DEBREMARKOS UNIVERSITY

Industrial project proposal

Ge’ez handwritten character recognition system with machine learning

PREPARED BY

Eba alemayehu

Software engineering academic program

Content

1. Introduction

1.1. Background of the project

Ge’ez is liturgical language of the Ethiopian church. Geʿez is a Semitic language of the Southern Peripheral group, to which also belong the South Arabic dialects and Amharic, one of the principal languages of Ethiopia. Ge’ez has its own writing style and alphabet. Both Geʿez and the related languages of Ethiopia are written and read from left to right, in contrast to the other Semitic languages. Extinct as a vernacular language, Geʿez is the ancestor of the modern Tigrinya and Tigré languages of Eritrea and Ethiopia. The oldest known inscription in the language dates from the 3rd or 4th century and is written in a script that does not indicate vowels.

Most of Ethiopian history and documentations were written in Ge’ez characters. Either with ge’ez language itself or other decendent languages of ge’ez like amharic. Before the era of computers and automation the government and other organizations used handwritten documents. This documents contain accumulated wisdom of our forefathers, the history of our country, the history of gov’t records like police records etc.. Therefore we need some automated way to help us digitize these documents in order to make storage and distribution of these documents match more easier.

1.2. Overview of existing system

Whenever we want to change a handwritten text in to computer understandable (editable text). We have to type the text again in to a computer system. People usually higher typists to do this work. In press industry usually authors like writing in handwriting. When they are done with writing they take their work to the typists to type done to a computer. Also media reporters take notes one some scenarios and type text to make news. We can mention a lot of areas where typing is used to change handwritten text to a computer.

Though there exist some character recognition system and a very few researches on handwritten amharic character recognition systems, we don’t think there still exist any amharic handwritten character recognition system yet. Especially systems that are available for developers to work on more application systems which are based on this character recognition systems.

1.3. Statement of the problem

Typing ge'ez characters into computer is relatively harder because the standard QWERTY keyboard which most of us use is not designed for amharic language. It also requires a lot of labour. It is a very time consuming task. These days people tend to use mobile or tablet devices rather than the conventional desktop computers. These devices have no convenient way of writing a long text. Especially writing amaharic text is much more difficult because though there exist few applications which allow as to write ge'ez there is no standard for the key layout.

On the other hand artificial intelligence and fully intelijent systems are growing. This systems are expected to dominate the world. This kind of systems incorporate artificial general intelligence which means they almost mimic a human mind. One of the issues in the part of this big general system call machine learning is the issue of diversity. Machine learning is an algorithm which takes data and tries to learn from that data. So that if it will not be trained with our language we will be left over in the digital divide. Which negatively affect us and our language. Therefore we need to develop aI systems which can mimic our culture and our language. Reading or recognizing character are one of the skills of language we need to train machines.

Humans can easily recognize characters one they have learned them in spite of how distorted they are but identifying handwritten characters is not an easy computer vision task in a traditional way of programming, because it would be impossible to be able to write rules about how each character are represented. Humans write characters in unpredictable way, style of writing differs from person to person. Therefore we need to use another approach which is machine learning. This puse a big computer since challenge but recently machine learning had become good at this kind of computer vision tasks because know we have good enough computational power to teach computers to detect this unpredictable writing of characters.

1.4. Literature review

Scholars have tried to solve this problem before. We have tried to review some of them hear:

1.5. Overview of proposed system

The proposed system is a trained machine learning model which can recognize any handwritten character recognition. The system uses neural networks in order to achieve this task. The network will be a classifier network.

Ones we could train a model with a satisfactory accuracy level we will make different applications and interfaces which uses this model. This includes a web app, mobile app, desktop app and an API for developers. There are a lot of applications which can be built on this model. Some of them are:

- Amharic writing learning app which teaches how to write amharic characters
- Amharic road signs reading and translation app
- Amharic optical character recognition systems

1.6. Objective of project

1.6.1. General objective

To build a computer system that can recognize any handwritten ge’ez characters.

1.6.2. Specific objective

The specific objective of this project are:

- Preparing ge’ez characters dataset which is publicly available and anyone who want to experiment on it can try out.
- To find the best learning algorithm and neural network architecture
- To train a model which can classify amharic characters
- To make application softwares which make use of the trained model

1.7. Scope of the project

This project consist of:

- Ge’ez character dataset collection
- Designing the learning algorithm and architecture
- Training the model with the collected dataset and Designed architecture
- Building an application on the top of the trained model
- Different small programs that help to automate some tasks on the process for instance data collection.

This project do not include:

- Any natural language processing. Our system does not understand the meaning of a text.
- No semantic analysis or data mining on text is done.
- Word or sentence based recognition. Our system is character based recognition.

1.8. Significance of the project

The significance of the project can be seen in different dimensions. On one side the system we are going to build an application which can solve some problems. It includes API that developers we amazing idea can build applications. On the other side when we see the big picture it could be one step forward for next AI projects that can be done with our juniors.

1.9. Tools and methodology

1.9.1. Data Collection methodology

To train our model we need a lot of data. We are planning two ways to collect data.

I. By distributing questionnaire paper to different people to get there handwriting.
II. Building an android app which help us collect characters data. The app will have a canvas to enable draw characters on screen.

1.9.2. Technologies to be used

I. Programming languages

- ➔ Python
- ➔ Javascript
- ➔ Java, swift or dart (optional)
- ➔ Html and css

II. Tools and technologies

- ➔ Opencv
- ➔ Tensor flow
- ➔ Flutter
- ➔ Keras
- ➔ Numpy
- ➔ Matplotlib
- ➔ django or flask

1.9.3. System requirements hardware and software

I. Operating system

- ➢ **Linux**: will used as the development and training operating system
- ➢ **Windows**: for documentation and some drawing
- ➢ **Mac os**: will be used for compiling ios apps

II. **Softwares**

- Android studio for android development
- Visual studio code as a text editor
- Apache or nginx serves
- Google chrome for testing and debugging javascript

III. **Hardwares**

- Two computers one for training the model one for a development. The specification for these two computers are listed as follows.
    A. Training server: core i7 processor, 16GB RAM 1 TB storage with GPU capability.
    B. Development pc: core i7 processor, 8GB RAM 1 TB storage.
- Android and iphone devices for testing
- Printed paper for data collection
- Other office apparatus for different purposes

1.9.4. System modeling tools

- Microsoft visio and project
- StarUML as modeling tool

1.10. Feasibility study

1.10.1. Technical

Thanks to the big improvements that deep learning bring into the computer vision world, we think the system is technically feasible with the resources we have now.

1.10.2. Operational

Through neural networks are relatively expensive in terms of computational resources we think it is feasible to operate our system in a computational power we have today

1.10.3. Economical

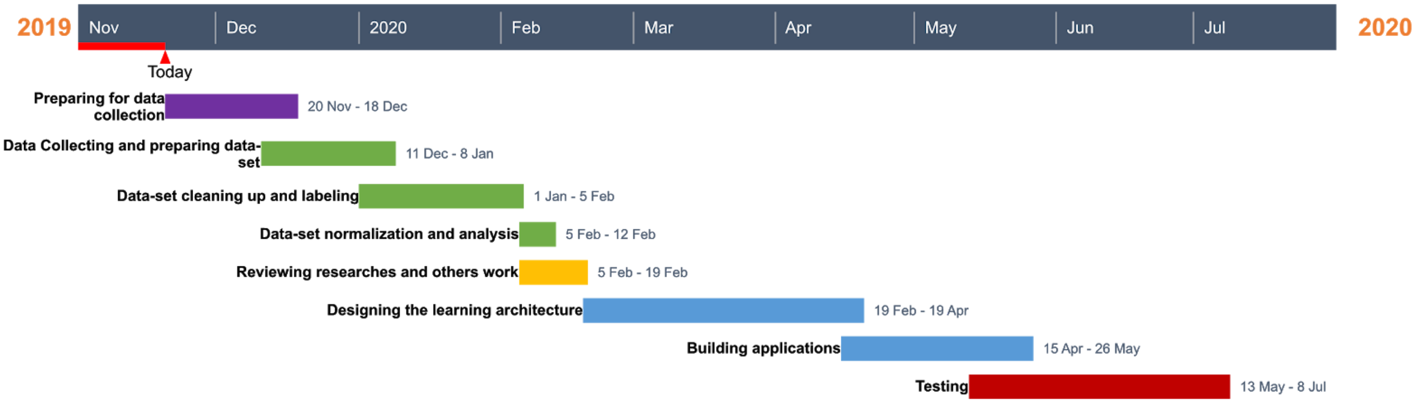
As we have mentioned earlier on the problem statement typing characters are time and resource consuming this project will allow our users to save a lot of time and resource so it is economically feasible.

1.11. Budget plan

Item	Measurement	Quantity	Unit price	Total
Computers	each	2	25,000 birr	50,000 birr
Copying paper	each	350	1 birr	350 birr
Sticky notes	each	2	25 birr	50 birr

1.12. Work breakdown

In general our work starts from collecting data and getting a pretty good data-sets. We will be dealing with the data set on the first phase (first semester). Data collection will be the next phase or the last semester.



Reference:

- https://www.britannica.com/topic/Geez-language#targetText=Ge%CA%BFez%20language%2C%20also%20spelled%20Geez,the%20principal%20languages%20of%20Ethiopia.