Arabic Islamic Manuscripts Digitization based on Hybrid K-NN/ SVM Approach and Cloud Computing Technologies

Our idea consists on:

- On the first hand, to consider cloud computing as an infrastructure (IaaS) to deploy our combination of algorithms K-NN/SVM for Arabic Islamic Manuscripts Recognition System AIMRS.
- On the second hand, to consider cloud Storage as a Service (SaaS) to store and retrieve large amounts of Arabic Islamic Manuscripts.

integration and cooperation of some strong complementary approaches. In addition, our approach offers a number of benefits, such as **the ability to store and retrieve large amounts of Islamic Manuscripts**, **fast processing**, **fast data access**, and **unlimited storage.**

paper components:

- general introduction to the K-NN and SVM algorithms and especially the hybrid approach KNN/SVM and the use of this algorithm in the Arabic pattern recognition system.
- Cloud computing and Cloud Storage.
- the distributed hybrid approaches are discussed.
- The design of the experiments, experimental results, and discussions
- The conclusions and future work

K-NN/ SVM

- K Nearest Neighbor (K-NN)
  KNN is an instance-based classification algorithm. The main idea of this classifier algorithm is quite straightforward. In order to classify a new character, the system finds the k nearest neighbors among the training data sets and uses the categories of the k nearest neighbors to weigh the category candidates.

- Support Vector Machine (SVM)
  SVM is a new type of pattern classifier. This promising classification technique is based on a novel statistical learning approach.

Cloud computing

Cloud computing technology is evolving as a key computing architecture for sharing resources that includes three types of resources infrastructures, software, and platform. Virtualization is a core technology for enabling cloud resource sharing.

Cloud Storage

The Cloud storage technology this intelligent storage solution is becoming an efficient business paradigm that combines software and industry-standard converged storage servers to deliver high-speed access in a much more modular and scalable solution.

The concept of a virtualized environment is introduced to decrease the existing storage inefficiencies. This intelligent storage solution offers compelling benefits and a way for the next-generation data center to address three important storage challenges: increasing the volume of data, facilitating the data management, and decreasing the cost, compared with traditional storage techniques that are often built around extremely costly, powerful and in some cases proprietary hardware.

THE PROPOSED APPROACH:

Our main idea is attending three objectives:

1. increase the recognition rate,
2. speed up the pattern recognition system
3. decrease the cost of management, handling, and archive backup, processing data.

To attend the first objective, we propose a hybrid approach K-NN/SVM similar to Bellili et al. This approach consists of using SVM as a decision classifier to exceed the limits of K-NN (sensible to different classes with similar attributes)

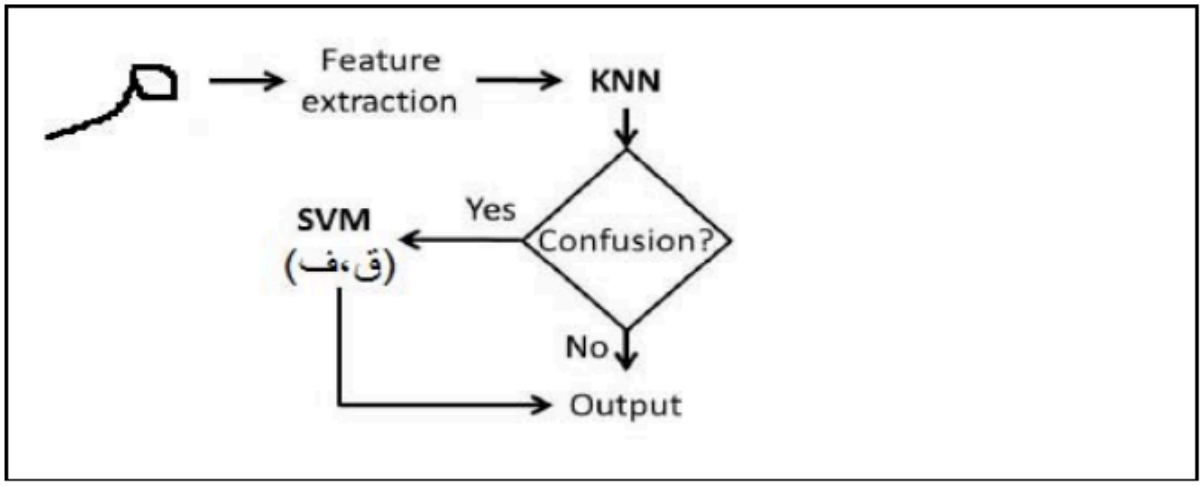

Figure 1. K-NN /SVM mechanism

to attend to the two other objectives, we propose to virtualize and distribute our Islamic Manuscript recognition application using distributed platforms such as cloud computing technologies.

Mapreduce, Hadoop, and cascading are used to manage, handle, and map the OCR application. S3 (Simple Storage Service) to manage the different data base (test and reference database) and the output.

We propose to use the master-slave model and SPMD (Single Process, Multiple Data) techniques to distribute our OCR application and data set to be recognized on a distributed-memory multiprocessor system. In this approach, each copy of the single program runs on processors independently and communication is provided by Hadoop. The big amounts of the document to recognize are better split into small parts (D1, D2, D3 …Dn) and assign each one to a slave node to achieve the recognition task. The OCR application will be implemented by a job flow for each node.

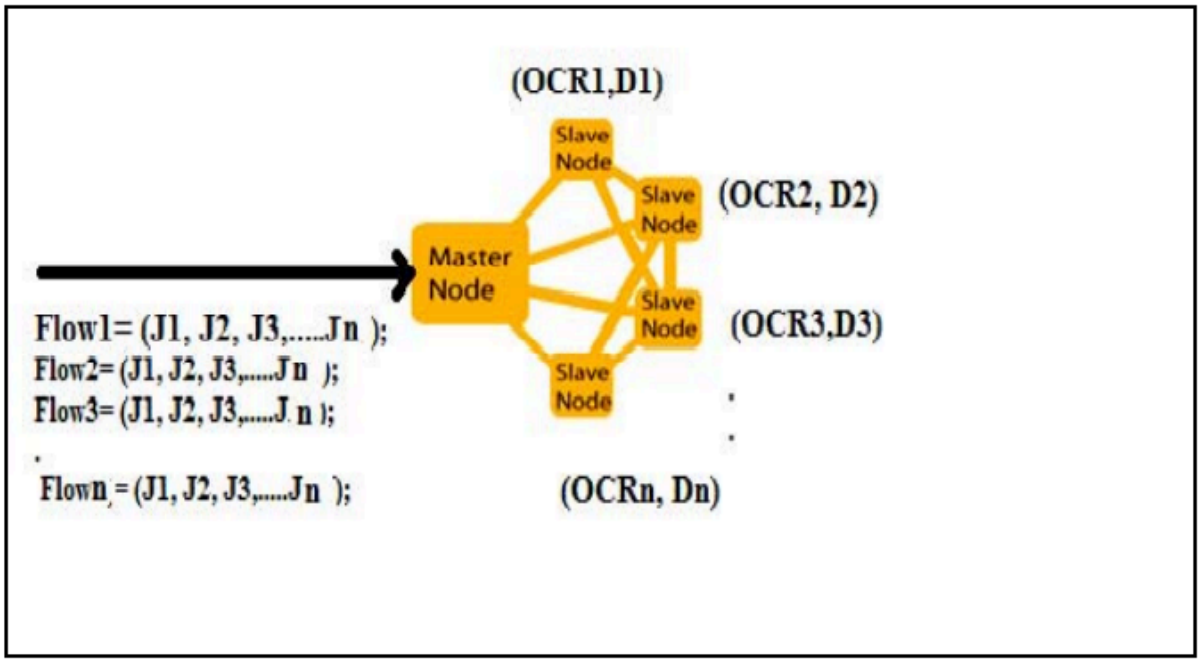

Figure 2. Example of job flow via Mapreduce

The experimental study

Datasets:

To evaluate the proposed approach we have used a corpus with **16000 pages** (**370 characters/page**) randomly chosen from the **Islamic Heritage Project (IHP)** of, Harvard University.

Harvard's Open Collections Program (OCP) has produced online digital copies of over 280 manuscripts that date from the 10th to the 20th century CE.

For the preprocessing image, the data set needs to be normalized where the OCP Dataset provides normalized images. Cropping, Filtering, and normalization are applied to our database.

For the segmentation step, Horizontal projection is applied to detect lines then each line into words using a drawing rectangle around each word and each word segmented into its primitives using vertical projection. Wavelet transform is used as a feature extraction technique. We have considered also a reference library composed of 345 characters representing approximately the totality of the Arabic alphabet (including the character's shape varies according to their position within words and with different positions (rotation and translation)).

Experimental environment:
The tests were conducted on a local Intel Core 2 Duo desktop having the configuration: 3.00 GHz *2, 2 GB of RAM running a Windows XP operating system. To run the Linux command, the shell Cygwin [22] is used. Java, JDK 1.6, and Eclipse 3.4 were used to program, implement, and build our OCR application.
Cascading and Hadoop were used to manage our application in a distributed environment. 100 Mbits/s was the network capacity.

Results and discussions:
To evaluate:
On the first hand the importance of the hybrid approach K-NN/SVM in the recognition rate
On the second hand the efficiency of the cloud computing technologies on the timely execution of the proposed hybrid K-NN/ SVM approach
we created six running Jobs flow in cascading on Amazon EC2 EMR clouds and conducted two comparison experiments on both **K-NN (with K=15)** and **K-NN/SVM** speedup in three instances of Amazon Elastic Computing Cloud service.

K-NN AND K-NN/SVM RECOGNITION RATE (%)
Average of K-NN 95.94%
Average of K-NN/SVM 96.92%

K-NN AND K-NN/SVM TIME EXECUTION (H)

Instance	Number of cores	K-NN(h)	K-NN/SVM(h)
Small instance of Amazon Elastic Computing	25	0.541	0.601
	50	0.420	0.501
	75	0.300	0.401
	100	0.200	0.387
Medium instance of Amazon Elastic Computing	25	0.500	0.514
	50	0.410	0.450
	75	0.250	0.300
	100	0.115	0.220
Large instance of Amazon Elastic Computing	25	0.400	0.420
	50	0.301	0.350
	75	0.115	0.200
	100	**0.126**	**0.140**

The average test time for **one computer** using K-NN and K-NN/SVM is **approximately 7 hours and 8 hours.**

the average test time for **100 computers** is **0.126 hours** and **0.140 hours** for K-NN and K-NN/SVM algorithms respectively.