Automatic Modeling and Recognition of Heterogeneous Logical Structures from Digitized Business Documents

Louisa KESSI, Frank LEBOURGEOIS, Christophe GARCIA
LIRIS - Laboratoire d'Informatique en Image et Systèmes d'information, UMR CNRS \INSA de Lyon 5205, F-69621, France.
{ louisa.kessi, franck.lebourgeois,christophe.garcia }@insa-lyon.fr

ICVSS 2014 — Sicily ~ 13-19 July — International Computer Vision Summer School

Abstract and Motivation

***The main objective*:** the automatic recognition of document structure by image analysis of highly heterogeneous business documents without any predefined model. Two main complementary directions are provided:

First direction, is to develop an automatic structure recognition system to generate specific model of business documents from the fusion of existing knowledge databases, complemented by exogenous information.

Second direction and most ambitious, is to automatically generalize all existing knowledge databases about business document structures, to derive a sufficiently generic model to recognize all new unknown document whatever its structure.

Ground Truth on Heterogeneous Documents

We have the access on 1 billion ITESOFT business documents per year from 600 clients worldwide.20

NLM inter-image for a pixel-based image registration

Currently Limits

❖Business documents, in particular invoices, are composed of an existing color template and an added filled-in text by the users.

❖The direct layout analysis without separating the preprinted form from the added text is difficult and not efficient.

Proposed solution

We have proposed a new approach for image registration at pixel level which can align non rigid images and tolerate spatial distortions. It is based on the NLM inter-images which aligned precisely two images with a high accuracy.

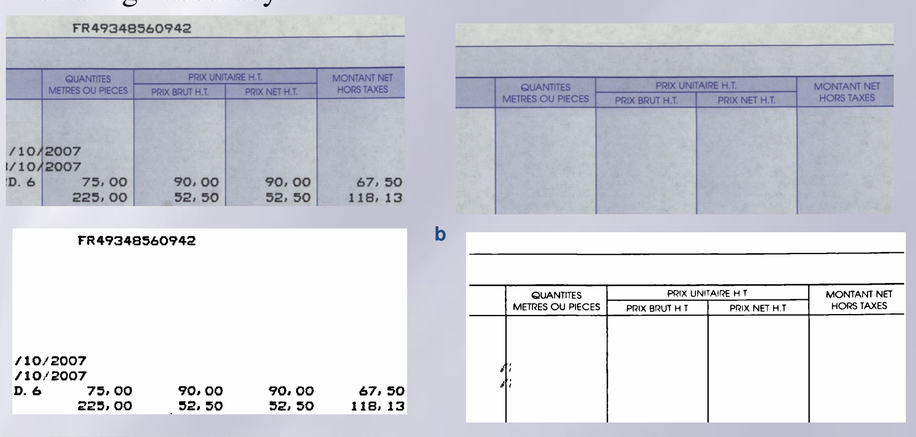


Fig2 shows the original image (a) and its model (b). The split process results into two binary image : the text added (c) and the preprint text (d). The Layout of the preprinted text in blue is displayed in the same image (e) the layout of the added text.

Currently limits of the state of the art

The complete state-of –the-art of the domain (732 references and 206 pages) is established [1,2].
Difficulties and Limit the state of the art:

❖Two trends since the last twenty years in Research : The layout segmentation of simple structures and the recognition to correct segmentation errors of the layouts. We think that the physical segmentation is inseparable from logical structure recognition and segmentation parameters must be adapted to content.

❖Developed system mainly based on specific documents and specialized recognition systems, can not be adapted to more generic documents.

❖No previous work on the recognition of completely heterogeneous documents without previous predefined model about their content.

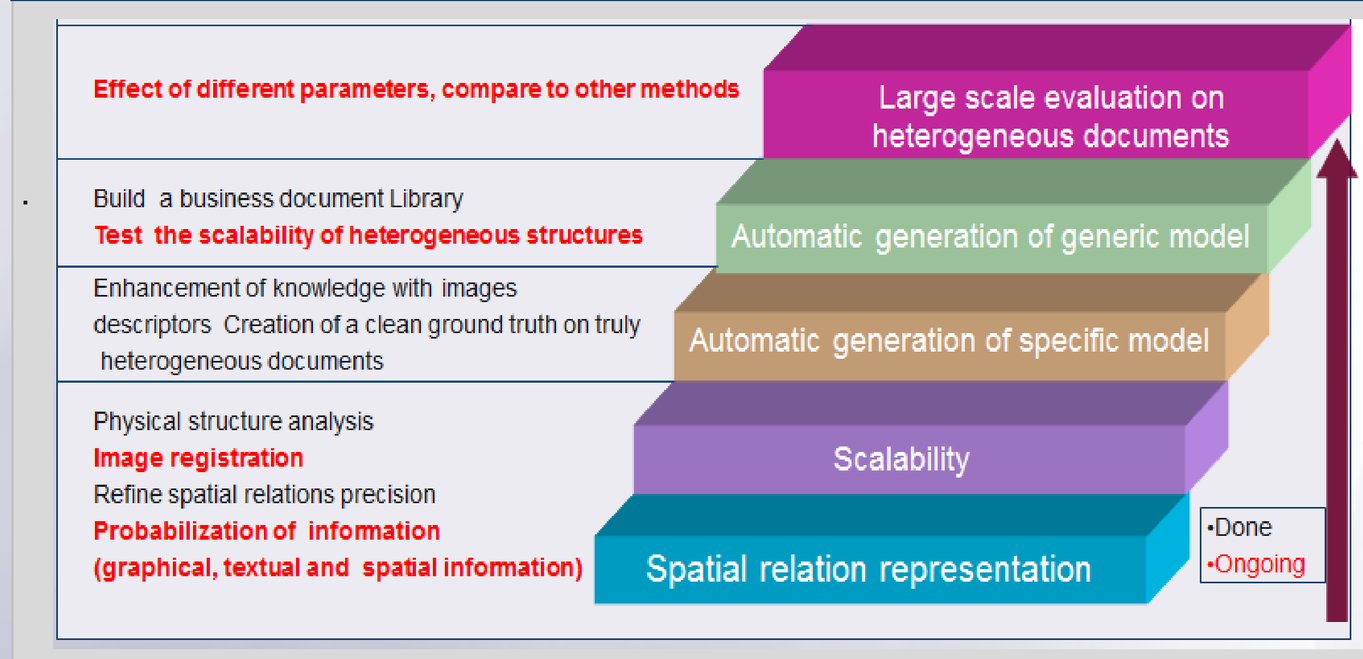
❖No previous work on large scale experimentation

❖No standardization in the organization of business documents whatever the sector of activity.

❖Confusion between documents classification and structure recognition.

❖No commercially available structures recognition systems which is completely generic.

My Ph.D. Work -Overview



Crossing the documents structures diversity

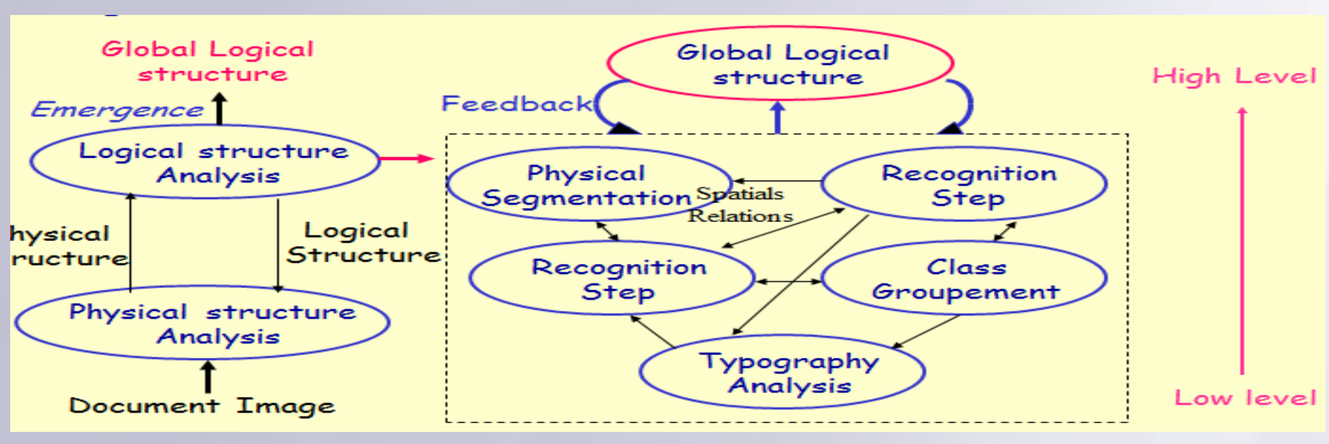


Fig1. Proposed Recognition System

Experiments

a) Automatic generation of a specific models

Test the enhancement of knowledge bases with descriptors images and data on the physical structure, improving the precision in spatial relations descriptions and the addition of probabilized information.

b) Automatic generation of a generic model for a class of documents

Test the generic model on the large class of invoices in order to measure performance against a manually generated model for this particular class of document.

c) Automatic generation of a generic model from the existing knowledge bases

Test the generic model of a database extended to other documents with diverse and varied content to test the scalability of heterogeneous structures of documents.

References

[1] L. Kessi, F. Le Bourgeois. Structure Recognition from digitized document: State-of-the-art. International Journal on Document Analysis and Recognition , 2013. (In preparation for submission.

[2]L. Kessi, F. Le Bourgeois. Structure Document Recognition: State-of-the-art. In preparation for submission.

[3] L. Kessi, F. Le Bourgeois, C. Garcia. Non Local Means Images registration : Application to Business Documents split between pre-printed forms and added text. Submitted to BMVC 2014

Contact : LOUISA KESSI (louisakessi@insa-lyon.fr) http://liris.cnrs.fr/membres?idn=lkessi