are verified at a higher resolution. This mechanism simplifies and improves document recognition:

- for noisy documents (archive registers, old newspapers), the perceptive vision makes it possible to select only relevant information,
- for low structured documents (mails, Bangla handwriting), the perceptive vision helps to rebuild the structure.

Thanks to the good results validated on 1,793 pages, we applied our work for the treatment at a large scale of more than 131,000 document pages. These results have been used in industrial transfers to archive centers.

6.3 Position in relation to the literature

We have shown on section 3 that the existing related work on perceptive vision was presenting two limits: the use of a single form of visual attention and the lack of genericity.

The novelty of DMOS-P is to combine the two forms of attention proposed by the neuropsychologists: the attention guided by salient data, and the attention guided by a precise goal. This combination increases the power of the method.

Moreover, our system is generic and flexible thanks to our formalism of perceptive layer. The method can be used to easily generate specific mechanisms of perceptive cooperation, adapted to each kind of document. Indeed, it is very important to exploit the large diversity of perceptive mechanisms to solve various problems of document analysis.

7. CONCLUSION

In this paper, we have presented the elements of the perceptive vision that can improve document analysis. Thus, we have pointed out that the perceptive cycle enables to combine several points of view of a same document, being guided by the knowledge. We have shown that using the visual attention guided by salient elements enables to recognize basic elements without *a priori* knowledge on the kind of document to recognize. The visual attention guided by a precise goal enables to recognize complex elements being guided by a model, even in a noisy environment. Consequently, it is important to introduce all these aspects to set up an efficient recognition system based on the perceptive vision.

In our work, we have proposed to improve an existing generic method of document structure analysis: DMOS. The new method DMOS-P is also generic and has been applied to four kinds of documents, in order to demonstrate the interest of the perceptive vision. Indeed we compared the efficiency of the two methods for the recognition of difficult images: old, damaged or irregular handwritten documents.

The obtained results demonstrate the interests of the perceptive vision for document structure analysis. Thus the combination of various points of view of the same document improves the recognition rates while simplifying the description. Moreover, the running time is reduced.

In the applications, we point out a mechanism of prediction/verification, linked to the perceptive vision: at low resolution, hypotheses on the contents are proposed, that are verified at a higher resolution. This mechanism simplifies and improves document recognition: for noisy documents, the perceptive vision makes it possible to easily select relevant information, whereas for low structured documents, the perceptive vision helps to rebuild the structure.

REFERENCES

[1] Antonacopoulos, A. and Downton, A. C., "Special issue on the analysis of historical documents," *International Journal on Document Analysis and Recognition* **9**, 75–77 (Apr. 2007).

[2] Mao, S., Rosenfeld, A., and Kanungo, T., "Document structure analysis algorithms: a literature survey," in [*Document Recognition and Retreval X, (Proceedings of SPIE/IST)*], **5010** (Jan. 2003).

[3] Neisser, U., [*Cognition and Reality : principles and implicatios of cognitive psychology*], W. H. Freeman and Company (1976).

[4] Treisman, A., "L'attention, les traits et la perception des objets," *Introduction aux sciences cognitives* (1992).

[5] Wolfe, J. M., "Guidance of visual search by preattentive information," *Neurobiolgy of Attention* **1**, 101–104 (2005).

[6] Bar, M., "Top-down facilitation of visual object recognition," *Neurobiology of Attention* , 140–145 (2005).

[7] Itti, L. and Koch, C., "Computational modeling of visual attention," *Nature Reviews Neuroscience* **2**, 194–203 (Mar 2001).

[8] Sarkar, S. and Boyer, K. L., "Perceptual organization in computer vision: a review and a proposal for a classificatory structure," *Systems, Man and Cybernetics, IEEE Transactions on* **23**(2), 382–399 (1993).

[9] Noton, D., "A theory of visual pattern perception," *Systems Science and Cybernetics, IEEE Transactions on* **6**, 349–357 (Oct. 1970).

[10] Wattenberg, M. and Fisher, D., "A model of multi-scale perceptual organization in information graphics," in [*Proceedings of IEEE Symposium on Information Visualization*], (October 2003).

[11] Likforman-Sulem, L., Hanimyan, A., and Faure, C., "A Hough based algorithm for extracting text lines in handwritten documents," in [*International Conference on Document Analysis and Recognition (ICDAR)*], 774–777 (1995).

[12] Cheng, H. and Bouman, C., "Multiscale bayesian segmentation using a trainable context model," *IEEE Transactions on Image Processing* **10**, 511–525 (Apr. 2001).

[13] Cinque, L., Forino, L., Levialdi, S., Lombardi, L., and Tanimoto, S. L., "Understanding the page logical structure.," in [*10th International Conference on Image Analysis and Processing (ICIAP 1999)*], 1003–1008 (1999).

[14] Déforges, O. and Barba, D., "Segmentation of complex documents multilevel images: A robust and fast text bodies-headers detection and extraction scheme 770," in [*ICDAR*], 770–773 (1995).

[15] Coüasnon, B., "DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *International Journal on Document Analysis and Recognition, IJDAR* **8(2)**, 111–122 (2006).

[16] Coüasnon, B., "DMOS: A generic document recognition method to application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems," in [*Proceedings of International Conference on Document Analysis and Recognition (ICDAR'01)*], 215–220 (2001).

[17] Lemaitre, A., Camillerapp, J., and Coüasnon, B., "Multiresolution cooperation improves document structure recognition," *International Journal on Document Analysis and Recognition (IJDAR)* **11**, 97–109 (November 2008).

[18] Lemaitre, A., Camillerapp, J., and Coüasnon, B., "Contribution of multiresolution description for archive document structure recognition," in [*Proceedings of International Conference on Document Analysis and Recognition (ICDAR'07)*], IAPR, ed., 247–251 (2007).

[19] Lemaitre, A., Camillerapp, J., and Coüasnon, B., "A generic method for structure recognition of handwritten mail documents," in [*Proceedings of SPIE Electronic Imaging 2008, Document Recognition and Retrieval (DRR XV)*], (2008).

[20] Lemaitre, A., Camillerapp, J., and Coasnon, B., "Use of perceptive vision for rulling recognition in ancient documents," in [*Eighth IAPR International Workshop on Graphics RECognition (GREC 2009)*], (2009).

[21] Lemaitre, A., Chaudhuri, B. B., and Coüasnon, B., "Perceptive vision for headline localisation in bangla handwritten text recognition," in [*Proceedings of International Conference on Document Analysis and Recognition (ICDAR'07)*], IAPR, ed., 614–618 (2007).