

Annual Report for Period:09/2003 - 09/2004**Submitted on:** 06/24/2004**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 is a Ph. D. student working on this project. His primary research is concerned with text compression and compressed domain pattern matching for text. Mr. Zhang 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:** Yes**Contribution to Project:**

Mr. Sun is working on compressed domain pattern matching and text data mining. He is a Ph. D. student. Mr. Sun 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

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 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 and we acknowledge the partial support from this grant for these efforts. Also, we are discussing the possibility of linking up our online compression utility website vlsi.cs.ucf.edu with the Canterbury Corpus website.

Activities and Findings

Research and Education Activities:

We improved our compressed pattern matching algorithms that achieve the complexity bounds of linear exact pattern matching algorithms for text compressed with the sorted context approach. The algorithms are based on binary search, and q-gram filtration, and make use of the sorted contexts generated by the Burrows-Wheeler transform.

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 studied the use of the BWT contexts in image shape retrieval. We have also started work on image compression using 2D-BWT.

We developed compressed pattern matching algorithms for three popular lossless image compression schemes: lossless JPEG, CALIC and JPEG-LS.

We developed new algorithms for compressed domain pattern matching based on LZW compression algorithms.

We have also developed a new two-pass modified LZW algorithm that support fast random access and partial decoding of the compressed text with application to information retrieval systems.

Findings:

We had previously developed two techniques for searching BWT transformed text using Boyer-Moore algorithm and binary search.

We improved the QGRAM algorithm into a new and improved algorithm, called QGREP. This improved on the sub-linear search time of the binary search and QGRAM algorithms. We provided a more rigorous complexity analysis of the performance of the proposed methods. Also, a detailed empirical comparison of the performance of the proposed methods, with other proposed methods was performed.

We investigated 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.

We developed a variant of the MFT algorithm that makes it more suitable for searching on the BWT transformed text. We analyzed the space and time complexity of the approximate pattern matching algorithms we propose. We also evaluate the empirical performance of the algorithms. Experimental comparison with other proposed methods for approximate pattern matching is currently underway.

We report our work on compressed pattern matching in LZW compressed files. The work is based on Amir's well-known 'almost-optimal' algorithm but has been improved to search not only the first occurrence of the pattern but also all other occurrences. The improvements also include the multi-pattern matching and a faster implementation for so-called 'simple pattern', which is defined as 'a pattern with no symbol appearing more than once'. Extensive experiments have been conducted to test the search performance and to compare with not only the 'decompress-then-search' approach but also the best available compressed pattern matching algorithms, particularly the BWT-based algorithms. The results showed that our method is competitive among the best algorithms.

We have developed modified LZW algorithms that support fast random access and partial decoding to the compressed text with application to information retrieval systems. The proposed approach, instead of fully decompressing the text and outputting the results selectively, allows random access and partial decoding of the compressed text and displaying only the relevant parts. The compression ratio is not degraded and even be improved slightly using the modified LZW algorithm. Preliminary results on the time and storage performance are encouraging.

For image search, we have studied the use of the sorted contexts produced by the BWT in efficient search and retrieval for very large shape databases (containing millions of shapes). We showed that the sorted contexts could in fact be exploited to provide some form of invariance in the retrieval, for instance, in the consideration of changes in orientation, and view point. The proposed methods were shown to provide an improved theoretical complexity, when compared with traditional approaches.

We have started work on ideas for two-dimensional Burrows-Wheeler Transform. The major objective is to use this for lossless (and perhaps lossy) image compression. We are also investigating methods that could exploit the results from such 2D transformations in image search. The 2D approach could also provide a more effective approach for general text compression. We are also studying how to exploit the many spatial contexts that can be observed in an image, and how these can be used in effective compression of an image and for the support of later search on the compressed image. One approach being investigated here is the use of block-based alphabets, rather than individual symbols from the prediction errors.

We developed compressed pattern matching algorithms for the three popular lossless image compression schemes: lossless JPEG, CALIC and JPEG-LS. We show that these algorithms can be made search-aware by minor modifications. We also present a modified JPEG-LS algorithm and the corresponding searching algorithm. Experimental results show that our method, in comparison with the 'decompress-then-searching' method, has nearly 30% improvement in search time for most natural images. The modified JPEG-LS algorithm also has shorter encoding and decoding time, with an improvement of about 12-15% and 8-12%, respectively, for most natural images. The tradeoff is the decrease of compression by about 2% -8%. To the best of our knowledge, this paper

reports the first JPEG-LS compressed matching algorithm.

Distribution of Activities

The work on pattern matching on BWT encoded text was performed jointly by the two PI's in collaboration with Professor Tim Bell of University of Canterbury, New Zealand. The work on BWT-based image compression was performed mainly at West Virginia University by Professor Adjero and his students. The work on pattern matching on prediction-based lossless image compression schemes was performed mainly at the University of Central Florida by Professor Mukherjee and his students.

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 seminar presentations. 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. Adjero and T. Bell, "Approximate Pattern Matching Using the Burrows-Wheeler Transform", Proceedings Data Compression Conference, p. 458, vol. , (2003). Published

Nan Zhang, Tao Tao, Ravi Vijaya Satya, and Amar Mukherjee, "Modified LZW Algorithm for Efficient Compressed Text Retrieval", Proc.International Conference on Information Technology: Coding and Computing, p. 224, vol. , (2004). Published

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

Tao Tao, Amar Mukherjee, "Compressed Pattern Matching for Predictive Lossless Image Encoding", Proc. International Conference on Distributed Multimedia Systems, p. 120, vol. , (2004). Published

Books or Other One-time Publications

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:

Contributions Beyond Science and Engineering:

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. 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.

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:

Activities and Findings: Any Outreach Activities

Any Book

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development

Contributions: To Any Resources for Research and Education