|  |  |  |
| --- | --- | --- |
| LOGO Peque–o | **MSc in Systems and Services Engineering for the Information Society (MSSEIS)**  Máster Universitario en Ingeniería de Sistemas y Servicios para la Sociedad de la Información (MISSSI) | E.U.I.T. Telecomunicación |

V1.2 January 2018

Master Thesis project

(Fill in only the parts in light yellow)

This document has to be registered at the students Secretariat, not later than three months in advance of the Master Thesis examination date

|  |  |
| --- | --- |
| **Title** | Dataflow Specification of a K-Means Clustering Algorithm |
| **Student** | Marta Rodríguez Ramos |
| **Student’s ID number**  (DNI, NIE, Passport No., …) | 43832935h |
| **Student’s mail** | marta.rodriguez.ramos@alumnos.upm.es |
| **Director** | Eduardo Juárez Martínez |
| **Director’s Department / Entity** | Departamento de Telemática y Electrónica |
| **Director’s mail** | eduardo.juarez@upm.es |
| **MSSEIS supervisor**  (If the director is not a MSSEIS’s member) |  |
| **MSSEIS supervisor’s Department** |  |
| **MSSEIS supervisor’s mail** |  |
| **Timeline**  (Estimated start and end dates) | 14/02/2019 – 18/06/2019 |
| **Location(s)**  (Where the work will be carried out) | CITSEM (Universidad Politécnica de Madrid) |
| **Budget in €**  (If applicable) | N/A |

If this Master Thesis is integrated in an EIT Digital MS Final Degree Project, fill in the following table the names of the people responsible of the curricular internship associated to the “Master Thesis Work Supplement” (See note (\*) below of interest to EIT Digital MS):

|  |  |
| --- | --- |
| **Professional responsible** (tutor profesional)  (Of the company or entity that hosts the curricular internship) |  |
| **Academic responsible** (tutor académico)  (Of the MSSEIS program) |  |

# Proposed board of examiners

|  |  |
| --- | --- |
| **Member 1 - President** | César Sanz Álvaro |
| **Member 2** | Antonio Carpeño Ruiz |
| **Member 3 - Secretary** | Juana María Gutiérrez Arriola |

# Description

Include a summary of the following aspects of the Master Thesis you plan to carry out: Objectives, description, methodology, tasks, materials to be used, time-schedule and bibliography. Recommended length: 2 pages. See note (\*) below of interest to EIT Digital MS.

|  |
| --- |
| **Objectives**  The main objectives of the project are as follows:   * Implement a K-Means clustering algorithm using a dataflow specification tool called PREESM * Parallelize a K-Means clustering algorithm to improve performance with regard to a serial implementation of the algorithm   **Description of the Project**  Proper and accurate detection of cancer tissues is a crucial task nowadays. There are multiple techniques used by surgeons during the operation. However, they are usually invasive and ionizing. Sometimes cancer invades surrounding healthy tissues. Therefore, during surgery, surgeons can either take some portions of this healthy tissue or leave some cancer tissue behind.  This project aims at the implementation of an unsupervised clustering method called *K-Means* on a parallel architecture in order to supply information in real time to surgeons.  The procedure is as follows; hyperspectral (HS) sensors attain hyperspectral cubes, and HS cubes are pre-processed in order to reduce dimensionality and noise. Afterwards, they are clustered employing *K-Means*, which defines different areas properly. After using this algorithm, an unsupervised segmentation map is generated.  Meanwhile in parallel, the system executes a number of algorithms belonging to supervised classification. These algorithms are PCA (Principal Components Analysis), SVM (Support Vector Machine) and KNN (K-Nearest Neighbour). After performing these algorithms, tissues are displayed using different colours in order to represent the associated classes.  Applying the majority voting, the unsupervised segmentation map obtained from K*-Means* clustering algorithm as well as the classification map obtained from supervised classification, are merged.  The implementation of this algorithm is carried out using a dataflow specification tool called PREESM. This tool is widely used for manycore architectures and signal processing applications. The objective of parallelizing this algorithm is to speed up computations for data clustering to target real time response.  **Methodology**  Hyperspectral images are processed and analysed by unsupervised as well as supervised classification methods. The algorithm of *K-Means* (unsupervised classification) determines the area of the different regions precisely and along with PCA, SVM and KNN (supervised classification) facilitates the detection of cancer tissues.  Firstly, *K-Means* is analysed and implemented using pseudo-code in order to understand it in detail. It should be noted that in contrast with all other algorithms (PCA, SVM and KNN), *K-Means* does not need prior knowledge of the data; therefore, it does not have a number of fixed steps.  Then, a serial version of *K-Means* iswritten in C language, first to recognize the most complex part of the code and detect the different functions, and then to test parallel implementation’s results.  The most complex part of the code can be parallelised in order to achieve real time response during surgery. After reviewing the code, it can be concluded that the function which computes the distances between pixels and centroids takes much more time than others in this algorithm; therefore, it would be interesting to parallelize this part first.  Dataflow specification is firstly carried out in order to ease the parallelization of this algorithm using a dataflow specification tool (PRESEEM). This tool has been selected since it represents the flow of data between different actors and it is widely used for parallelism extraction.    As was mentioned before, PREESM eases the implementation of the *K-Means* clustering algorithm on manycore architectures. The proposed requirement that constraint the system for real time processing is 60 seconds.  To check out the proper operation of the algorithm, a dataset of in-vivo hyperspectral human brain image database is used (test set), first to verify on a single core (serial code) in order to validate the results and afterwards on a manycore architecture. In other words, a set of real HS images will be employed to test and simulate the *K-Means* clustering algorithm on a manycore architecture to reach real-time processing during surgical operations.  **Tasks**  This Master Thesis has been divided into several tasks:  Task 1   * State of the art   Hyperspectral images  PCA,SVM and KNN algorithms  K-Means algorithm  K-Means on parallel architectures (OpenMP, Cuda, OpenCL)  Task 2     * Tutorials of PREESM tool   Tutorial Introduction  Parallelize an Application on a Multicore CPU   * Implementation of K-Means pseudo-code   Task 3   * Tutorials of PREESM tool   Code generation for Multicore CPU   * K-Means clustering algorithm in C (serial version) * Analysis of the parallelism of the algorithm   Task 4   * Tutorial of PREESM tool   Software Pipelining for throughput Optimization  Memory Footprint Reduction  Advanced Memory Footprint Reduction   * Design of the K-Means clustering algorithm dataflow graph, architecture graph and scenario on a single core using PREESM   Task 5   * Tutorial of PREESM tool   Automated actor execution Time measurement  Spider  Papify  HW/SW Code Generation for ZedBoard  Throughput Evaluation for Hierarchical Graphs   * Design of the K-Means clustering algorithm dataflow graph, architecture graph and scenario on a single core using PREESM   Task 6   * Parallelize the algorithm using PREESM * Write the report of the Master Thesis   Task 7   * Finalize Master Thesis and tests   **Materials to be used**   * PREESM Open Source Software * Libraries CMake, PThread and SDL2 and SDL2\_ttf * SVM,PCA and KNN algorithms implemented in PREESM * K-Means C Serial Version   **Time Schedule**    In order to follow a schedule and meet all requirements set out in the project, the duration of each task is as follows:   * Task 1: 14/02/2019 - 28/02/2019 * Task 2: 28/02/2019 - 14/03/2019 * Task 3: 04/03/2019 - 18/03/2019 * Task 4: 11/03/2019 - 01/04/2019 * Task 5: 16/03/2019 - 21/04/2019 * Task 6: 21/04/2019 - 25/05/2019 * Task 7: 25/05/2019 - 18/06/2019   **Bibliography**  **PCA, SVM and KNN using PREESM**  [1] <https://gitlab.citsem.upm.es/dmadronal/HSI_cancer_detection>  **PREESM Tutorials**  [2] <https://preesm.github.io/tutos/intro/>  **K-Means C Serial Version**  [3] <https://drive.upm.es/index.php/s/K8ZmqNOivQoOxKE>  **K-Means**  [4] Emanuele Torti , Giordana Florimbi, Francesca Castelli, Samuel Ortega, Himar Fabelo , Gustavo Marrero Callicó, Margarita Marrero-Martin and Francesco Leporati. (2018). Parallel K-Means Clustering for Brain Cancer Detection Using Hyperspectral Images. *Electronics. 7,* 283.  **Manycore Arquitectures**  [5] R. Lazcano, D. Madroñal, R. Salvador, K. Desnos, M. Pelcat, R. Guerra, H. Fabelo, S. Ortega, S. López, G. M. Callicó, E. Juárez and C. Sanz. (2017). Porting a PCA-based Hyperspectral Image Dimensionality Reduction Algorithm for Brain Cancer Detection on a Manycore Architecture. *Syst. Archit, 77, 101–111.*  [6] R. Lazcano, D. Madroñal, H. Fabelo, S. Ortega, R. Salvador, G.M. Callicó, E. Juárez and C. Sanz. (2017). Parallel implementation of an iterative PCA algorithm for hyperspectral images on a manycore platform. *Conference on Design and Architectures for Signal and Image Processing, DASIP.*  [7] D. Madroñal, R. Lazcano, R. Salvador, H. Fabelo, S. Ortega, G. M. Callico, E. Juarez, C. Sanz. (2017). SVM-based real-time hyperspectral image classifier on a manycore architecture. *Journal of Systems Architecture. 80,* 30-40*.*  [8] D. Madroñal, R. Lazcano, S. Ortega, H. Fabelo, R. Salvador, G. M. Callicó, E. Juárez and C. Sanz. (2017) .Implementation of a spatial-spectral classification algorithm using medical hyperspectral images. *32nd Conference on Design of Circuits and Integrated Systems (DCIS).*  **S-LAM Architecture**  [9] Maxime Pelcat, Jean François Nezan, Jonathan Piat, Jerome Croizer, Slaheddine Aridhi. (2009). A System-Level Architecture Model for Rapid Prototyping of Heterogeneous Multicore Embedded Systems. *Conference on Design and Architectures for Signal and Image Processing (DASIP)*, Nice, France. 8 p.  **Master Thesis**  [10] Lazcano López, Raquel. (2015). *Parallelization of the PCA algorithm on the Kalray MPPA-256 Platform*  (Master Thesis). Universidad Politécnica de Madrid.  [11] Madroñal Quintín, Daniel. (2015)*. Implementación de una Support Vector Machine en RVC-CAL para imágenes hiperespectrales* (Master Thesis). Universidad Politécnica de Madrid.  **Papers**  [12] D. Madroñal Quintín, R. Lazcano López, E. Juárez Martínez, C. Sanz Álvaro. (2015). Dimensionality reduction and endmember extraction for hyperspectral imaging using an RVC-CAL library. [*SPIE Remote Sensing*](https://www.spiedigitallibrary.org/conference-proceedings-of-spie/browse/SPIE-Remote-Sensing/2015)*, Toulouse, France.*  [13] R. Salvador, H. A. Fabelo, R. Lazcano, S. Ortega, D. Madroñal, G. Marrero, E. Juárez, C. Sanz. (2016). HELICoiD tool demonstrator for real-time brain cancer detection. *Conference on Design and Architectures for Signal and Image Processing (DASIP).*  [14] D. Madroñal, R. Lazcano, H. Fabelo, S. Ortega, G. M. Callicó, E. Juárez, C. Sanz. (2016). Hyperspectral image classification using a parallel implementation of the linear SVM on a Massively Parallel Processor Array (MPPA) platform. *Conference on Design and Architectures for Signal and Image Processing, DASIP, 154-160.* |
|  |

# Internship and Master Thesis relationship

(This section is of interest only to EIT Digital MS Final Degree Project). Include a ½ page description of the Internship and Master Thesis relationship, covering the following: Internship and Master Thesis Industrial context; Internship and Master Thesis topic description (theoretical framework); How the Internship and Master Thesis are connected; How will you be exposed to the reality of professional life outside university. See note (\*) below of interest to EIT Digital MS.

|  |
| --- |
|  |

# Dates and signatures

|  |  |  |
| --- | --- | --- |
|  | **Date** | **Signature** |
| **Student** | 15/04/2019 |  |
| **Director** |  |  |
| **MSSEIS supervisor**  (If different from the Director) |  |  |

# Approval from the program responsible

|  |  |  |
| --- | --- | --- |
|  | **Date** | **Signature** |
| **Program responsible** |  |  |

Note (\*): Information of interest to EIT Digital MS:

|  |
| --- |
| EIT Digital MS students have to fulfil a 30-ECTS Final Degree Project (FDP), mainly during their spring semester, which has to be both scientifically as well as industrially relevant. This FDP is a combination of a Master Thesis and an Internship, both within the same thematic focus.  In MSSEIS program, this FDP is implemented through two spring semester courses:   * Master Thesis (15 ECTS). * Master Thesis Work Supplement (15 ECTS). This is done through the curricular internship modality (*prácticas externas curriculares*) of UPM’s COIE (*Centro de Orientación e Información de Empleo*). The internship has to be long enough to cover 15 ECTS.   In consequence, in the case of a FDP for an EIT Digital MS student, you have to fill in these two sections of this document:   * “Description”. See the aspects to be covered in this section above. * “Internship and Master Thesis relationship”, with information that the EIT Digital MS asks to the students as part of their proposal. See the aspects to be covered in this section above. |