

Annual Report for Period:08/2002 - 08/2003**Submitted on:** 05/15/2003**Principal Investigator:** Mukherjee, Amar .**Award ID:** 0207819**Organization:** U of Central Florida**Title:**

Collaborative: Compressed Domain Search for Text and Images by Sorted Contexts

Project Participants

Senior Personnel

Name: Mukherjee, Amar**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Post-doc

Graduate Student

Name: Zhang, Nan**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Mr. Zhang is a Ph. D. student working on this project. His primary reserach is concerned with text compression and compressed domain pattern matching for text. Mr. Zhang is working under Prof. Mukherjee's supervision.

Name: Satya, Ravi**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Ravi Vijaya Satya is working on related area of DNA compression and bioinformatics using some of the BWT-based techniques being developed in this grant. Mr. Satya is working under Prof. Mukherjee's supervision.

Name: Tao, Tao**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Tao Tao is working on compressed domain pattern matching for image problems. He is a Ph. D. student. . Mr. Tao is working under Prof. Mukherjee's supervision.

Name: Sun, Weifeng**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Weifeng Sun is a prospective Ph. D. student and is working on lossles text compression problems. Mr. Sun is working under Prof. Mukherjee's supervision.

Name: Powell, Matt**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Matt Powell is a student of Prof. Tim Bell, a co-PI located at Canterbury University, New Zealand. He worked on compressed domain pattern matching problem using binary search t4echniques.

Name: Firth, Andrew**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Firth worked on compressed domain pattern serach using BWT transform. Mr. Firth is worked under Prof. Bell's supervision.

Undergraduate Student

Technician, Programmer**Other Participant**

Name: Bell, Tim

Worked for more than 160 Hours: Yes

Contribution to Project:

Prof. Bell is a co-PI for this project and is playing a critical role in developing ideas, writing papers and supervising students contributing to the project objectives.

Research Experience for Undergraduates**Organizational Partners****Other Collaborators or Contacts**

We have been collaborating with a well-known researcher in the data compression field: Professor Tim Bell of Computer Science Department, University of Canterbury, New Zealand. Tim Bell is one of the co-Principal Investigators of the project although he has been listed as a Senior Personnel on the budget page of the proposal for technical reasons. His two students Matt Powell and Andrew Firth have contributed to the project. NSF does not fund their activities; they are supported by the resources available to them from the University of Canterbury, New Zealand. We have been working on several joint papers on compressed domain pattern matching. Also, we are discussing the possibility of linking up our online compression utility website vlsi.cs.ucf.edu with the Canterbury website.

Activities and Findings**Research and Education Activities:**

We developed compressed pattern matching algorithms that achieve the complexity bounds of linear exact pattern matching algorithms for text compressed with the sorted context approach and also algorithms based on binary and q-gram search of sorted context generated by the Burrows-Wheeler transform.

We studied the feasibility and utility of introducing error prediction (as used in context-based lossless image compression) at one or more stages of the compression process. We also started preliminary work on compressed domain approximate pattern matching.

Findings:

We develop two techniques for searching BWT transformed text using Boyer-Moore algorithm and binary search. The first technique applies the Boyer-Moore algorithm on characters that match while they are decompressed and skips the part of the compressed file that cannot possibly lead to any match. The second technique is based on the observation that the BWT transform contains a sorted list of all substrings of the original string, which can be exploited for rapid searching using a variant of binary search. Both techniques are faster than decompress-then-search approach for small number of queries, and binary search is even faster for large number of queries. The attached file gives further details of our approach.

The sorted context of the BWT transformed text also forms the basis of a pattern search algorithm which uses the q-grams of the pattern against the sorted q-grams of the text. The decoder only has limited information about the sorted context, but fast q-gram (context) generation and matching algorithms have been developed to exploit this with the help of auxiliary index arrays built in linear time. The algorithm (we call it the QGRAM algorithm) first computes the index arrays for the correspondence between F, the first column of the sorted

cyclic matrix in BWT and T. All exact matches are grouped in consecutive rows of the sorted matrix, which makes the binary searches on F and/or q-grams of the matrix very. We also developed a new algorithm, called QGREP, which improved on the sublinear search time of the binary search and QGRAM algorithms.

We have started working on ways to extend the algorithms for approximate pattern matching, especially the k-mismatch problem, and the k-approximate matching problem.

For image compression, we have been studying algorithms for adaptive scanning-path for BWT-based lossless image compression. The methods use image statistics to predict the activity in the image. Based on this, and the nature of transformed output from the BWT, the algorithms determine the scanning path to use for the given part of the image. This provides adaptability in the scanning path without the time consuming problem of explicit edge detection or image segmentation.

Training and Development:

Several Ph.D. and Masters students have participated and contributed in this research project, but not all of them received direct support from the grant. Individual Co-PIs meet with graduate students at their respective universities on a regular basis to discuss research problems. The students acquire the necessary skills to search literature and carry on an in-depth study and research in a field. The students are also asked to make presentations on their work. This gives the students experience of teaching graduate level courses and seminars. The overall effect of these activities is to train graduate students with the current research on the forefront of technology. Each one of them acquired valuable experience in undertaking significant programming tasks.

Outreach Activities:

Journal Publications

N. Zhang, A. Mukherjee, D. Adjeroh and T. Bell, "Approximate Pattern Match Using the Burrows-Wheeler Transform", Proceedings Data Compression Conference, p. 458, vol. , (2003). Published

D. Adjeroh, M. Powell, N. Zhang, A. Mukherjee and T. Bell, "Pattern Matching on BWT Text: Exact Pattern Matching", IEEE Transactions on Computers, p. , vol. , (). Submitted

Books or Other One-time Publications

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

With the huge amounts of data often involved, efficiency considerations (in terms of both space and time) make it important to consider ways to keep the data in the compressed form for as much as possible, even when it is being searched. Our objectives in this proposal is to develop techniques for compressed domain pattern matching, i.e. to search for the required information directly on the compressed data, with minimal (or no) decompression. We proposed a class of new compressed domain pattern matching algorithms that exploits the sorted contexts of the Burrows-Wheeler transform. Our proposed methods are applicable to both text and images compressed based on the BWT.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

At the University of Central Florida, we have taught a graduate level course entitled 'CAP5937:Multimedia Compression on the Internet'. This has a new URL location: <http://www.cs.ucf.edu/courses/cap5015/>. This particular topic has grown directly out of the research that we have been conducting for the last couple of years on data compression. Lecture topics have included both text and image compression, including topics from the research on the current NSF grant. The course has now being revised for next offering in Fall of 2003.

At the West Virginia University, two graduate courses that relate to the project have been ongoing. CS591K û Multimedia Systems have sections that discuss applications of compression to images and general multimedia data. EE591 û Introduction to Information Theory have sections that treat the fundamental basis and limitations of data compression. In the current report period, both courses (CS591K, Fall 2002; EE591 Spring2003), have involved projects on lossless image compression, which are very relevant to the project.

Contributions Beyond Science and Engineering:

Text searching is an important problem in diverse areas of human endeavor. With the emergence of the Internet, and the pervasive nature of email communication, we are just starting to appreciate the importance of fast text searching û for both exact and inexact. With time, again thanks to the Internet and other improvements in communications and storage technology, images will become much more prevalent as they are today. And thus, people will want to search on images with the same ease that they us to search text data. Thus, the results from the proposed work will have impact far beyond the realms of computer science, or engineering, but in different aspects of or day to day activities as a society.

Special Requirements

Special reporting requirements: None

Change in Objectives or Scope: None

Unobligated funds: less than 20 percent of current funds

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners

Activities and Findings: Any Outreach Activities

Any Book

Any Web/Internet Site

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development