

**Annual Report for Period:**09/2005 - 09/2006**Submitted on:** 07/29/2006**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 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 worked on several joint papers on compressed domain pattern matching and we acknowledge the partial support from this grant for these efforts. We have been exploring the possibilities of writing a book on Burrows-Wheeler Transform and its applications jointly with Profs. Tim Bell and Peter Fenwick based on our work supported by this and prior NSF projects.

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

For texts, our studies focused on the connection between the BWT and other text index structures, such as suffix trees and suffix arrays.

We developed new algorithms for the direct suffix sorting problem, with far reaching implications for the BWT itself, and for its applications in BWT-based compressed search, and in BWT-based search-aware text compression.

For image compression and search, we are studying the use of the BWT contexts in image shape retrieval. In image compression, we further studied our proposed method for image compression based on a generalization of the context modeling approach used by the PPM (prediction by partial matching) family of algorithms.

**Findings:**

A major bottleneck in BWT-based analysis (both in the original transformation, and in its applications) is the suffix sorting stage. Given an input sequence, the suffix sorting problem is to construct a lexicographic ordering of all suffixes in the text. This is the most time consuming stage in BWT construction, but also establishes the very important connection between the BWT and other index structures, such as suffix trees and suffix arrays.

We studied this connection between the BWT, suffix trees and suffix arrays. We developed two sort-and-merge direct suffix sorting algorithms, based on a recursive even/odd partitioning of the input text. While traditional suffix sorting relies on an initial construction of a suffix tree, a direct suffix sorting algorithm computes the suffix array from the given sequence directly, without using the suffix array. The first algorithm uses a symmetric consideration of the partitions at each recursion step. It sorts all the suffixes in linear time and space on average, but in  $O(n \log n)$  worst case. The second algorithm uses a non-symmetric consideration of

the partitions at each recursion step, combined with radix sort to construct the suffix array. The result is a direct suffix sorting algorithm with a linear time and space complexity. This direct suffix sorting approach is important in BWT compression and analysis, since the suffix tree is not really needed, if we can get the suffix array. The approach proposed, especially, the second algorithm also provides important data structures that will be used in our investigation of BWT-based search-aware compression.

We developed a variant of the MTF 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.

For Images, we are continuing our investigation on the use of BWT for efficient search and retrieval for very large shape databases. The sorted contexts in BWT could in be exploited to provide some form of invariance in the retrieval, for instance, in the consideration of changes in orientation, and view point. At this point, we are focusing on 2D shapes. We will study the theoretical complexity of this approach, and compare with traditional approaches.

Earlier under the project, motivated by PPM (Prediction by Partial Matching) used for text, we introduced the notion of approximate contexts, and based on this proposed PPAM  $\hat{u}$  Prediction by Approximate Partial Matching, a new method for context modeling in image compression. The intriguing idea is that these approximate contexts (which are lossy and not exact) could be used to perform lossless compression of an image. We have performed a deeper study of PPAM and approximate contexts, and extended the method to compress both natural images, and application specific images, such as micro arrays.

We have performed a detailed theoretical analysis of the complexity of the PPAM model, and proposed algorithms and efficient data structures for maintaining and searching approximate contexts. The empirical performance was shown to be better than those of standard lossless compression algorithms, such as JPEG-LS and JPEG2000, and competitive with other high performance systems such as CALIC and EDP.

The connection between the BWT and PPM, and between PPM and PPAM could provide further ideas on improving both efficiency and compression performance.

### **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.

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. 1217, vol. 35, (2005). 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

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. 128, vol. , (2005). Accepted

Tao Tao and Amar Mukherjee, "Pattern Matching In LZW Compressed Files", *IEEE Transactions on Computers*, Vol.54, No. 8, August, 2005, pp.929-938, p. 929, vol. 54, (2005). Published

Tao Tao and Amar Mukherjee, "Multiple-Pattern Matching in LZW Compressed Files", *Proc. International International Conference on Information Technology, Coding and Computing*, p. 91, vol. 1, (2005). Published

F. Nan and D. A. Adjeroh, "An Algorithm for suffix sorting and its applications", *Proc. Computational Science Bioinformatics*, Stanford CA, p. 1, vol. , (2006). Accepted

R. Vijaya Satya and A. Mukherjee, "An optimal algorithm for perfect phylogeny haplotyping", *Journal of Computational Biology*, p. 897, vol. 13, (2006). Published

Weifeng Sun and Amar Mukherjee, "Generalized Wavelet Product Integral for Rendering Dynamic Glossy Objects", *Proc. SIGGRAPH*, p. , vol. , (2006). 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

Ravi Vijaya Satya, "Algorithms for Haplotype Inference and Block Partitioning", (2006). Thesis, Published  
Bibliography: Doctoral Dissertation

### 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](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.

New algorithms for compressed domain pattern matching based on the Burrows-Wheeler transform (BWT) have been developed. The sorted context of the BWT transform gives rise to very efficient binary, Boyer-Moore and q-gram based search strategy for performing exact match, and k-mismatch and approximate pattern matching operations. The search times of binary search, suffix arrays and q-grams have been improved by around 20%, as well as reduce the memory requirement of the latter two by 40% and 31%, respectively. Our k-mismatch and approximate search algorithms out perform the well known agrep algorithm. Search-aware compression schemes that support compressed-domain search directly on the compressed data with minimal or no decompression has been developed. Search engines have been developed using compressed keywords or inversion dictionaries to expedite search operations for terabyte scale text and image repositories. For image compression, new pattern matching algorithms in LZW compressed files and two-dimensional pattern matching algorithms on images compressed by JPEG-LS compression algorithm have been developed. The multiple-pattern matching algorithm is the fastest when the number of patterns is not very large.

A new paradigm of two-pass compression algorithms very effective for applications in archival information retrieval systems has been advanced. This method yields search-aware compression for compressed domain search with random access property suitable for parallel/distributed system. The methodology has been implemented with respect to the widely used LZW algorithm. The idea is then extended to lossless image compression. In particular, a two-pass variation of

JPEG-LS using this principle has been developed and implemented. The methodology has been successfully applied for search and retrieval for very large text and image repositories.

We have also explored the techniques of wavelet based compression of images to real time rendering of dynamic image scenes which has opened up a new avenue for research.

Thus, our progress of work so far has been consistent with our stated objectives. This is also reflected by the number of publications listed in this and our previous annual reports, production of doctoral and masters dissertations and our professional recognition as researchers in this area. During next year we will continue to work on compressed domain search and retrieval problems for text and images, and consider extending our methodology to related field of computer graphics. We will complete implementation of the algorithms proposed.

#### **Contributions to Other Disciplines:**

Our work on data compression which is related and partially based on pattern matching and context analysis, has led to undertake some projects in the area of Bioinformatics. Using pattern matching and suffix tree approaches we have developed algorithms for DNA compression and haplotype inference and block partitioning problems.

We have published a few papers and graduated a Ph. D. student in this area with partial support from this grant.

#### **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 is scheduled for Spring of 2007.

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 and image 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 and imaging science. 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 and process images with the same ease that they us to search text data. Thus, the results from the proposed work will have impact beyond the realms of computer science, or engineering.

### **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 Adjeroh 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 Donal Adjeroh 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 Human Resource Development