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

Hamdi Hassen[1,3]
[1]Computer science departement,
College Of Science And arts at Al
Ola, Taibah University, KSA.
[3]Mir@cl Lab, FSEGS
University of Sfax
BP 1088, 3018 Sfax, Tunisia
(216) 74 278 777
hhassen2006@yahoo.fr

Maher Khemakhem[2,3]
[2]Computer Science Department,
Faculty of Computing and
Technology,
University of King Abulaziz, KSA
[3]Mir@cl Lab, FSEGS
University of Sfax
BP 1088, 3018 Sfax, Tunisia
(216) 74 278 777
maher.khemakhem@fsegs.rnu.tn

***Abstract*— In many national libraries and archive centers, most of Islamic Manuscripts are still in their initial form and not digitized yet. These documents are indeed very rich in knowledge and constitute a part of the heritage of Muslims.**

The main reasons behind the digitization of these Islamic Manuscripts are, consequently and at least, to enhance and ease their accessibility by people who are interesting in order to exploit the corresponding knowledge and to improve their durability.

Such a mission is not yet easy to achieve because of the weaknesses of the existing approaches and algorithms. Indeed, several researchers around the world have proposed a variety of approaches, algorithms and techniques in order to build a powerful system able to digitize such documents. Unfortunately, these efforts didn't succeed since they realized that such problem cannot be solved without the integration and cooperation of several strong complementary approaches, algorithms and techniques at the same time. But, such an idea and a system requires at the same time the knowledge of the good complementary approaches, algorithms and techniques which can lead to an acceptable recognition rate of the Arabic handwriting in one hand, and the adequate hardware infrastructure which can host such complex and greedy software to achieve the mission in a reasonable time on the other hand.

Our idea consists on: in 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. In the second hand, to consider cloud Storage as a Service (SaaS) to store and retrieve large amounts of Arabic Islamic Manuscripts.

Our approach provides indeed an adequate platform for the expected powerful digitization system based on the 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, the fast processing, the fast data access, and the unlimited storage.

Keywords — Arabic Islamic Manuscripts, Cloud Computing, Pattern Recognition System

I. INTRODUCTION

The Manuscript literature of the Islamic world is a vast area of study. The Islamic Manuscripts contain an as yet almost untapped source of the rich Islamic heritage. Islamic Manuscripts have been studied for quite a while and many are well-known. However, even more are as unknown or at least insufficiently appreciated yet.

The combination of the multiple methods produce the most precise final result because it exploited the advantages of each one. This combination seems to constitute an interesting approach in Arabic Islamic Manuscripts digitization.

The major problem of this hybrid approach is the need of a large computing and storage capacity.

Distributed platform such as cloud computing and exactly map reduce, hadoop, cascading and S3 can be a solution to resolve the problem.

Today's cloud computing is primarily used to deliver infrastructure, platform, and software as services, which are made available as subscription-based services in a pay-as-you-go model to consumers. These services in industry are respectively referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

978-1-4799-2823-1/15 $31.00 © 2015 IEEE
DOI 10.1109/NOORIC.2013.77

The remainder of the article is organized as follows: in section II, we give a general introduction to the K-NN and SVM algorithms and especially hybrid approach KNN/SVM and the use of this algorithm in Arabic pattern recognition system. Cloud computing and Cloud Storage are presented in section III. In section IV, the distributed hybrid approaches are discussed. The design of the experiments, experimental results and discussions are presented in section V. The conclusions and future work are discussed in section VI.

II. K-NN/ SVM

A. 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 weight the category candidates.

The K-NN algorithm can be described using the following equation:

$$C(x) = \arg\max_{c\in C} \sum_{i=1}^{k} \sigma(c, c(y_i)) \qquad (1)$$

Where

c(y_i) is the class of y_i

σ is a function that σ (u,v) = 1 if u = v.

KNN can be also viewed as a probability-based classifier, shown in equation 2.

$$c(x) = \arg\max_{c\in C} \bar{P}(C|x)) \qquad (2)$$

Where

$\bar{P}(C|x))$: Is the fraction of instance of class C in the nearest neighbors, shown in the following equation:

$$\bar{P}(C|x)) = \frac{\sum_{i=1}^{k} \sigma(c, c(y_i))}{k} \qquad (3)$$

Without prior knowledge, most K-NN classier apply euclidean distances as the measurement of the closeness between examples. Since it has already been shown that treating the neighbors that are of low relevance as the same importance as those of high relevance could possibly degrade the performance of K-NN procedures [1].

Many researchers have found that the k-NN algorithm achieves very good performance for character recognition in their experiments on different data sets [2].

K-NN represents a very intersecting classifier for Arabic handwriting recognition because of their great adaptability and versatility in handwriting sequential signals, but K-NN recognition rate is sensible to different classes with similar attributes [3].

B. Support Vector Machine (SVM)

SVM is a new type of pattern classifier. This promising classification technique is based on a novel statistical learning approach [4].

SVM have been successfully applied in different applications such as face detection, face recognition object detection, object recognition, handwritten character and digit recognition [5, 6, 7].

SVM is the technique used to minimize the risk [8] that can be described using the following equation:

$$R = \frac{1}{l} \sum_{i=1}^{l} |f(xi) - yi| \qquad (4)$$

Where

R: The empirical Risk,

l: The size of the class,

f: The classification function decision.

III. CLOUD COMPUTING AND CLOUD STORAGE: AN OVERVIEW

A. 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[9]. Virtualization is a core technology for enabling cloud resource sharing. Infrastructure as a Service (IaaS) is the first type of resources which include computing power, storage, and machine provisioning such as the infrastructure provider Amazon EC2 which is a web service interface that easily request and configure capacity online [10]

Software as a Service (SaaS) is the second type of resources in Cloud computing technology including middleware and development resources. In the first hand, the middleware consists of cloud centric operating systems, application servers, databases, and others. In the second hand, the development resources comprehend design platforms, development tools, testing tools, deployment tools, and open sources-based reference projects.[10]

Platform as a Service (PaaS) is the third service in a cloud computing technology that delivers an application development and deployment platform as a service to developers over the Web [10]

B. *Cloud Storage*

The volume of digital structured or unstructured content such as Islamic Manuscripts continues to increase putting heightened pressure on storage capacity and data center networks [11]. More efficient storage techniques are an imperative data must be available to meet business, regulatory, and compliance need.

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

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

Amazon S3[12], Gigaspaces [13], ElephantDrive [14], are small companies (providers) that offer large Web applications and consequently, can avoid large capital expenditures in infrastructure by renting distributed storage and pay per use. The storage capacity employed may be large and it should be able to further scale up. However, as data scales up, hardware failures in current datacenters become more frequent [15]

III) THE PROPOSED APPROACH

There are many problems that hindering the task of Arabic handwritten character recognition and especially Islamic Manuscripts such as the similarity between different characters (for example, ه and م, ر and و, among others) causing confusion at moment of classifying such patterns and the large amount of Manuscrits to recognize.

Our main idea is attending three objectives: increase the recognition rate, speedup the pattern recognition system and decrease the cost of management, handle, and archive backup, processing data.

Hybrid approaches have been applied widely in pattern recognition system to adapt the increasing requirement of real-time high accuracy and robustness. There are a lot of successful hybrid image recognition algorithms.

To attend the first objective, we propose a hybrid approach K-NN/SVM similar to Bellili et al. This approach consists on using SVM as a decision classifier to exceed the limits of K-NN (sensible to different classes with similar attributes). Fig.1.

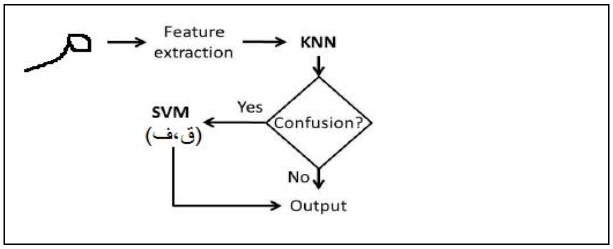


Figure 1. K-NN /SVM mechanism

In order to attend the two other objectives, we propose to virtualizes and distribute our Islamic Manuscript recognition application using distributed platform such as cloud computing technologies. Mapreduce [16], Hadoop [17] and cascading[18] are used to manage, handle, and map the OCR application. S3 (Simple Storage Service) to manage the different data base (test and reference data base) and the output.

We propose to use the master–slave model and SPMD (Single Process, Multiple Data) technique 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 document to recognize is better to split it 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. Fig.2.

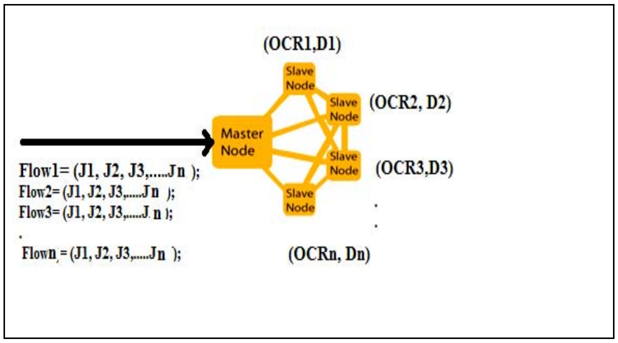

(OCR1,D1)
(OCR2, D2)
(OCR3,D3)
(OCRn, Dn)
Master Node
Slave Node
Slave Node
Slave Node
Slave Node
Flow1= (J1, J2, J3,.....Jn);
Flow2= (J1, J2, J3,.....Jn);
Flow3= (J1, J2, J3,.....J n);
Flown = (J1, J2, J3,.....Jn);

Figure 2. Example of job flow via Mapreduce

This figure is from the user guide book of Amazon MapReduce [14].

IV) The experimental study

A. *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 [20]. The Harvard's Open Collections Program (OCP) has produced online digital copies of over 280 manuscripts that date from the 10th to the 20th centuries CE. Fig.3 . This open dataset represents different regions, including Saudi Arabia, North Africa, Egypt, Syria, Lebanon, Palestine, Iran, Iraq, Turkey, and South, Southeast, and Central Asia.

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

For segmentation step, Horizontal projection is applied to detect lines then each line into words using drawing rectangle around each word and each word segmented into its primitives using vertical projection.

Wavelet transform [21] 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 characters shape variation according to their position within words and with different positions (rotation and translation)).

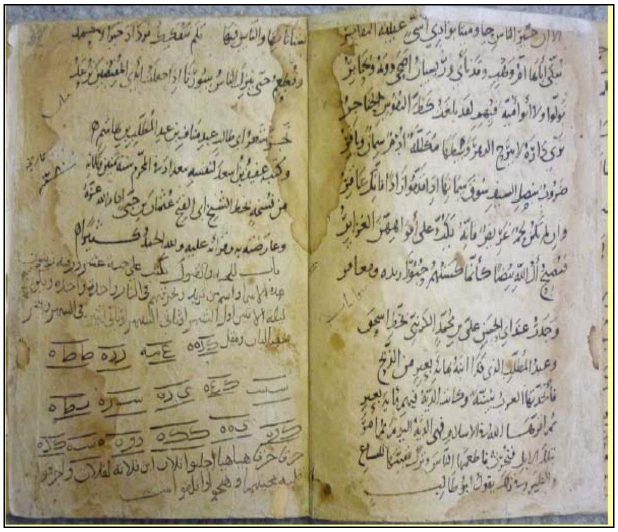


Figure 3. A sample of the studied corpus

B. 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 Linux command, the shell Cygwin [22] is used. Java, JDK 1.6 and Eclipse 3.4 were used program, implement and built our OCR application.

Cascading and Hadoop were used to manage our application in a distributed environment. 100 Mbits/s was the network capacity.

Based on a state of art of the cloud computing providers [23], Amazon Elastic Computing Cloud, is an IaaS cloud computing service that opens Amazon's large computing infrastructure to its users was selected for the implementation of our approach.

To verify that distributed hybrid approach K-NN/SVM functions correctly distributed platform such as cloud technologies, we created six running Jobs flow on the Amazon Elastic Computing Cloud service.

We have allocated 100 cores using the three Standard Amazon EC2 Instances. First the "small" instances each with 1.7 GB of memory, 160 GB of instance storage, and 32-bit platform. Second the Large Instance 7.5 GB of memory, 850 GB of instance storage, 64-bit platform and finally the Extra Large Instance 15 GB of memory, 1690 GB of instance storage and 64-bit platform. Amazon S3 is used to managed the input and output data.

The Experimental environment setup is based on two steps: The first step is the setup and configuration of the experimental environment while the second is to submit the application in a real cloud such as Amazon EC2.

In the first step, we setup the mentioned development environment to develop the application, and then we implement our OCR application and build the executable jar file with Eclipse. Finally we Run and debug our executable jar file in Cygwin with Hadoop.

In the second step, when our application was successfully running in Hadoop to submit it to Amazon cloud EC2. First we should Sign Up for Amazon S3 to create a bucket using the AWS Management Console were every object in Amazon S3 will be stored. Then we upload the data base and the OCR application in the created bucked using client tools of the CloudBerry Explorer for Amazon S3 installed on the local machine. The third sub step is to create a job flow using Cascading by specifying the input and output data, the processing application, the number of EC2 instances and finally we launch the job flow.

Finally, when the job flow finished processing the data, we pick up the results from our S3 bucket.

The picture below is from the user guide book Amazon cloud computing [22] describes the steps of using Amazon Elastic MapReduce to execute our approach.

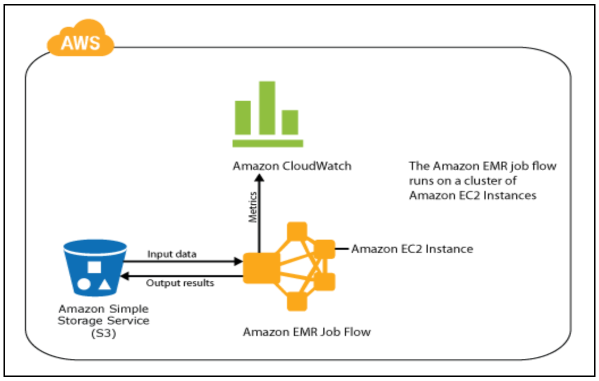

AWS
Amazon CloudWatch
The Amazon EMR job flow runs on a cluster of Amazon EC2 Instances
Metrics
Input data
Output results
Amazon EC2 Instance
Amazon Simple Storage Service (S3)
Amazon EMR Job Flow

Figure 4. Job flow execution in Amazon MapReduce Model

This figure is from the user guide book of Amazon MapReduce [15].

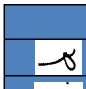
C. *Results and discussions*

In this section are published the results of several experiments performed in the Map reduce cloud technologies. All of them follow the same experimental methodology of previous proposed work whose results were reported over this same technology. Therefore, it is possible to compare the proposed method against other ones. Additionally, the preliminary and final experiments follow a cross validation methodology in order to grant a high confidence for the reported rates.

To evaluate, in the first hand the importance of the hybrid approach K-NN/SVM in the recognition rate and in the second hand the efficiency of the cloud computing technologies on the time 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.

The table below presents some confusion state with K-NN classifier.

TABLE I: SOME CONFUSION STATE

	×					
			×	×		
	×					
	×				×	×
					×	×
			×	×		
		×	×			

Hybrid approach K-NN/ SVM is used to eliminate this confusion.

The 28 Arabic characters written with different scripter in different positions in the word represent all the classes used in our experiments.

The recognition rate using K-NN and K-NN/SVM is presented respectively in the table below.

TABLE II. K-NN AND K-NN/SVM RECOGNITION RATE (%)

Classes	K-NN (%)	K-NN/SVM (%)
ا	96.00	97.30
ب	95.20	96.91
ت	95.10	96.70
ث	95.30	96.10
ج	96.20	97.40
ح	96.30	97.50
خ	95.50	96.60
د	95.50	96.10
ذ	96.10	96.90
ر	95,80	96,40
ز	95.00	96.37
س	96.16	96.79
ش	95.80	96.45
ص	96.00	96.70
ض	96.29	97.19
ط	96.00	96.87
ظ	96.10	96.80
ع	95.90	96.70
غ	96.00	97.10
ف	95.90	97.00
ق	96.00	97.10
ك	95.80	96.90
ل	96.00	97.20
م	95.89	97.00
ن	96.00	97.12
ه	96.20	97.30
و	96.00	97.13
ي	95.90	97.00
Average (%)	**95.94**	**96.92**

This table shows that the hybrid K-NN-SVM classifier improves the performance in terms of recognition rate compared with a single K-NN model for Arabic Islamic characters classification process. One can also observe that the results obtained using K-NN/SVM for Islamic manuscripts recognition are better than existing products [24][25] [26].

The table below presents the execution time using K-NN and K-NN/SVM with different instances of Amazon Elastic Computing.

TABLE III. 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 are approximately 7 hours and 8 hours and the average test time for 100 computers are 0.126 hours and 0.140 hours for K-NN and K-NN/SVM algorithm respectively. In conclusion, the sequential system is able to recognize only 21 and 18 characters per second for the two algorithms K-NN and K-NN/SVM.

Consequently, obtained results confirm that the execution time decrease when we ameliorate the instance of Amazon Elastic Computing: if we use 100 cores with large instance of Amazon Elastic Computing, then the execution time reaches the value 432 and 504 seconds. This result is very interesting, because in this case our proposed OCR system is able to recognize more than 1200 and 1041 characters per second for K-NN and K-NN/SVM.

Our approach presents many advantages: first , the automatic parallelization of our application, second obtained results confirm that distributed systems and more specifically cloud computing present a very interesting framework to speed up the Arabic handwriting recognition system based on the K-NN/SVM algorithm knowing this time include data partitioning, task scheduling, handling machine failures, managing inter-machine communication, and all is completely transparent to the programmer/analyst/user.

V) CONCLUSION AND PERSPECTIVE

This paper proposes an hybrid approach based on K-NN/SVM via cloud computing technologies which can increase the recognition rate and speed up the Arabic OCR system in order to digitalize large amounts of Arabic Islamic manuscripts.

Performance evaluation of the proposed approach confirms that Mapreduce, hadoop and cascading technologies provide an adequate platform to build much more powerful and scalable handwritten OCR systems. Indeed, this platform facilitates the data partitioning, task scheduling, handling machine failures and managing inter-machine.

As a future work, the proposed design model requires further investigations. In particular, we examine how to deploy our OCR application on a multi-cloud infrastructure and how to create a cloud computing specialized in the Arabic Islamic Manuscripts Digitization (AIMD- As A- Service).

ACKNOWLEDGMENT

The authors wish to thank the reviewers for their fruitful Comments. The authors also wish to acknowledge the members of the Miracl Laboratory, Sfax, Tunisia and the members of computer science department, College Of Science And arts at Al-Ola, Taibah University, KSA.

REFERENCES

[1] Friedman, J.: Flexible metric nearest neighbor classification, technical report 113, stanford university statistics department (1994)

[2] Li Baoli1, Yu Shiwen1, and Lu Qin2, "An Improved k-Nearest Neighbor Algorithm for Text Categorization", Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, 2003.

[3] Cleber Zanchettin, Byron Leite Dantas Bezerra and Washington W. Azevedo AKNN-SVM Hybrid Model for Cursive Handwriting Recognition, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[4] B. Boser, I. Guyon, and V. Vapnik, A training algorithm for optimal margin classifiers, In Proceedings of Fifth Annual Workshop on Computational Learning Theory, New York, (1992).

[5] C. Choisy and A. Belaid, Handwriting recognition using local methods for normalization and global methods for recognition, In Proceedings of Sixth Int. Conference On Document Analysis and Recognition, pp. 23-27, (2001).

[6] L.N. Teow and K.F. Loe, Robust vision-based features and classification schemes for off-line handwritten digit recognition, Pattern Recognition, January, (2002).

[7] B. Zhao, Y. Liu, and S.W. Xia, Support vector machines and its application inhandwritten numerical recognition, In Proceedings of 15th Int. Conference on Pattern Recognition, vol. 2, pp. 720-723, (2000).

[8] V. V. C. Cortes. Support vector network. Machine Learning, 20:273–297, 1995.

[9] Liang-Jie Zhang, Carl K Chang, Ephraim Feig, Robert Grossman, Keynote Panel, Business Cloud: Bringing The Power of SOA and Cloud Computing , pp.xix, 2008 IEEE InternationalConference on Services Computing (SCC 2008), July 2008

[10] Amazon Elastic Compute Cloud (Amazon EC2), http://aws.amazon.com/ec2/, 2009

[11] http://www.al-jazirah.com/1243255/ln37d.htm

[12] Available at: http://aws.amazon.com/s3/

[13] Available at: http://www.gigaspaces.com/

[14] Available at: http://home.elephantdrive.com/

[15] E. Pinheiro, W.-D. Weber, and L. A. Barroso. Failure trends in a large disk drive population. In Proc. of 5th USENIX Conference on File and Storage Technologies (FAST '07), San Jose, CA, USA, February 2007

[16] Available at: http://aws.amazon.com/elasticmapreduce/

[17] Available at: http://hadoop.apache.org/

[18] Available at: http://www.cascading.org

[19] Available at: http://aws.amazon.com/elasticmapreduce/

[20] Available at: http://ocp.hul.harvard.edu/

[21] Hassen Hamdi, Maher Khemakhem, A Comparative study of Arabic handwritten characters invariant feature. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011

[22] Available at: http://www.cygwin.com/

[23] R. Prodan and S. Ostermann, "A Survey and Taxonomy of Infrastructure as a Service and Web Hosting Cloud Providers," Proc. Int'l Conf. Grid Computing, pp. 1-10, 2009.

[24] Iman Zangeneh, Mostafa Moradi, Ali Mokhtarbaf, The Comparison of Data Replication in Distributed Systems, World Academy of Science, Engineering and Technology 59 2011.

[25] CiyaICR product, http://www.Ciyasoft.com,2004

[26] M.Khemakhem and A. Belghith. Towards A Distributed Arabic OCR Based on the DTW Algorithm: Performance Analysis The International Arab Journal of Information Technology, Vol. 6, No. 2, April 2009.