Google Award Project

Pattern Redundancy Analysis for Document Image Indexation and Transcription

http://code.google.com/p/paradiit/

Computer Science Laboratory of Tours - France

Progess of the PaRADIIT Project

RAYAR Frédéric
2011-07-26

1. Generalities

1.1. Goal of the project

The main goal of this project is to produce a software suite, an open-source forge for RETRO and AGORA with:

- An improved clustering method (pattern redundancy analysis) ,
- An interactive and collaborative transcription system,
- And new functionalities concerning typographical studies: creation of typographical families to generate learning datasets

1.2. Human resources

- Gathering of a team for this project: professor, associate professor, postdoc, PhD students
- An R&D has been hired for this project and started in April 2011

1.3. Technical Environment

- Installation of a SVN for collaborative work (for the first work)
- Creation of a Google Code account to share the open source of the project (for the final work)
  http://code.google.com/p/paradiit/
- Creation of a Google Site (presentation, news, …)
  https://sites.google.com/site/paradiitproject/

1.4. Events

- Various article in the French press
  https://sites.google.com/site/paradiitproject/press

- Participation in the Impact Workshop in Rouen (3/31/2011): *"Recent Developments in OCR dor Digital Librairies"*

- Participation in the International Conference *Digital Humanities 2011* (June 19-22)

2. WP1 – Extraction (AGORA)

2.1. Work done since January 2011

- Specification of AGORA2011 Engine using C# language and Aforge.NET image processing library

- Implementation of Graph management of Element of Content of a document, Basic Operators of manipulation of our structure

- Unit tests for this functionalities

2.2. Work to do

- Development of the GUI

- Development of Document Image Processing Operations

- Discussions about high level scenario integration

- Software test and validation

3. WP3 – Redundancy Analysis (Clustering)

3.1. Work done since January 2011

- Specification of the pattern extraction algorithm: From bounding box to convex hull

- Specification of the features extraction algorithm:
  - Combination of Hue's and Zernike's moments
  - Computation time analysis

- Specification of the Pattern comparison algorithm
  - Distance selection: L-norm, Cosinus, Jacquard index, ...
  - Experiments to select the best one
  - Computation time analysis: need to merge the patterns in sets before feature comparison

- Specification of the clustering algorithm
  - Computation of the prototype for a set of patterns
  - Method to cluster prototypes (and then contained patterns)
  - Computation time analysis

3.2. Work to do

- Development of the specified algorithm

- Parallelization of the algorithm using ReduceMap to reduce the time of these tasks

4. WP2 – Exploitation (Retro)

4.1. Work done since April 2011

- Discussions about new usecase and scenario
- Redaction of a Software Requirements Specification Document
- Validation of a enhanced RETRO2011 Design Draft
- Choice to exploit WPF (Windows Presentation Foundation) for possible future porting of Retro as a Web service
- Research of OCROpus and Tesseract software as possible OCR Engine for RETRO2011
- Discussion about Super Resolution methods for text images regarding our application field

4.2. Work to do

- Development of a first prototype for visualization purpose
- Integration of OCR Engine for Automatic Transcription and Dictionary for Contextual Transcription
- Integration of the typology consideration to improve transcription
- Possible porting as a web service for an online-use

5. Future Works Planning

- PaRADIIT project has really started in April 2011

- For the end of 2011

September	AGORA2011 Beta
October	Clustering2011 Beta
December	RETRO2011 Beta

- For April 2012

  AGORA2011 final version
  RETRO2011 final version
  Clustering2011 final version

  Availability of open source of the project on the Google Code

- We hope the main goals will be reached in April 2012