

Annual Report for Period:09/2004 - 09/2005**Submitted on:** 06/13/2005**Principal Investigator:** Mukherjee, Amar .**Award ID:** 0312724**Organization:** U of Central Florida**Title:**

ITR Collaborative Research: Compressed Search and Retrieval for Very Large Text and Image Repositories

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's primary research is concerned with text compression and compressed domain pattern matching for text. Mr. Zhang finished his Ph. D. dissertation Spring 2005 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 finished his Ph. D. dissertation Spring 2005 under Professor Mukherjee's supervision. Mr. Tao is working under Prof. Mukherjee's supervision

Name: Sun, Weifeng**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Mr. Sun is working on image compression problems using wavelets. He is a Ph. D. student. Mr. Sun is working under Prof. Mukherjee's supervision

Name: Satya, Ravi**Worked for more than 160 Hours:** Yes**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. Although Mr. Satya has participated in data compression research. his dissertation will be in the Bioinformatics area. Mr. Satya is working under Prof. Mukherjee's supervision

Undergraduate Student**Technician, Programmer****Other Participant****Research Experience for Undergraduates****Organizational Partners**

West Virginia University

Professor Don Adjeroh, Co-PI

Title: ITR Collaborative Research: Compressed Search and Retrieval for Very Large Text and Image Repositories

Proposal Number: IIS-0312484

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 student 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 and we acknowledge the partial support from this grant for these efforts.

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

We studied compressed domain 'almost optimal' pattern matching algorithms for multiple patterns on text compressed using LZW algorithm. We studied methods to extend the exact pattern matching algorithms to the problem of compressed domain approximate pattern matching.

For image compression and search, we are studying the use of the BWT contexts in image shape retrieval. In image compression, developed a method for image compression by generalizing the context modeling approach used by the PPM (prediction by partial matching) family of algorithms. To improve the overall compression scheme, we also developed an adaptive edge-based prediction for natural images.

Findings:

We have developed a novel compressed pattern matching algorithm for multiple patterns using Aho-Corasick algorithm. The algorithm takes $O(mt+n+r)$ time with $O(mt)$ extra space, where n is the size of the compressed file, m is the total size of all the patterns, t is the size of the LZW trie and r is the number of occurrences of the patterns. The algorithm is particularly efficient when being applied on archival search if the archives are compressed with a common LZW trie.

We continued our investigation on the problem of approximate pattern matching, especially the k -mismatch problem, and the k -approximate matching problem. We addressed the problems using the exact pattern matching algorithms as our starting point. We use a two-stage approach. In stage one, we use the exact pattern matching algorithms to hypothesize areas with potential approximate match to the pattern. In stage two, we verify the potential matches using Ukkonen's DFA algorithm.

All the above algorithms have been implemented and extensive experiments have been conducted to compare our results with the best existing algorithms. The experimental results show that our compressed domain pattern matching, k -mismatch and approximate algorithms are among the best algorithms and very fast. Two doctoral dissertations have been completed based on this work (see publications later).

We have also completed an extensive report on comparison of BWT (Burrows-Wheeler transformed) based approaches to string pattern matching. A paper based on this report has been published in a reputable journal.

For image compression, we have studied 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.

We have started work on ideas for two-dimensional image compression using methods motivated by the PPM algorithm. We call this PPAM -

Prediction by Partial Approximate Matching. We introduced the notion of approximate contexts. The intriguing idea is that these approximate contexts (which are lossy and not exact) could be used to perform lossless compression of an image. The empirical performance was shown to be better than those of standard lossless compression algorithms, such as JPEG-LS and JPEG2000, and high performance systems such as CALIC and EDP. To further improve general compression, we developed a method for adaptive edge-based prediction. Here, we analyze each given position in the image, and based on the nature of the edges around this point, the system adaptively selects a simple edge-based prediction, or the more time consuming Least Square prediction. The result is an improved compression performance with minimal computation.

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

Tim Bell, Andrew Firth, Amar Mukherjee and Donald Adjeroh, "A Comparison of BWT Approaches to String Pattern Matching", Software-Practice and Experience, p. 1, vol. 35, (2005). Published

Tao Tao and Amar Mukherjee, "Pattern Matching In LZW Compressed Files", IEEE Transactions on Computers, p. , vol. , (2005). Accepted

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

T. Tao and Amar Mukherjee, "Multiple-Pattern Matching in LZW Compressed Files Using Aho-Corasick Algorithm", Proc. Data Compression Conference, p. 482, vol. , (2005). Published

Tao Tao and Amar Mukherjee, "LZW Based Compressed Pattern Matching", Proc. Data Compression Conference, p. 568, vol. , (2004). Published

Tao Tao and Amar Mukherjee, "Multiple-Pattern Matching For LZW Compressed Files", International Conference on Information Technology: Coding and Computing, p. 91, vol. , (2005). Published

Y. Zhang and D. Adjeroh, "Prediction by partial approximate matching for lossless image compression", Proceedings Data Compression Conference, p. 494, vol. , (2005). Published

Y. Zhang, R. Parthe and D. Adjeroh, "On compression of DNA microarray images", Proceedings IEEE Conference on Computational Science Bioinformatics, p. , vol. , (2005). Accepted

Books or Other One-time Publications

Amar Mukherjee, N. Zhang, T. Tao, R. Vijaya Satya and W.Sun, "Search and Retrieval of Compressed Text", (2005). Book, Published
 Editor(s): A. Hurson
 Collection: Advances in Computers
 Bibliography: Vol.63

Nan Zhang, "Transform Based and Search Aware Text Compression Schemes and Compressed Domain Text Retrieval", (2005). Thesis, Published

Bibliography: Doctoral Dissertation, University of Central Florida

Tao Tao, "Compressed Pattern Matching for Text and Images", (2005). Thesis, Published

Bibliography: Doctoral Dissertation, University of Central Florida

Web/Internet Site

URL(s):

<http://vlsi.cs.ucf.edu/>

Description:

This site presents a complete description of our research and activities conducted under all the NSF sponsored research grants on data compression and compressed domain pattern matching. It posts all our publications electronically, it makes available all our annual and final reports submitted to NSF and people involved in the projects. If you follow the old website link (http://vlsi.cs.ucf.edu/old_root/index.html), it leads to M5 Online Compression Utility site where all our compression software and other compression software downloaded from different sites are made available online.

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 2005.

At the West Virginia University, two graduate courses that relate to the project have been ongoing. EE558 û Multimedia Systems have sections that discuss applications of compression to images and general multimedia data. EE568 û Information Theory have sections that treat the fundamental basis and limitations of data compression. In the current report period, both courses (CS558, Fall 2003; EE568 Spring2004), have involved projects on lossless image compression, which are very relevant to the project. EE568 also involved projects and written reports on general data compression.

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 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 use 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 our day to day activities as a society.

Special Requirements**Special reporting requirements:**

The work on compressed domain pattern matching using LZW compression was conducted by Amar Mukherjee and his students at the University of Central Florida. The work on image compression using BWT approach was done by Professor Donald Adjero and his students at the West Virginia University. The work on comparison of BWT approaches to string pattern matching was conducted by Professor Bell and his students at the University of Canterbury at Christchurch, New Zealand with input from all three groups of researchers. The work on exact pattern matching, k-mismatch and approximate pattern matching were conducted jointly by Professors Amar Mukherjee and Donald Adjero with inputs from Professor Bell and student groups working under the PI's of the project.

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:

Activities and Findings: Any Outreach Activities

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development