Ecole Polytechnique de l'Université de Tours
Département Informatique
64 avenue Jean Portalis
37200 Tours, France
Tél. +33 (0)2 47 36 14 14
polytech.univ-tours.fr

Projet Recherche & Développement

2022-2023

Deep-Agora

Incremental segmentation of images of old documents

Entreprise
Centre d'études supérieures de la Renaissance

Étudiant
Théo BOISSEAU (DI5)

Tuteur académique
Jean-Yves RAMEL

Tuteur entreprise
Rémi JIMENES

24 novembre 2022

Liste des intervenants

Entreprise

Centre d'études supérieures de la Renaissance
59, rue Néricault Destouches
37020 Tours, France
cesr.univ-tours.fr

Nom	Email	Qualité
Théo BOISSEAU	theo.boisseau@etu.univ-tours.fr	Étudiant DI5
Jean-Yves RAMEL	jean-yves.ramel@univ-tours.fr	Tuteur académique, Département Informatique
Rémi JIMENES	remi.jimenes@univ-tours.fr	Tuteur entreprise

Avertissement

Ce document a été rédigé par Théo BOISSEAU susnommé l'auteur.

L'entreprise Centre d'études supérieures de la Renaissance est représentée par Rémi Jimenes susnommé le tuteur entreprise.

L'Ecole Polytechnique de l'Université de Tours est représentée par Jean-Yves RAMEL susnommé le tuteur académique.

Par l'utilisation de ce modèle de document, l'ensemble des intervenants du projet acceptent les conditions définies ci-après.

L'auteur reconnait assumer l'entière responsabilité du contenu du document ainsi que toutes suites judiciaires qui pourraient en découler du fait du non respect des lois ou des droits d'auteur.

L'auteur atteste que les propos du document sont sincères et assume l'entière responsabilité de la véracité des propos.

L'auteur atteste ne pas s'approprier le travail d'autrui et que le document ne contient aucun plagiat.

L'auteur atteste que le document ne contient aucun propos diffamatoire ou condamnable devant la loi.

L'auteur reconnait qu'il ne peut diffuser ce document en partie ou en intégralité sous quelque forme que ce soit sans l'accord préalable du tuteur académique et de l'entreprise.

L'auteur autorise l'école polytechnique de l'université François Rabelais de Tours à diffuser tout ou partie de ce document, sous quelque forme que ce soit, y compris après transformation en citant la source. Cette diffusion devra se faire gracieusement et être accompagnée du présent avertissement.

Pour citer ce document

Théo BOISSEAU, *Deep-Agora: Incremental segmentation of images of old documents*, Projet Recherche & Développement, Ecole Polytechnique de l'Université de Tours, Tours, France, 2022-2023.

```
@mastersthesis{
    author={BOISSEAU, Théo},
    title={Deep-Agora: Incremental segmentation of images of old documents},
    type={Projet Recherche \& Développement},
    school={Ecole Polytechnique de l'Université de Tours},
    address={Tours, France},
    year={2022-2023}
}
```

Document mis en forme sous LaTeX
grâce à la classe de document `polytech.cls` V2.5.2(2019/08/07)

Table des matières

Table des figures

1 Introduction

1 Actors, issues and context

The Research & Development Project (R&D project) is the final work that the student engineer must complete to obtain his diploma. It places the future engineer in a project situation by making him/her produce personal work and invites him/her to show initiative and maturity regarding a specific high-level problem. The R&D project, which lasts at least two days a week throughout the fifth year, i.e. 26 weeks, is the subject, each semester, of a dissertation and an oral presentation to a jury.

This report aims to provide both the main document that everyone can read and all the technical and methodological elements. It consists mainly of complete sections of the different documents produced, with the technical sections in the appendix.

The actors of this project are :

— the client, which here are Centre for Advanced Renaissance Studies (fr. Centre d'études supérieures de la Renaissance) (CESR), for which a contact is Rémi Jimenes, lecturer and researcher.
— the Project/Product Owner (fr. Maître d'ouvrage) (fr. MOA), who is Jean-Yves Ramel, professor of computer science, director of Laboratory of Fundamental and Applied Computer Science of Tours (fr. Laboratoire d'Informatique Fondamentale et Appliquée de Tours) (LIFAT) and academic tutor for this project. He is responsible for representing the client by ensuring that the deadlines are met and that the product conforms.
— the Project Manager / Scrum Master (fr. Maître d'œuvre) (fr. MOE), who is me, Théo Boisseau, an engineering student in his final year of study. I decide on the technical means used to design the product by what was defined by the product owner.

To convert these historical books into accessible digital libraries, LIFAT is developing image processing software that participates in a complete processing chain, including layout analysis, text/illustration separation (i.e. segmentation of content elements), optical character recognition (i.e. OCR) and text transcription. This project focuses on layout analysis and segmentation of content elements of historical documents.

The client expressed the need for easy-to-use interactive software so that its users, historians, could create their own scenarios for extracting elements of content (EOCs) from images of historical documents. These historical documents are mainly Renaissance corpora, accessible

from the CESR database, and contain mainly printed or manuscript text, illustrations and page ornaments.

The simplicity of creating extraction scenarios, their reuse and their adaptation to different documents are essential dimensions of the requirement.

However, this simplicity should not unduly compromise the reliability and performance of the software. Image processing of historical documents is a particularly difficult task notably because of broken characters, stains, and poor paper quality.

In recent years, the performance of some deep learning techniques has surpassed that of shallow methods established by experts on various image processing tasks. As this progress has made many computer vision tools available, it now seems possible to meet this need with a completely new approach.

2 Objectives

This project aims to propose a new approach based on deep learning neural networks to solve this image processing problem.

To this end, the Deep-Agora R&D project aims to build a prototype of an optimisation software capable of extracting textual and decorative elements of content from images of historical documents.

The user should not be responsible for training the models. Therefore, several deep learning models can be created and trained to extract the content elements required in the different use cases of the software.

Due to its nature as a prototype, the system will need to be composed of computational documents combining scripts and good documentation. It must also provide access to training datasets and parameter storage files to reproduce the deep learning models created.

If the objective is achieved, the project can be continued and a scenario creation subsystem can be implemented to deploy the models created within it.

3 Hypotheses

Blablabla.

4 Methodological bases

An Agile project management method will be used to create learning loops to quickly gather and integrate feedback. Therefore, the Scrum method should be preferred in which ideology is to :

— learn from experience
— to self-organise and prioritise
— to reflect on gains and losses to continuously improve

Therefore, contact with the product owner should be maintained as much as possible, as it will help me to improve and learn considerably as the project progresses.

To this end, we set sprints with a fixed duration of 2 weeks. At least one deliverable, containing an e-mail, should be sent to the product owner at least every two weeks and preferably once a week. During the implementation phase, a meeting to get feedback about the product should be scheduled at the end of each sprint.

I use GitHub for configuration management, by creating two different repositories :

— Deep-Agora, which contains the source code of the project
— Deep-Agora_DOC, which contains all the deliverables of the specification, analysis and modelling part of the first semester

As a project management tool, I will also use GitHub. As of this year, it offers a similar feature to Trello called Projects, an adaptable spreadsheet that can also integrate with my issues and pull requests on GitHub to help me plan and track my work efficiently.

? programming rules, references to supporting documents such as the quality assurance and/or test plan ?

2 Description générale

1 Environnement du projet

This project is part of a larger research project between CESR and LIFAT. It is currently being carried out as part of a programme for the regional valorisation of old books (mainly dating from the Renaissance), namely the *Humanist Virtual Libraries* controlled by the CESR.

CESR does not have powerful computing machines capable of training deep neural networks, but it has several machines and a large amount of remote and on-premises storage.

Agora, the software developed and published ten years ago by LIFAT to process images of historical documents, is undergoing a complete overhaul in this project. Its technologies need to be updated and, above all, its overhaul should meet the previously unattainable need for simplicity in scenario creation.

Therefore, no takeover of the existing system is planned, as it has to be completely redesigned.

The developer will train deep neural networks, whose task is not intended for the end users of Deep-Agora. This part of the project is to be carried out outside the software system, but within the environment, as an engineer's system.

2 Caractéristiques des utilisateurs

End users of Deep-Agora are all historians of CESR.

They have a sufficient but moderate command of computer tools. They often use them but need either extensive training or very solid documentation to use them in the case of advanced tools with complex functions. They did not have a satisfactory experience with Agora, as its interface was too complex. They do not need user access rights to use Agora.

3 Fonctionnalités du système

blablabla....

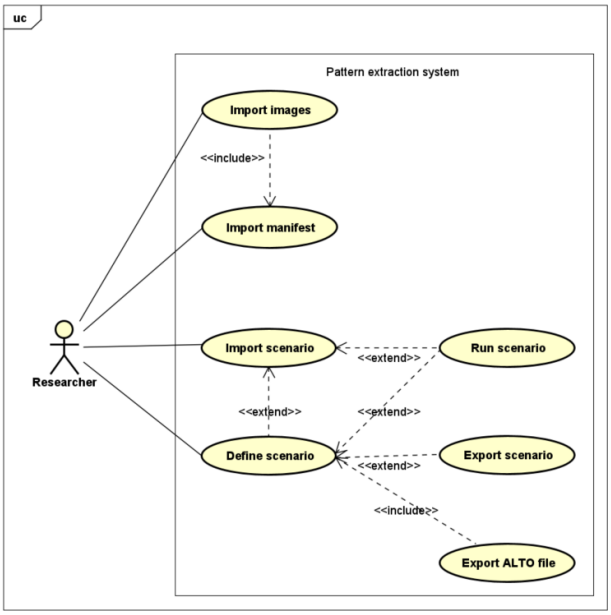


Figure 2.1 – *Use cases diagram*

4 Structure générale du système

blablabla....

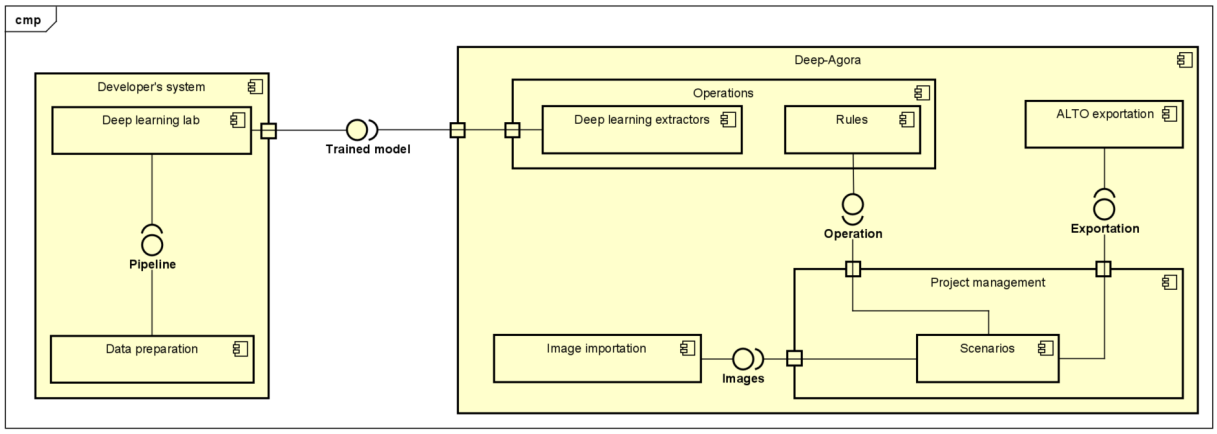


Figure 2.2 – *Component diagram*

blablabla....

3 État de l'art / Veille technologique

Sujet a définir en concertation avec votre encadrant Polytech

1 Section 1

PENSEZ à bien insérer TOUTES les références bibliographiques utilisées dans votre bibliographie.

Voici un exemple de citation : [**DBLP:journals/corr/abs-1804-02767**]. Et une autre : [**DBLP:journals/corr/RedmonDGF15**].

Paragraphe 1

blablabla

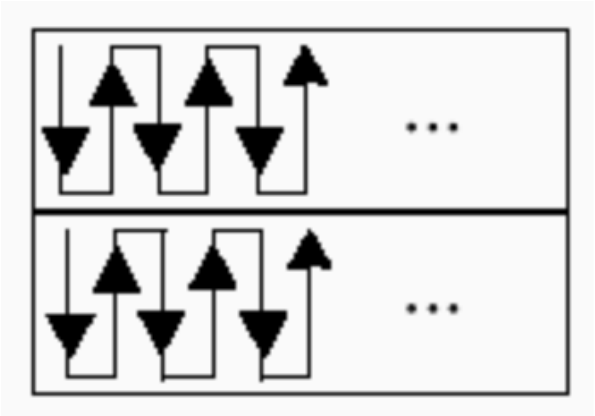


Figure 3.1 – *Fouille de données et visualisation*

Paragraphe 2

blablabla

2 Section 2

blablabla

4 Analyse et conception

1 Analyse

1.1 Hypothèses utilisées

blablabla

1.2 Spécifications

Inclure ici un résumé du cahier de spécification qui sera inséré en ANNEXE

2 Modélisation proposée

Inclure ici une description du système à développer pouvant notamment inclure les principaux diagrammes UML non détaillés et démontrant sa faisabilité durant la phase de mise en œuvre.

Les modes de validation prévus pour les différents éléments à produire pourront être précisés ici.

5 Mise en oeuvre

Description de vos productions et de leurs modes de réalisation.

(résumé du cahier de développement inséré en ANNEXE)

blablabla

1 Outils et librairie utilisés

blablabla

2 Éléments d'implémentation, choix techniques

```
#include <iostream>
using namespace std;

int main () {
    cout << "Hello, world !";
    return 0;
}
```

Un exemple de PHP :

```
class pdfOrder extends FPDF
{
 function _check($x,$y,$width,$checked) {
   if ($checked)
     $this->rect($x,$y,$width,$width,'F');
   else
     $this->rect($x,$y,$width,$width);
 }
 function LI($sansFrais = false) {
   $LI = 'LI';
   $coord = 'Laboratoire informatique
```

```
64, avenue Jean Portalis
37200 Tours
Tél. : 02 47 36 14 42
Fax. : 02 47 36 14 22';
    $this->Image(dirname(__FILE__)    . '/li.jpg',10,2,20);
    $this->SetFont('Times','B',20);
    $this->SetFont('Times','',9);
    $this->setXY(35,3);
    $this->Multicell(80,4,utf8_decode($coord),0,'LT');
 }
```

3 Analyse des résultats, évaluation, qualité

blablabla

4 Principales IHM

4.1 IHM 1

Résumé des principaux éléments présent dans le Guide de l'utilisateur avec d'éventuels compléments d'information sur leur mode de mise en œuvre.

6 Bilan et conclusion

1 Bilan du semestre 9

Liste des taches faites, en cours, à faire CF Planning S9 et S10 à fournir en annexe

2 Bilan du semestre 10

Bilan global ⇒ respect du cahier des charges (fait / à faire)

3 Bilan sur la qualité

blablabla

4 Bilan auto-critique

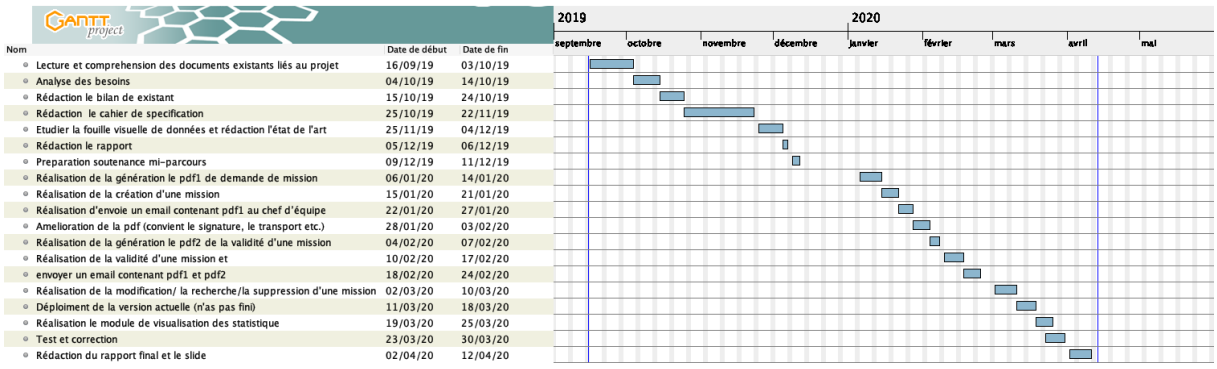
blablabla

Annexes

A Planification, gestion de projet

1 Evolution du projet

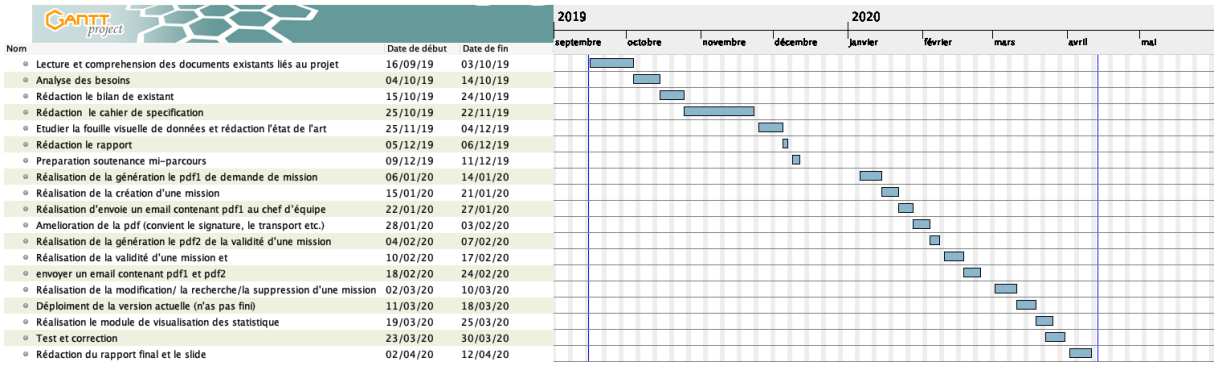
Le diagramme de Gantt Initial pour la planification de ce projet



Nom	Date de début	Date de fin
Lecture et comprehension des documents existants liés au projet	16/09/19	03/10/19
Analyse des besoins	04/10/19	14/10/19
Rédaction le bilan de existant	15/10/19	24/10/19
Rédaction le cahier de specification	25/10/19	22/11/19
Etudier la fouille visuelle de données et rédaction l'état de l'art	25/11/19	04/12/19
Rédaction le rapport	05/12/19	06/12/19
Preparation soutenance mi-parcours	09/12/19	11/12/19
Réalisation de la génération le pdf1 de demande de mission	06/01/20	14/01/20
Réalisation de la création d'une mission	15/01/20	21/01/20
Réalisation d'envoie un email contenant pdf1 au chef d'équipe	22/01/20	27/01/20
Amelioration de la pdf (convient le signature, le transport etc.)	28/01/20	03/02/20
Réalisation de la génération le pdf2 de la validité d'une mission	04/02/20	07/02/20
Réalisation de la validité d'une mission et	10/02/20	17/02/20
envoyer un email contenant pdf1 et pdf2	18/02/20	24/02/20
Réalisation de la modification/ la recherche/la suppression d'une mission	02/03/20	10/03/20
Déploiment de la version actuelle (n'as pas fini)	11/03/20	18/03/20
Réalisation le module de visualisation des statistique	19/03/20	25/03/20
Test et correction	23/03/20	30/03/20
Rédaction du rapport final et le slide	02/04/20	12/04/20

Figure A.1 – *Le diagramme de Gantt Final*

Le diagramme de Gantt Final de ce projet est comme Figure A.2.



Nom	Date de début	Date de fin
Lecture et comprehension des documents existants liés au projet	16/09/19	03/10/19
Analyse des besoins	04/10/19	14/10/19
Rédaction le bilan de existant	15/10/19	24/10/19
Rédaction le cahier de specification	25/10/19	22/11/19
Etudier la fouille visuelle de données et rédaction l'état de l'art	25/11/19	04/12/19
Rédaction le rapport	05/12/19	06/12/19
Preparation soutenance mi-parcours	09/12/19	11/12/19
Réalisation de la génération le pdf1 de demande de mission	06/01/20	14/01/20
Réalisation de la création d'une mission	15/01/20	21/01/20
Réalisation d'envoie un email contenant pdf1 au chef d'équipe	22/01/20	27/01/20
Amelioration de la pdf (convient le signature, le transport etc.)	28/01/20	03/02/20
Réalisation de la génération le pdf2 de la validité d'une mission	04/02/20	07/02/20
Réalisation de la validité d'une mission et	10/02/20	17/02/20
envoyer un email contenant pdf1 et pdf2	18/02/20	24/02/20
Réalisation de la modification/ la recherche/la suppression d'une mission	02/03/20	10/03/20
Déploiment de la version actuelle (n'as pas fini)	11/03/20	18/03/20
Réalisation le module de visualisation des statistique	19/03/20	25/03/20
Test et correction	23/03/20	30/03/20
Rédaction du rapport final et le slide	02/04/20	12/04/20

Figure A.2 – *Le diagramme de Gantt Final*

2 Description des tâches

Tâche 1 : Intitulé parlant

- Date de début : 16/09/2019
- Date de fin : 03/10/2019
- Durée : 17 jours
- Description : éléments à faire, liste des entrées (pré-requis) et sorties (livrables)s de la tache.

Tâche 2 : Intitulé parlant

- Date de début : 16/09/2019
- Date de fin : 03/10/2019
- Durée : 17 jours
- Description : éléments à faire, liste des entrées (pré-requis) et sorties (livrables)s de la tache.

B Description des interfaces

1 Interfaces matérielles/logicielles

blablabla

2 Interfaces homme/machine

blablabla

C Cahier de Spécifications

1 spécifications Fonctionnelles

1.1 Fonctionnalités à développer

1.2 Définition de la fonction 1 : intitulé parlant

Description de la fonction 1 :

Élément 1

Entrée : ? ? ? ?
Sortie : ? ? ? ?
Préconditions : ? ? ? ?
Postconditions : ? ? ? ?

Élément 2

Entrée : ? ? ? ?
Sortie : ? ? ? ?
Préconditions : ? ? ? ?
Postconditions : ? ? ? ?

1.3 Définition de la fonction 2 :intitulé parlant

Présentation de la fonction 2 :

- Nom de la fonction : Visualisation des statistiques
- blablabla
- Primordiale

Description de la fonction 2 :

blablabla

2 Spécifications non fonctionnelles

2.1 Contraintes de développement et conception

blablabla

2.2 Contraintes de fonctionnement et d'exploitation

2.2.1 Performances

blablabla

2.2.2 Capacités

blablabla

2.2.3 Contrôlabilité

blablabla

2.2.4 Sécurité

blablabla

D Cahier du développeur

1 Introduction

blablabla

2 Diagrammes architecturaux et UML

blablabla

3 Descriptions détaillées de données exploitées

blablabla

4 Descriptions détaillées des classes, modules, réalisations

blablabla

E Document d'installation

Ce document regroupe toutes les informations nécessaires pour l'installation du projet sur les machines, ainsi que pour sa mise en production.

F Document d'utilisation

blablabla

G Cahier de test

Les tests visent à garantir l'exactitude, l'intégrité, la sécurité et les performances du logiciel.

1 Tests unitaires

blablabla

IDENTIFICATION OF COMPONENT
Afficher toutes les missions de l'utilisateur identifié
DESCRIPTION OF THE TEST (granularity, scenario, values, actions)
Action :blablabla. blablabla.
EXPECTED RESULTS
Cas 1 : blblablaa. Cas 2 : blablabla.
OBTAINED RESULTS
blablabla

2 Tests d'intégration

blablabla

Projet Recherche & Développement

Deep-Agora : Incremental segmentation of images of old documents

Théo BOISSEAU

Encadrement : Jean-Yves RAMEL

université de TOURS

En collaboration avec Centre d'études supérieures de la Renaissance

Objectifs

— point 1
— point 2
— point 3

Mise en œuvre

1. point 1
2. point 2
3. point 3

Résultats attendus

Voici du texte. Voici du texte. Voici du texte. Voici du texte. Voici du texte. Voici du texte.

Ecole Polytechnique de l'Université de Tours
Département Informatique
64 avenue Jean Portalis, 37200 Tours, France
polytech.univ-tours.fr

POLYTECH TOURS
Informatique

Projet Recherche & Développement

Deep-Agora : Incremental segmentation of images of old documents

Théo BOISSEAU

Encadrement : Jean-Yves RAMEL

université de TOURS

En collaboration avec Centre d'études supérieures de la Renaissance

Objectifs

— point 1
— point 2
— point 3

Mise en œuvre

1. point 1
2. point 2
3. point 3

Résultats attendus

Voici du texte. Voici du texte. Voici du texte. Voici du texte. Voici du texte. Voici du texte. Voici du texte.

LIFAT EA 6300

LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS

LIFAT EA 6300

LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS

LIFAT EA 6300

LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS

Ecole Polytechnique de l'Université de Tours
Département Informatique
64 avenue Jean Portalis, 37200 Tours, France
polytech.univ-tours.fr

POLYTECH TOURS

Informatique

Deep-Agora

Incremental segmentation of images of old documents

Résumé

Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD. Voici le résumé de ce PRD.

Mots-clés

motcle1, motcle2, etc.

Abstract

Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project. Here is the abstract of this project.

Keywords

word1, word2, etc.

Entreprise
Centre d'études supérieures de la Renaissance

université de TOURS

Tuteur entreprise
Rémi Jimenes

Étudiant
Théo BOISSEAU (DI5)

Tuteur académique
Jean-Yves RAMEL