5

Grid'5000 Based Large Scale OCR Using the DTW Algorithm: Case of the Arabic Cursive Writing

Mohamed Labidi[1], Maher Khemakhem[2] and Mohamed Jemni[3]
[1]*Research Unit UTIC, ESSTT/ University of Tunis, Tunis*
[2]*MIRACL Lab, FSEGS/ University of Sfax, Sfax*
[3]*Research Unit UTIC, ESSTT/ University of Tunis, Tunis*
Tunisia

1. Introduction

Large scale optical character recognition (OCR) refers to or means the computerization of large amounts of documents such as news papers. Despite the diversity of commercial OCR products, this task still remains too far from the mature especially if the input documents are insufficient quality or cursive writing such as the Arabic documents (Vinciarelli, 2002). Indeed, in their project (Holley, 2009), the national library of Australia reports that existing OCR systems are, commonly, weak. Moreover, their conducted experiments on historical newspapers show that the corresponding accuracy raw varied from 71% to 98.02%. This is surely due to the weakness of the approaches and techniques used in these systems.

Printed cursive written documents such as the Arabic one presents, in addition, other difficulties which are behind the weaknesses of the existing commercialized systems especially when the quality of the input binary image of the document is not good enough. The first difficulty encountered for such writing is the segmentation of any given input word or sub-word into isolated characters given that the size of each of which is variable. In practice, if the segmentation process is conducted successfully, then it eases the recognition step to a large extent. That is why Latin printed OCR systems are, commonly, more powerful compared to those devoted to the cursive writing documents.

Dynamic Time Warp (DTW) algorithm is a well known procedure especially in pattern recognition (Alves et al., 2002; Khemakhem et al., 2005; Philip, 1992; Vuori et al., 2001), (Khemakhem et al., 2009; Kumar et al., 2006; Tapia et al., 2007). The DTW algorithm is the result of the adaptation of dynamic programming to the field of pattern recognition. Printed cursive writing OCR by the DTW algorithm provides very interesting recognition rates without prior character segmentation (such as: the Arabic, Persian, Urdu, latin connected characters,...), (Khemakhem et al., 2005). The purpose of the DTW algorithm is to perform optimal time alignment between a reference pattern and an unknown pattern and evaluate their difference. Intensive experiments show that the recognition rate of the DTW algorithm remains acceptable compared to the existing commercialized systems even when the quality of the input documents is not good enough. Intensive tests on more than 100.000 connected characters (most of them are Arabic cursive and including some important noise) show that the segmentation average rate is greater than 98% and the recognition average rate is greater than 97% (Abedi et al., 2004). Consequently, we think that it is possible to build a powerful OCR system based on the DTW algorithm. Unfortunately, the enormous amount of computing to be achieved constitutes, however, the main drawback and hence restricts the use of the DTW algorithm.

Several works and approaches have been proposed to solve this problem (Philip, 1992) (Alves et al., 2002; Khemakhem et al., 2007; 2005; 1993). We are rather interested in this paper in the distributed systems and their possibilities to provide, costless and on demand, enough computing power which can ensure the substantial reduction of the DTW response time. Indeed, we found in our previous work that the response time of this algorithm is proportional to the provided computing power (Khemakhem et al., 2009). This quite means that if we own a grid computing then we can reach very interesting and increasing speedup factors.

Grid computing is an attractive infrastructure that provides a huge computing power, (Buyya et al., 2005; Foster et al., 2002; IBM, 2003; Khemakhem et al., 2009) without any prior investment. This is due to its ability to interconnect many computer networks of several organizations at the same time. Consequently, users can share many heterogeneous computer resources such us computing power. In our previous work, we have shown through an analytical and experimental studies (Khemakhem et al., 2009), that grid infrastructures can provide an adequate solution to very large quantities of Arabic documents OCR. Unfortunately, in our previous work we did not find the opportunity to make experiments on a scalable grid where we can achieve more significant experimental performance evaluation. Recently and fortunately the occasion has been provided to us to make intensive experiments on the French Grid’5000. Consequently, we report in this chapter these results which confirm, indeed, the results of our previous works.

This chapter starts with a brief presentation of the state of the art in terms of large scale OCR systems. Then, it presents and formulates the printed cursive Writing OCR by the DTW algorithm. In the forth part, a brief overview on grid computing and Grid’5000 platform are stated. The fifth part details the experimental conditions and the performance evaluation. Finally, we conclude and present some perspectives of this work.

2. Related work

A few solutions for large scale OCR are provided by computer scientists. Representative examples including OCRGrid (Goto, 2006), OCRopus (http://www.ocropus.org), Kirtas(http://www.kirtas.com) and The Australian Newspaper Digitization Project (Holley, 2009).

OCRGrid is a platform for distributed and cooperative OCR systems. The main idea of OCRGrid consists of deploying a lot of OCR servers on a network to allow end users to search for and use the adequate server. As servers can cooperate with each other, clients can benefit from a distributed parallel environment and consequently accelerate OCR tasks. On the other hand, applications searching for improving accuracy can benefit from OCRGrid due to the use of Majority logic technique which requires the running of many OCR engines. A multilingual processing environment can be also realized by combining a lot of community-supported OCR servers for various languages with localized dictionaries.

OCRopus is an open source OCR system sponsored by Google. It targets with its service the research community by improving the state of the art in optical character recognition and sought to serve also the large scale commercial document conversions. The system applications include modern digital library and the recognition of classical literature. Its main perspective consists of familiarizing the system with more languages so as to become omni-lingual and omni-script through contributing in open source community. Although the system has been evolving, it has not yet incorporated the Arabic language into its framework. Kirtas technology is an automatic book scanner which can do batch OCR for large volumes of books and other documents. By using innovative "automatic page-turning scanner" technology and a high-resolution Canon digital camera, Kirtas ensures image processing, quality control, OCR and Metadata. It can handle 15 left-to-right languages including English and French, 5 right-to-left languages including Arabic and 3 bilingual languages including Arabic/English. Kirtas OCR processing rates are too fast (about 1 page per second). Many public and university libraries decided to exploit Kirtas technology to digitize their old books. For all this, such a technology is too expensive and only rich institutions can afford to implement it.

National library of Australia has used OCR software to establish a large scale Historic Digitization Project. The Australian Newspaper Digitization Program (ANDP) claimed that 'acceptable' OCR was still to be improved. Besides, the poor quality of the original source documents urged the National Library of Australia to work with what was at hand but to anticipate the lack of OCR accuracy. In order to improve the quality of OCR accuracy, the committee of the Library adopted from the thirteen methods they came out with only five new ones, which are going to be tested and investigated. These methods are mostly to compare image optimization software, to experiment using greyscale files, to use Australian dictionaries, to clean/correct OCR text manually, and to use confusion matrix and language modeling post/during OCR processing. They searched for improving OCR accuracy by using both a combination of methods and manual methods of humans correcting the mistakes of machines. The Australian library was considered leading in that it was the first worldwide to involve public users in the correction of texts instead of the contractor. Such a solution was considered labor intensive for the Library before the emergence of web 2.0 technologies. Although the public was not informed that they could introduce any corrections to the texts, they embarked on this correction immediately. There were measures decided to check the accuracy of the OCR-corrected text via counting the number of lines corrected and number of different articles corrected. However, the interventions of public users may badly affect the content of the articles so they thought of monitoring and moderation to make sure that no data has been added to the original text.

In the next section, we will present the basics of the DTW algorithm which is considered the corner stone of our proposal to solve the problem of large scale OCR.

3. The DTW algorithm

This algorithm is a well known procedure especially in pattern recognition (Philip, 1992) (Alves et al., 2002; Khemakhem et al., 2007; Tapia et al., 2007; Vinciarelli, 2002; Vuori et al., 2001). The purpose of this procedure is to perform an optimal time alignment between a reference pattern and an unknown pattern and evaluate their difference. What makes the DTW procedure very attractive is its ability to recognize properly cursive characters (connected blocks of characters such as words or parts of words in Arabic) without need of a prior segmentation into characters according to a given reference library of isolated characters. The adaptation of this procedure to the Arabic cursive OCR has shown to provide very interesting results (Khemakhem et al., 2007; 2005).

3.1 Cursive writing OCR by the DTW algorithm

Words in any cursive writing, such as the Arabic language, are inherently written in blocks of connected characters. While the segmentation of the text into blocks of connected characters is a preliminary phase to the recognition process, a further segmentation of these blocks into separate characters is usually adopted. Indeed, many researchers have considered the segmentation of Arabic words into isolated characters before performing the recognition phase (AlBadr et al., 1998; Cheung et al., 2001; Vuori et al., 2001). The crux of the viability of the use of DTW technique is then its ability and efficiency to perform the recognition without the prior segmentation of blocks into separate characters.

Let V represents a reference library of$R$ trained characters $Cr, r = 1, 2, , R$. defining the Arabic alphabet in some given fonts. We here stress the fact that several fonts could be considered even simultaneously It suffices to get them trained which is easily done at the learning phase while constructing the reference library V. Let T represents a block of connected Arabic characters to be recognized. T is then composed of a sequence of N feature vectors Ti that are actually representing the concatenation of some subsequences of feature vectors representing each an unknown character to be recognized. The text T is seen as lying on the time axis (the X-axis) in such a manner that feature vector Ti stands at time i on this axis. The reference library V is portrayed on the Y-axis, where the reference character Cr is of length $lr, 1 \leq r \leq R$. According to (Khemakhem et al., 2007). Let $S(i, j, r)$ represents the cumulative distance at point (i, j) relative to the reference character Cr. The objective is then to detect simultaneously and dynamically the number of characters making T and recognizing these characters. There exists surely a number k and indices $(m1, m2, ..., mk)$ such that $Cm1 + Cm2 + ... + Cmk$ represents the optimal alignment to text T where denotes the concatenation operation. The path warping from point $(1, 1, m1)$ to point (N, lmk, k) and representing the optimal alignment is therefore of minimum cumulative distance that is:

$$S(N, l_{m_k}, k) = \min_{1 \leq r \leq R} \{S(N, l_r, r)\} \tag{1}$$

This path, however, is not continuous since it spans many different characters. We therefore must allow at any time the transition from the end of one reference character to the beginning of a new character. The end of reference character C_r is first reached whenever the warping function reaches the point (i, l_r, r) where $i = \lceil \frac{l_r+1}{2} \rceil, ..., N$. The warping function always reaches the ends of the reference characters. At each time i, we allow the start of the warping function at the beginning of each reference character along with the addition of the smallest cumulative distance among the end points found at time $(i - 1)$. The resulting functional equations are:

$$S(i, j, r) = D(i, j, r) + \min_{\substack{1 \leq i \leq N \\ 1 \leq j \leq l_r \\ 1 \leq r \leq R}} \left\{ \begin{array}{c} S(i-1, j, r), \\ S(i-1, j-1, r), \\ S(i-1, j-2, r) \end{array} \right\} \tag{2}$$

with the boundary conditions :

$$S(i,1,r) = D(i,1,r) + \min_{\substack{1+\lceil \frac{1+\min\limits_{1\leq r\leq R}\{l_r\}}{2} \rceil \leq i \leq N \\ 1 \leq k \leq R \\ 1 \leq r \leq R}} S\left(i-1, l_k, k\right) \tag{3}$$

To trace back the warping function and the optimal alignment path, we have to memorize the transition time from one reference character to the others (Khemakhem et al., 2007). This can easily be accomplished by the following procedure:

$$b(i,j,r) = trace \min_{\substack{1 \leq i \leq N \\ 1 \leq j \leq l_r \\ 1 \leq r \leq R}} \left\{ \begin{array}{c} b(i-1,j,r), \\ b(i-1,j-1,r), \\ b(i-1,j-2,r) \end{array} \right\} \tag{4}$$

Where *trace* min is a function that returns the element corresponding to the term that minimizes the functional equations.

4. Grid computing

A grid is a collection of machines, sometimes referred to as nodes, resources, members, donors, clients, hosts, engines, and many other such terms. They all contribute any combination of resources to the grid as a whole. Some resources may be used by all users of the grid while others may have specific restrictions, (Bahi et al., 2006; Buyya et al., 2005; Foster et al., 2002; IBM, 2003; Shi et al., 2006).

In most organizations, there are large amounts of under utilized computing resources. Most desktop machines are busy less than 5 percent of the time. In some organizations, even the server machines can often be relatively idle. Grid computing provides a framework for exploiting these under utilized resources and thus has the possibility of substantially increasing the efficiency of resource usage (IBM, 2003).

Often, machines may have enormous unused disk drive capacity. Grid computing, more specifically, a data grid, can be used to aggregate this unused storage into a much larger virtual data store, possibly configured to achieve improved performance and reliability over that of any single machine (IBM, 2003).

Consequently, a grid computing is an infrastructure that allows to many institutions (regardless their geographical locations) to interconnect a large collection of their heterogeneous computer networks and systems to share together a set of software and/or hardware resources, services, licences ... (Buyya et al., 2005; Foster et al., 2002; IBM, 2003). This huge ability of sharing resources in various combinations will lead to many advantages such as: $*$Increase the efficiency of resource usage;

$*$Facilitate the remote collaboration between: institutions, researchers, ...

$*$Give to users a huge computing power;

$*$Give to users a huge storage capacity, etc.

Some researchers attempt now to model and realize this infrastructure in the right manner and some others attempt to anticipate and to take the expected advantages of such infrastructure to solve several problems, (Bahi et al., 2006; Buyya et al., 2005; Foster et al., 2002; IBM, 2003; Shi et al., 2006).

4.1 Grid’5000

Grid’5000 is a French research effort developing a large scale nation wide infrastructure for Grid research. 17 French laboratories are involved, nation wide, in the objective of providing the community of Grid researchers a test bed allowing experiments in all the software layers between the network protocols up to the applications (Bolze et al., 2006). The Grid’5000 platform is intended to support research in all areas of computer science related to large scale distributed processing and networking. Researchers should use Grid’5000 in the perspective of large scale experiments (at least 3 sites and 1000 CPUs). They may generate useful results for other communities, as long as the community of computer science researchers learns something from Grid’5000 experiments. It is a shared tool, used by many people with different and varying needs. The administrators pursue the following objectives, the main one being the first of this list:

1.Make the tool available to experiments involving a significant number of nodes (in the 1000’s). To make this possible, reservation fragmentation must be avoided as much as possible.

2.Keep the platform available for the development of experiments during the day. Therefore, reservations using all the nodes available on one site during work hours (in France) should be avoided in general.

Amongst the advantages of this grid is the ease of its reconfiguration. Indeed, whenever a user wants to perform experiments on Grid’5000, he has to make a prior reservation of a fixed number of nodes for a precise duration of time. Once the request of this user is accepted, he will be the lone authorized to exploit reserved nodes during the required period. A minimal set of software is installed on the nodes of each site. Sometimes a user should create a “Kadeploy” [1] image in which other software package that he needs can be included. He should also deploy it on the authorized nodes before its utilization.

The aim of this chapter is to prove trough intensive experiments over Grid’5000 that such infrastructure constitutes maybe the only way to solve the problem of costless computerization of big quantities of cursive writing documents using the DTW algorithm.

4.2 Deployment of the Arabic OCR over the Grid’5000

Recall that the objective of the present chapter is to validate the analytical results found in our previous work which reveal that grid computing is an adequate infrastructure to build a powerful distributed system which is able to perform large scale OCR, (Khemakhem et al., 2009).

It is commonly known that if we would like to distribute any given application over any given distributed infrastructure, one could make a decision about the adequate way to achieve it. Indeed, we can exploit either the inherent parallelism of the application algorithms or the corresponding data distribution or maybe both. For our application, we found in our previous work (Khemakhem et al., 1993) that the only way to reach interesting and customized speedup for the recognition process of the Arabic OCR based on the DTW algorithm is to proceed to

[1] Kadeploy is a fast and scalable deployment system towards cluster and grid computing. It provides a set of tools, for cloning, configuring (post installation) and managing a set of nodes. Currently it deploys successfully linux, *BSD, Windows, Solaris on x86 and 64 bits computers.

the data distribution. It means that we have to assign to each processor participating in the work, both, the application executable and a part of the data to be processed. Such solution is simpler and induces a very large degree of distribution. Our strategy consists of using one node of the grid as master and the remaining nodes as workers. During our experiments, each worker has been assigned by the same number of binary images which represents a part of the overall documents to be processed because all of the used grid workers are homogeneous (see table 1). For each experiment, the number of binary images to be assigned to each worker is calculated according to the number of workers participating in the work. The master starts by assigning to each worker the Arabic OCR Application and a part of the input documents to be computerized (in the form of binary image files). Next, it launches remotely the recognition process on each worker. For each input binary image file the target worker generates an output ascii file giving the recognition result. these files are turned back to the master.

5. Performance evaluation of massive cursive writing OCR over the GRID'5000

To ascertain our thesis which considers that grid infrastructures constitute an adequate solution (and maybe the only way) to solve the problem of large scale Arabic OCR without any prior investment, we have conducted intensive experiments over the grid'5000. These experiments have concerned the distribution of a corpus test composed of 10000 binary image files. Each image contains around 1200 characters.
Recall that the principle of our deployment consists to take the input binary images of the arabic documents to be processed (each image corresponds to a single document page) and then assign them optimally or pseudo optimally among the targeted computers (denoted by workers) of the grid. This means that every computer participating in the work will be assigned, naturally, according to its own computing power.

5.1 The experimental conditions

Cluster	Location	Cluster type	Nodes	CPU Type	Frequency	Memory
Capricorne	Lyon	IBM eServer 326	56	AMD Opteron 246	2.0GHz	2 GB
Sagittaire	Lyon	Sun Fire V20z	79	AMD Opteron 250	2.4GHz	2GB
Gdx	Orsay	IBM eServer 326m	312	AMD Opteron 246	2.0 GHz	2 GB
Netgdx	Orsay	IBM eServer 326m	30	AMD Opteron 246	2.0 GHz	2 GB

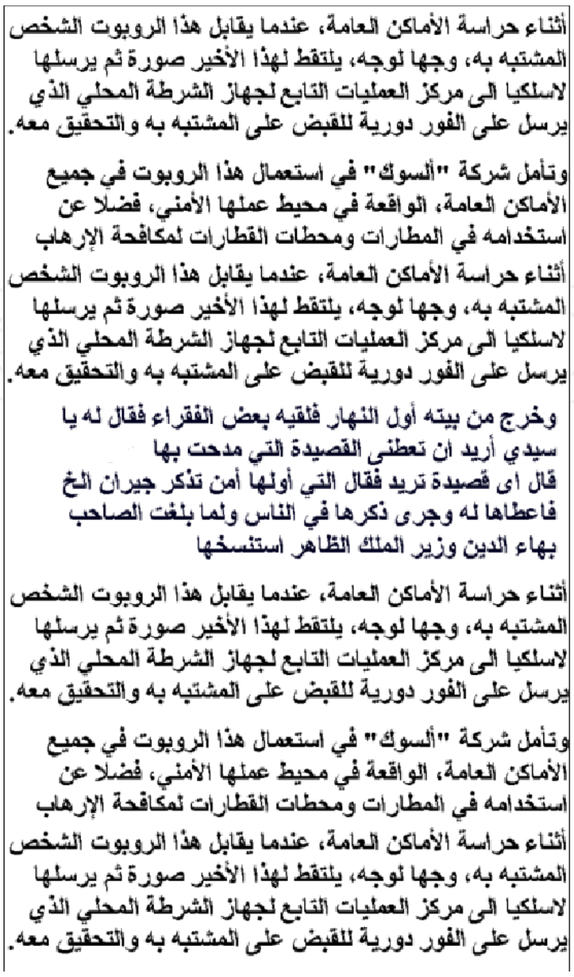
Table 1. Hardware configuration of the Grid testbed

Experiments were conducted using 4 clusters of 2 different sites in France. The hardware configuration is shown in table 1. Debian GNU/Linux 4.0 is the operating system of all used workers. For the software, we used specialized scripts shell to assign tasks and keep the recognition results. As described at the Section 4.1, the nodes must be reconfigured at each utilization.

5.2 Results and prformance evalaution

5.2.1 The studied corpus

We have used a corpus of 10000 images which have been assigned optimally and pseudo optimally among the authorized workers. Figure 1 illustrates a sample of the studied corpus.



أثناء حراسة الأماكن العامة، عندما يقابل هذا الروبوت الشخص
المشتبه به، وجها لوجه، يلتقط لهذا الأخير صورة ثم يرسلها
لاسلكيا الى مركز العمليات التابع لجهاز الشرطة المحلي الذي
يرسل على الفور دورية للقبض على المشتبه به والتحقيق معه.

وتأمل شركة "ألسوك" في استعمال هذا الروبوت في جميع
الأماكن العامة، الواقعة في محيط عملها الأمني، فضلا عن
استخدامه في المطارات ومحطات القطارات لمكافحة الإرهاب

أثناء حراسة الأماكن العامة، عندما يقابل هذا الروبوت الشخص
المشتبه به، وجها لوجه، يلتقط لهذا الأخير صورة ثم يرسلها
لاسلكيا الى مركز العمليات التابع لجهاز الشرطة المحلي الذي
يرسل على الفور دورية للقبض على المشتبه به والتحقيق معه.

وخرج من بيته أول النهار فلقيه بعض الفقراء فقال له يا
سيدي أريد ان تعطنى القصيدة التي مدحت بها
قال اى قصيدة تريد فقال التي أولها أمن تذكر جيران الخ
فاعطاها له وجرى ذكرها في الناس ولما بلغت الصاحب
بهاء الدين وزير الملك الظاهر استنسخها

أثناء حراسة الأماكن العامة، عندما يقابل هذا الروبوت الشخص
المشتبه به، وجها لوجه، يلتقط لهذا الأخير صورة ثم يرسلها
لاسلكيا الى مركز العمليات التابع لجهاز الشرطة المحلي الذي
يرسل على الفور دورية للقبض على المشتبه به والتحقيق معه.

وتأمل شركة "ألسوك" في استعمال هذا الروبوت في جميع
الأماكن العامة، الواقعة في محيط عملها الأمني، فضلا عن
استخدامه في المطارات ومحطات القطارات لمكافحة الإرهاب

أثناء حراسة الأماكن العامة، عندما يقابل هذا الروبوت الشخص
المشتبه به، وجها لوجه، يلتقط لهذا الأخير صورة ثم يرسلها
لاسلكيا الى مركز العمليات التابع لجهاز الشرطة المحلي الذي
يرسل على الفور دورية للقبض على المشتبه به والتحقيق معه.

Fig. 1. A sample of the studied corpus

5.2.2 Results found

In this section we present the obtained results of the studied application over the Grid'5000. Indeed, the next results cover the distribution of the already presented corpus over a variable number of homogeneous nodes of Grid'5000 varying from 39 to 426.

Figure 2 illustrates the recognition time against the number of workers used. The horizontal axis represents the number of used workers and the vertical axis shows the corresponding recognition time. We observe that the recognition time decreases significantly with the increase of the number of workers. Indeed, we show that the time required to recognize the totality of the studied corpus decreases from 334 minutes by using 39 workers to 35 minutes by using 426 workers. Recall that the recognition of the corpus in the sequential mode requires approximately 10797 minutes which is equivalent to more than one week! such results confirm that the response time of the recognition process is proportional to the provided computing power, (Khemakhem et al., 2009). Consequently, we can confirm that grid infrastructure is adequate to solve the large scale OCR problem such as the computerization of library of books by using the DTW algorithm which is almost impossible to perform sequentially.
Figure 3 illustrates the speedup of the described experiment against the number of workers used. The horizontal axis represents the number of workers used and the vertical axis gives the speedup factor. We observe that the speedup is an increasing function of the

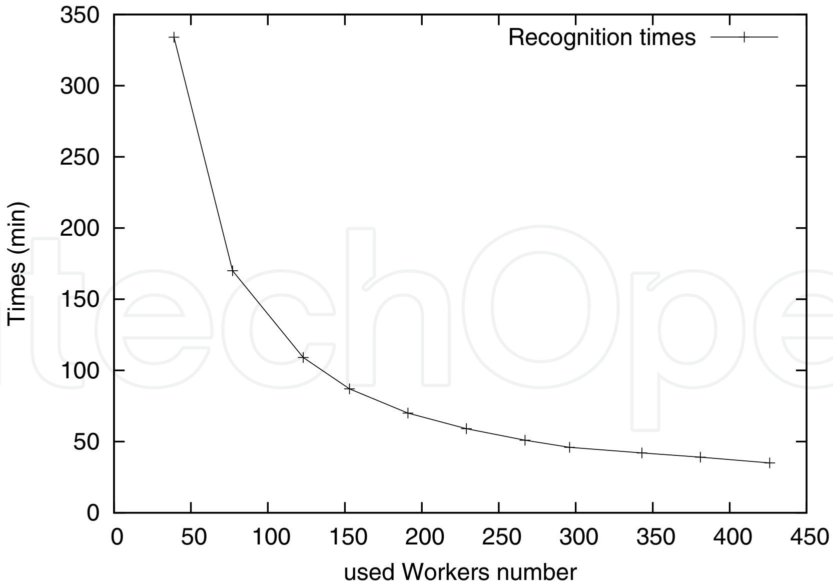


Fig. 2. The total recognition time of the studied corpus

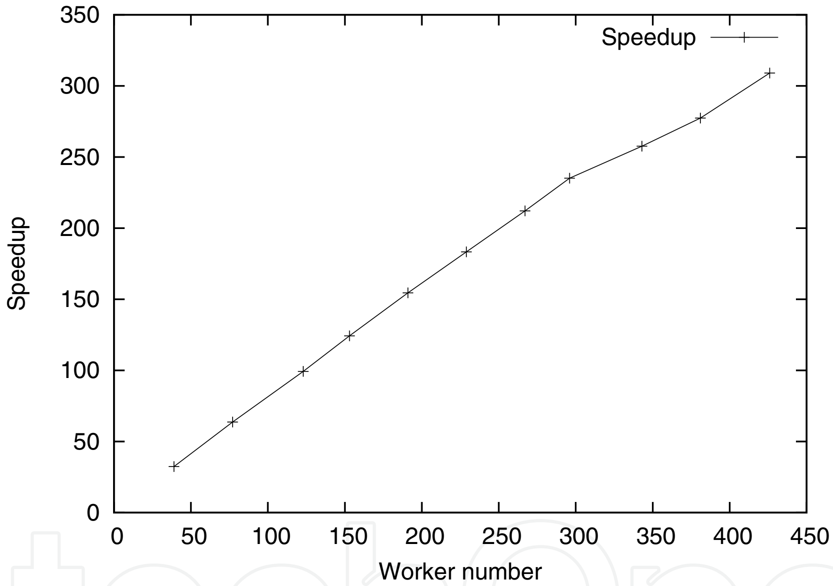


Fig. 3. Speedup factor

used workers. This figure shows, indeed, that we can reach interesting speedup factors; for example by using 426 workers, we reached 309 as speedup factor. This means that our proposed distributed system is able to recognize more than 5500 characters a second. We note that our sequential system is able to recognize only 18 characters a second. Such results confirm that grid'5000 is an adequate infrastructure to speedup drastically the response time of the DTW algorithm. Furthermore, with such enough computing power, we can improve the recognition rate by adding some complementary approaches and techniques. As a way to reach this expectation, we can use, for example, rich lexicons where we can check all recognized words and sub words and try to overcome those which don't have any linguistic meaning. Besides, we are convinced that volunteer grids (Khemakhem et al., 2009) can be considered also as amongst the adequate infrastructures which can solve large scale OCR given that the corresponding resources (workers) aren't condemned like those of dedicated grids. That is why the experimental study of these infrastructures presents for us also another perspective despite the corresponding challenges especially the volatility of nodes participating in the work and the security of data to be processed. Consequently we are faced to two issues: the fault tolerance and the data integrity (such as data alteration) problems. To solve these problems, the use of middleware seems to be mandatory to deploy efficiently large scale cursive writing OCR over volunteer grids.

6. Conclusion and perspectives

In this chapter we have shown how grid environments are able to solve both problems; the large scale OCR and the complex computing power of the DTW algorithm. The enough computing and storage power provided by any grid environment make very attractive our proposal. Indeed, these can help a lot to consider other types of complementary post treatment approaches and techniques to the DTW algorithm which will lead, surely, to the improvement of its recognition rate. Several investigations are under study especially the selection and then the integration of some other complementary approaches and/or techniques which can enhance the recognition rate of our proposal. In addition, the use of an appropriate middleware to manage data scheduling and ensure the fault tolerance seems to be mandatory if we wish to migrate to volunteer grid.

7. Acknowledgment

Experiments presented in this chapter were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

8. References

Abedi, N. et Khemakhem, M.(2004) Reconnaissance de Caractères Imprimés Cursifs Arabes Par Comparaison Dynamique et Modèles Cachés de Markov, *in Proceedings of GEI Tunisia*, pp. 56-63, Monastir, Tunisia, March 2004.

AlBadr, A.& Haralick, R.(1998). A segmentation free approach to text recognition with application to Arabic text, *International Journal On Document Analysis And Recognition (IJDAR),*, Vol. 1, No. 3, pp.147-166, 1998.

Alves, C, E, R. Caceres, E, N.& Dehne,F.(2002). Parallel Dynamic Programming for solving the String Editing Problem *Proceedings of the fourteenth annual ACM symposium on Parallel algorithms and architectures*, ISBN:1-58113-529-7, Winnipeg, Manitoba, Canada, August 10-13, 2002.

Bahi, J, M. Couturier, R. Mazouzi, K.& Salomo, M. (2006) Synchronous and asynchronous solution of a 3D transport model in a grid computing environment, *Applied Mathematical Modelling,*, Vol. 30, pp. 616-628 2006

Bolze, R.& all. (2006). Grid'5000: a large scale and highly reconfigurable grid experimental testbed, *International Journal of High Performance Computing Applications,*, Vol. 20, No. 4, pp. 481-494 Winter 2006.

Buyya, R.& Venugopal, S. (2005). A Gentle introduction to Grid Computing and Technologies, *CSI Communications*, Vol. 29, No. 1, pp.9-19, India, May 7-19, 2005.

Cheung, A. Bennamoun, M.& Bergman, N. (2001). An Arabic optical character recognition system using recognition based segmentation, *Pattern Recongnition,*, pp. 215-233, 2001.

Foster, I. Kesselman, C. & Tuecke, S.(2002). The Anatomy of the Grid, *Intl J. Supercomputer Applications*, Vol. 15, (2001) pp.200-222.

Goto, H. (2006). OCRGrid: A Platform for Distributed and Cooperative OCR Systems, *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 2, pp.982-985, 2006

Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, *D-Lib Magazine* March/April 2009 Vol. 15, No. 3/4, ISSN 1082-9873.

IBM. (2003). Introduction to Grid Computing with Globus. *IBM RedBook*, URL: http://www.redbooks.ibm.com/redbooks/pdfs/sg246895.pdf, September 2003.

Khemakhem, M. et Belghith, A. (2009). Towards A Distributed Arabic OCR Based on the DTW Algoriyhm: Performance Analysis, *The International Arab Journal of Information Technology*, Vol. 6, No. 2, pp. 153-161, April 2009.

Khemakhem, M. et Belghith, A. (2009). A P2p Grid Architecture For Distributed Arabic OCR Based On The DTW Algorithm, *The International Journal of Computers and Applications (IJCA)*, Vol. 31.,No. 1, ACTA PRESS, 2009.

Khemakhem, M. & Belghith, A. (2007). Agent based architecture for Parallel and Distributed Complex Information Processing, *the International Revue on Computers and Softwares (IRECOS)*, Vol. 2, No. 1, January, 2007, 38–44.

Khemakhem, M.& Belghith, A. (2005). A Multipurpose Multi-Agent System based on a loosely coupled Architecture to speedup the DTW algorithm for Arabic printed cursive OCR, *Proceedings of the ACS/IEEE 2005 International Conference on Computer Systems and Applications*, pp. 121-vii, ISBN, Egypt, January 2005, aiccsa, Cairo

Khemakhem, M. Belghith, A.& Ben Ahmed,M. (1993). Modélisation architecturale de la Comparaison Dynamique distribuée, *Proceedings of the Second International Congress On Arabic and Advanced Computer Technology*, Casablanca, Morocco, December 1993.

Kumar, A. Balasubramanian, A. Namboodiri, AM.& Jawahar, C. (2006). Model-Based Annotation of Online Handwritten Datasets, *Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition*, http://hal.inria.fr/inria-00105158/en/, 2006.

Philip, G. Bradford, (1992). Efficient Parallel Dynamic Programming, *Proceedings of the 30th Annual Allerton Conference on Communication, Control and Computing*, 185-194, University of Illinois, 1992.

Shi, Z. Huang, H. Luo, J. Lin, F.& Zhang, H.(2006). Agent Based Grid Computing, *Journal of Applied Mathematical Modelling*, Vol. 30, pp. 629-640 2006

Tapia,E.& Rojas, R. (2007). A Survey on Recognition of On-Line Handwritten Mathematical Notation *Technical Report B-07-01 Freie University at Berlin, Institut fur Informatik Takustr.*, 9, 14195 Berlin, Germany, January 26, 2007.

Vinciarelli, A.(2002). A survey on offline cursive word recognition, *Pattern Recognition*, Vol. 35, pp: 1433-1446, 2002.

Vuori, V. Laaksonen, J. Oja, E. et Kangas, J.(2001). Experiments with adaptation strategies for a prototype-based recognition system for isolated handwritten characters, *IJDAR*, Vol. 3, pp: 150-159, 2001.

Recent Advances in Document Recognition and Understanding
Edited by Dr. Minoru Mori

ISBN 978-953-307-320-0
Hard cover, 94 pages
Publisher InTech
Published online 17, October, 2011
Published in print edition October, 2011

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821