

The SCRIBO Module of the OLENA Platform: a Free **Software Framework for Document Image Analysis**

Guillaume Lazzara, Roland Levillain, Thierry Géraud, Yann Jacquelet, Julien Marquegnies, Arthur Crépin-Leblond

EPITA Research and Development Laboratory (LRDE), France

olena@lrde.epita.fr

At a Glance

- The Issue A Document Image Analysis (DIA) processing chain cannot handle all types of documents.
- The Point It is necessary to provide specific treatments for each kind of documents.
- Our Contribution A framework to design DIA software, preserving flexibility and efficiency.
- The Outcome The implementation of our proposal, the SCRIBO module, illustrates the benefits of this approach.



- adapt processing chains to specific documents.
- Efficiency Handle large amounts of documents.

Graphical Interfaces available.

Easy to integrate high-level Application Programming Interface (API) and support for various platforms.

BUSINESS LIFE

et-run of the

About 1 A.M.

ng the highw



- Provide easy-to-use applications for DIA
- Make research progress in DIA accessible to end-user applications. Using our Image Processing (IP) library in concrete use cases.

The Olena platform



Applications and Uses Cases

eds, that it took them hours to get leep. Where were w ble place to do a c as how to work the heating s as tanks and finally figured it out. Th AIRSTREAM ext morning, however. I learned that d been too slow: My 2-(, exits a mol alker, awoke with a nice head c spital in 196 rked in Red Sau lakehed of El Mirage to watch the in 1960; and settin

instead we braved the ten-mile driv the flat-screen TV

o scale back on all that is bi **MY PLEA TO ALAN MULALLY**

hen you customize them (se box). To that point, 40% (an growing) of today's Airstream



ds, that it took them hours to get was how to work the heating system. umbled with a flashlight and the outsid finally figured it out. Th wever, I learned that

84 ' F O R T U N E November 26.

Instead we braved the ten-mile driv o a waterside park in Newport Beach. Calif. And although a questionable insimple minute of it. We even forgo ause they are high-end folk art

MY PLEA TO ALAN MULALLY

In which the author begs Ford 3 CEO to produce the Ford Airslream

had the opportunity to spend an afternoon with concept actually achieved something that I honest ou customize them (se your advanced-design team and their brilliant Ford thought would never be possible. It made me desire box). To that point, 40% (and Airstream hybrid hydrogen fuel-cell concept that you what is essentially a minivan. If it can win over a growing) of today's Airstream unveiled in Detroit. I was once again struck by its family-vehicle skeptic like me, imagine hovy easy i buyers are "design aficionado

Document reconstruction in PDF.



C++ Library (Milena) More information Online demos http://olena.lrde.epita.fr/Demos Website http://olena.lrde.epita.fr/ Contact olena@lrde.epita.fr

The SCRIBO Project [1]



Project conducted in the context of the "System@tic Paris-Région" Cluster (France). ■ 9 Partners : AFP, CEA-List, EPITA, INRIA-Alpage, Mandriva, Nuxeo, Proxem, Tagmatica, XWiki.



Original document image.

fort, and its svvish yet simple interior. You may agree tem. But you must already know this. So when



culptures that represent Amer

an pride and skill. In an age

here people at the pointy end

of the earning curve are startin

to scale back on all that is big

mple life without sacrificin

GUI for DIA and reconstruction.

The SCRIBO module: a DIA Framework

Provides

- Basic routines
- Basic DIA toolchains
- Text in document
- Document layout analysis
- Text in picture
- High-level data structures
- Novel algorithms and techniques
- Standard I/O
- GUI and Command Line Interface (CLI)

Facts

- 3 years of development
- 40K lines of C++
- Open Source GPL v2
- Used in Nepomuk/KDE

Assets

- End-to-end tools \rightarrow From digital document to HTML and PDF reconstruction.
- Based on a well established IP library.

Milena: a Generic Image Processing Library [2]

■ 3 years of development.

■ Budget of 3,5M€

Provides

Data structures Safe data types

- More than 70 algorithms
- Memory management

Facts

10 years of development Version 1.0 released on July 2009 • 120K lines of C++Open Source GPL v2

References

1 SCRIBO, Semi-automatic and Collaborative Retrieval of Information Based on Ontologies. http://www.scribo.ws.

[2] Roland Levillain, Thierry Géraud, and Laurent Najman. Why and how to design a generic and efficient image processing framework: The case of the Milena library.

In Proc. of the IEEE Intl. Conference on Image Processing (ICIP), 2010.