Final Technical Report for DURIP grant for Infrastructure for Large-Scale Multimedia Information Indexing, Retrieval and Organization - grant number # F49620-99-1-0138

Principal Investigator: R. Manmatha
Co-Principal Investigators: J. Allan, W. B. Croft, J. Callan

A Introduction

The equipment acquired through DURIP has enabled NSF's National Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts to further its mission of carrying out leading research in the areas of organizing, classifying and retrieving text and more recently multimedia. CIIR's goal is to provide organizations and individuals tools to organize information, so that relevant information may be easily obtained and new relationships discovered. The center has a number of grants from the defense department (see the list in section D) and the DURIP equipment has aided in the research performed as part of these grants. The Center has trained over 60 graduate students and 50 undergraduate students to do leading edge research and development in information retrieval. Currently, the center has 3 faculty, 10 technical staff (including one posdoctoral research associate) and 16 graduate and 12 undegraduate students are being trained.

We list below the equipment acquired and follow with a description of the research performed using this equipment.

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-
0303

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 30 May 00	3. REPORT TYPE AND DA... Final Technical Report 1 Mat 99 to28 Feb 00

4. TITLE AND SUBTITLE
Infrastructure of Large Scale Multimedia Information Indexing, Retrieval and Organziations

5. FUNDING NUMBERS
F49620-99-1-0138

6. AUTHOR(S)
R. Manmath

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts
Office of Grant & Contract Administration
Goodell Building Room 408
Amherst, MA 01003

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
801 N. Randolph St, Rm 732
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The equipment acquired through DURIP has enabled NSF's National Centem for Intelligent Information Retrieval (CIIR) at the University of Massachusetts to further its mission of carrying out leading research in the areas of organizing, classifying and retrieving text and more recently multimedia. CIIR's goal is to provide organizations and individuals tools to organize information, so that relevant information may be easily obtained and new relation-ships discovered. The center has a number of grants from the defense department (see the list in section D) and the Db-RIP equipment has aided in the research performed as part of these grants. The Center has trained over 60 graduate students and 50 undergraduate students to do leading edge research and development in information retrieval. Currently, the center has 3 faculty, 10 technical staff (including. one postdoctoral research associate) and 16 graduate and 12 undegraduate students are being trained.

14. SUBJECT TERMS

15. NUMBER OF PAGES
10

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

B Equipment Acquired

Equipment Description	P.O. No.	Amount $	Vendor
Server - 4 Xeon 450 Mhz cpu's, 2GB RAM, 2TB RAID 5 disk	113,328	Y214605	Gateway
Server - 4 400 Mhz ultrasparc processors, 4 GB RAM 800 GB RAID disk,	48,347	Y201802	Sun Microsystems
2 Spectra-Logic TreeFrog-618 Tape Libraries 15-slot AIT-2 (36GB) drives	44,290	Y219790	Sanstart Inc.
Sun 218 GB Disk SubSystem	13,960	Y214608	Sun Microsystems
6 Gateway PC workstations	16,707	Y214605	Gateway
1 Gateway PC workstations	1,840	Y220226	Gateway
1 Gateway PC workstations	1,741	Y214888	Gateway
1 Sun ultra-10 workstation	4,320	Y214608	Sun Microsystems
1 IBM thinkpad laptop	5,072	Y185935	Insight
1 Sony Vaio PC	2,653	Y21119	Insight
1 Sony Vaio laptop	4,103	Y214283	Insight
1 Sony Digital VCR	4,103	Y214287	Valley Comm
1 Compact VHS Cam Corder	393	Y214229	Valley Comm
1 Sony Monitor/VCR	446	Y214867	Valley Comm
1 Sony Digital Video Camera	1092	Y214867	Valley Comm
1 HP Officejet multifunction printer/scanner/copier	715	Y185935	Yes Computers
3 HP Officejet multifunction printer/scanner/copier	1682	Y213925	Insight
1 HP Officejet multifunction printer/scanner/copier	500	Y214883	Insight
1 Olympus Digital MegaPixel Camera	848	Y213934	Insight

Table 1: List of Equipment Purchased.

Total for Equipment purchase $265,557
Grant 5-28159 $245,577
Matching Funds 1-02663 $20,000
Unspent $2,023

B.1 Changes between Proposed and Actual Equipment Acquired

There were a few changes between the list of proposed equipment and the actual equipment acquired. They were caused partly by the fact that the cost of equipment had come down by the time the grant started as well as the fact that sum of our needs had changed.

The changes include:

1. The quad processor server was originally estimated to cost $135K with 2GB of ram and 1 TB of RAID 5 disk space. We were able to obtain a quad processor server with 2 GB of ram and 2 TB of RAID 5 disk (double the original disk space) for a little more than $113K.
2. The savings from the above server was used to purchase an additional tape library so that both servers could be backed.
3. We decided to not buy the CD Jukebox because we found that the additional disk space could be used instead.
4. We also decided not to buy the MPEG2 hardware encoder and instead bought a digital video cassete player to store the output of the digital video camera.
5. We were able to buy additional workstations and PC's because of price drops.
6. Instead of the two scanners we bought multifunction devices (allowing scanning, printing and copying). These turned out to be useful for research purposes allowing us not only to scan colored images for research. We were also able to buy additional devices for research use because of price drops of other equipment.

C Accomplishments

The center has accomplished significant research advances in the areas of distributed information retrieval, information filtering, topic detection, multimedia indexing and retrieval, document image processing, terabyte collections, data mining, summarization, resource discovery, interfaces and visualization, and cross-lingual information retrieval.

C.1 Research into indexing and organizing images, scanned documents and video

The equipment obtained from this grant has been used for extending our research in the areas of indexing and retrieving images based on their content. It has also been used to continue our work on detecting text printed against image backgrounds as well as work on indexing handwritten text using word matching. Finally, the equipment has been used to start some new and innovative work in video indexing and retrieval. This work is supported by a grant from DARPA and the US Patent and Trademark Office as well as an interagency Stimulate grant.

We will briefly discuss some of these areas:

C.1.1 Image Indexing and Retrieval

At CIIR we have developed techniques to index and retrieve images using their content. We have focused on methods to retrieve images based on appearance (the shape of the greylevel surface of the image) as well as on color.

We have developed fast methods to retrieve images based on their global similarity to a query images. Global similarity is determined by comparing the distributions of local features such as 3D curvature and phase at multiple scales [6, 10]. Previously this approach has been used to retrieve images from a database of 1500 images of assorted images (cars, faces, apes etc.) as well as on trademark databases of size 2000 and 60000 [6, 10]. The work has also been used to retrieve trademarks using both image and text retrieval. Retrievals were done in real time. Recent additions to our work have included comparing the performance of a moment-based technique and our technique on a set of artificially generated shapes as well as on a set of more than 10000 geometric trademarks from the British patent office. The comparison has shown that our techniques perform better. This work has been published in [9]. We have also created a user interface for the purpose of generating relevance judgements for the British trademarks. We hope to obtain these judgements this summer in collaboration with the University of Glasgow. This approach to retrieving images based on distributions of curvature and phase is now being exploited to do face recognition.

We have also used color to retrieve images based on domain specific constraints. The aim of this work is to index images in a database using features computed from the object of interest only, instead of the whole image. This provides improved retrieval performance since the effect of background elements is eliminated. The main problem encountered in this task is the segmentation of the region of interest from the background. In the case of a database of flower patent images a lot of color domain knowledge was available (for example, flowers are rarely black, gray, green or brown) which was exploited to detect flower regions. The flowers were then indexed by color. Other domain knowledge such as, the background colors are usually visible along the periphery of the image can also be used to produce a list of candidate background colors and check each hypothesis by eliminating the color and evaluating the remaining image.

In more recent work we are expanding our flower patent database from 300 images to a much bigger set of about 5000 images. We note that our segmentation is quite accurate (in the initial database of 300 images there were only 2 errors). Our current work involves a database of bird images. In addition to automatically detecting and eliminating possible background colors from the image, we use an edge image of the bird to detect the focus of attention of the image. Combining the edge information with color-based background elimination produces good results (correctly identified bird regions) in most cases. The work with flower patents is reported in [4].

C.1.2 Text-in-images

Many images often contain associated textual material which may be useful for indexing. This textual material may be present within the image itself—for example, the number 33 on a Celtics basketball player's jersey identifies the player as Larry Bird —or the text may be in the form of captions, text material in close proximity to the picture, or closed captions for a video clip. This latter type of textual material can be indexed using INQUERY provided it is in ASCII. These cannot be normally handled by convential optical character recognition (OCR) systems).

We have developed methods at CIIR [11, 12]. The techniques utilize certain characteristics of text which differentiate itself from most of the image. Text shows spatial coherence—i.e., rarely are individual characters found by themselves; rather they are found in association with other characters forming words or groups of words. In addition, text has definite frequency characteristics—i.e., there is a characteristic spacing between letters. The spacing between letters in a word is less than that between adjacent words. Adjacent lines of text are often separated by spaces. These characteristics are exploited in our methods to detect text in images [11]. The text is essentially detected in two stages. The first stage exploits the frequency characteristics and assumes that text may be treated as a special kind of texture. A standard texture segmentation scheme is used to delineate areas of the image which are possibly text. This is done by filtering the image with Gaussian derivatives and clustering the filtered outputs. The second stage involves strokes generated using vertical connected edges. Constraints are used to eliminate strokes which are unlikely to be part of a text segment. For example, the strokes specify that other strokes must be present close by, and that the strokes must be approximately aligned (see [11] for more details). Neighboring strokes are clustered into chips. The chips are then filtered using aspect ratio and size and the chips left are filtering are considered to be text. Larger font sizes can be handled by using scale space ideas. That is, the same procedure is applied to the image at multiple resolutions and the results from different resolutions fused.

Our text detection work has been published in a journal [12]. The work describe in this journal includes more extensive evaluation than other work reported in the literature. In recent work we have been experimenting with testing this algorithm on additional databases. For example, we are looking at how text may be detected in trademark images.

C.1.3 Word Spotting: Indexing Handwritten Documents

The indexing of handwritten documents is another important area of research. Examples of such handwritten documents include the early Presidential papers at the Library of Congress and the W.E.B. DuBois collection at the University of Massachusetts Library. Currently there are few tools for automating the indexing of such manuscripts. Handwritten manuscripts and even unusual machine fonts do not give good results when an OCR is used. In pioneering work [5] we showed the feasibility of producing an index (like the index at the back of a printed book) for single author handwritten manuscripts. The method essentially consists of segmenting a page into words and then matching the words to find lists of similar words. A person then assigns ASCII equivalents for one member of each list. The machine then automatically classifies all members of the list and links them to the appropriate page. We call this technique word spotting - since it involves finding words similar to a given word.

In recent work, we have examined better ways of segmenting such handwritten pages into words. Handwriting unlike machine printed text is not as regularly spaced. Further, ascenders and descenders on different lines may meet. Previous work on segmenting handwritten words has largely been confined to special situations like addresses on postal envelopes. Our technique for segmenting handwritten words involves segmenting the lines first followed by word segmentation. Line segmentation is carried out by projecting the intensities onto the vertical and then examining maxima and minima (a kind of projection profile technique) . Word segmentation is carried out by smoothing the image with anisotropic Gaussian derivatives and then looking for blobs - the blobs constitute the words.

The work on segmenting handwritten manuscript pages has been tested on some of George Washington's manuscripts and is reported in [7]. It is shown that 86% of the words can be automatically segmented using this technique. We will be testing this on the complete collection of George Washington's manuscripts as well as on some manuscript pages from Thomas Jefferson.

C.1.4 Video Indexing and Retrieval

We have begun work on indexing and retrieving compressed (MPEG) video by creating a new representation of video. Current MPEG encoders do not respect shot boundaries. In MPEG video, I or anchor frames are stored and differences between the I frames and neigboring frames are also stored. Thus, the same video shot may be compressed in different ways depending on which video stream it is a part of. In particular, the I frames and the number of I frames for the same video shot depend on the rest of the video. Recently, we have created a new MPEG encoder which always creates the same I frames as well as the same number of I frames for a particular video shot. This unique representation of a video shot makes it easier to index and retrieve compressed videos. We are now currently exploring how to exploit this representation for retrieving videos.

We are also exploring a new user interface for browsing videos which consists of a number of videos all running simultaneously. We believe the user can rapidly select the video of interest from this interface. Although such a representation has not been used for video browsing or retrieval, a similar representation is found in the control room of a television station where many monitors are playing simultaneously and the editor chooses the appropriate shot to show.

C.2 Research into Organizing, Browsing, Visualizing, Indexing and Retrieving Text

The equipment has also aided in research into organizing, browsing, visualizing, indexing and retrieving text. This work is funded amongst others by a DARPA grant on topic detection and tracking as well as a more recent DARPA grant for researching and developing tools for rapidly adaptable translingual information retrieval and organization.

We now briefly discuss some of the work in this area;

Topic detection and tracking (TDT) is the study of event based information organization tasks. These include segmenting the audio of broadcast news shows into single-topic stories, detecting the onset of a new topic in the stream of news, grouping stories together into topically-coherant clusters, tracking follow-up stories given a few stories on a single topic, and deciding whether or not two randomly selected stories discuss the same topic.

As part of this work we have shown that first story detection in TDT is hard when it is based on tracking stories. It is hard because requiring high-quality first story detection requires improving tracking effectiveness to a degree that experience suggests is unlikely. This work has been submitted for publication [2]. Many research tasks in TDT have counterparts in more traditional information retrieval (IR) research. The TDT tracking task can be compared to the IR filtering task and recently we have shown that they have nearly identical effectiveness [1]. CIIR's system and results on the rigorous TDT evaluations conducted is discussed in [3]. We also discuss two methods for learning parameters automatically for event classification in broadcast news. This is discussed in [8].

We have also been recently awarded a grant for developing new algorithms and tools for effective translingual information retrieval and organization. This work emphasizes using readily available language resources, and the ability to rapidly adapt to new languages. The core of our approach is the use of bilingual dictionaries or lists of word (phrase) pairs from two languages, together with statistical language models derived from corpora. We are currently researching, implementing and evaluating a number of approaches in this area.

D List and Abstract of Grants for Which Equipment was Used

1. Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

   Defense Advanced Research Projects Agency/ITO and United States Patent and Trademark Office, contract #F19628-95-C-0235, PI: Bruce Croft, Amounts $2,592,913

   The aim of this grant was to develop techniques for browsing, discovery and search in large distributed databases. Although the techniques are applied to patent and trademark data, the research is relevant to large distributed databases of text, images and scanned documents. The grant has a number of components. The first one involves research and development into searching and browsing patent text data. It involves finding how As part of the grant a system is being researched and developed for searching patent text data. There is also ongoing research into how patents may already be classified into new categories or misclassified patents may be detected. Part of the work involves working with images and scanned documents. This involves research into how Specifically, we are researching how a trademark retrieval system may be built by combining image and text retrieval.

2. Multimodal Indexing, Retrieval and Browsing, Combining Content Based Image Retrieval with Text Retrieval

   Part of the Interagency STIMULATE program. Agencies involved are the National Science Foundation, Central Intelligence Agency, Defense Advanced Research Projects Agency, and National Security Agency, contract #IRI-9619117, PIs: James Allan and R. Manmatha, Amounts $749,121.

   This research involves developing techniques for retrieving images using content based indexing, text present in the image as well as text associated with the image. Thus the thrust of the research is on developing new techniques for finding images "similar" in appearance or color to a query image and for finding new techniques to detect text against image backgrounds. At a later stage, the work will also involve research into multi-modal retrieval. Multimodal retrieval cannot simply involve combining image and text retrieval. One must also decide which is more appropriate for a given query. Research is also being carried out to develop new interfaces for visualizing the outputs of information retrieval systems. Since the user is usually involved in interacting with an information retrieval system, the use of a good user interface is essential.

3. Topic Tracking in Broadcast Data Using Hidden Markov Models

   TDT: (Subcontract from Dragon Systems, Inc.) Defense Advanced Research Projects Agency, contract #MDA904-97-C-0408, PIs; Bruce Croft and James Allan, $651,778.

   This contract explored event-based organization of broadcast news in a program called Topic Detection and Tracking. The project focused on defining rigorous evaluation techniques for five TDT tasks: segmenting the audio of broadcast news shows into single-topic stories, detecting the onset of a new topic in the stream of news, grouping stories together into topically-coherant clusters, tracking follow-up stories given a few stories on a single topic, and deciding whether or not two randomly selected stories discuss the same topic. This work was done in the context of a multi-university and -company research program that included rigorous "competitive" evaluations.

4. Tools for Rapidly Adaptable Translingual Information retrieval and Organization

   DARPA/SPAWARSYSCEN-SD, Contract #N66001-99-1-8912, PIs; W. Bruce Croft and James Allan, $2,248,899. Period of Performance: 07/01/99 to 06/30/03

   In this contract, we are developing new algorithms and tools for effective translingual information retrieval and organization, with a strong emphasis on the use of readily available language resources, and the ability to rapidly adapt to new languages. The core of our approach is the use of bilingual dictionaries or lists of word (phrase) pairs from two languages, together with statistical language models derived from corpora.

References

[1] J. Allan, V. Lavrenko, and H. Jin. Comparing effectiveness in tdt and ir. Technical Report CIIR Technical Report IR-197, Center for Intelligent Information Retrieval, Computer Science Dept., University of Massachusetts, Amherst, 2000.

[2] J. Allan, V. Lavrenko, and H. Jin. First story detection in tdt is hard. In *submitted to Ninth International Conference on Information and Knowledge Management CIKM)*, Nov 2000.

[3] J. Allan, V. Lavrenko, D. Malin, and R. Swan. Detections, bounds and timelines: Umass and tdt-3. In *presented at the Topic Detection and Tracking Workshop (TDT-3)*, Feb 2000.

[4] M. Das, R. Manmatha, and E. M. Riseman. Indexing flowers by color names using domain knowledge-driven segmentation. *IEEE Intelligent Systems*, 14(5):24–33, 1999.

[5] R. Manmatha and W. B. Croft. Word spotting: Indexing handwritten manuscripts. In Mark Maybury, editor, *Intelligent Multi-media Information Retrieval*, pages 43–64. AAAI/MIT Press, 1997.

[6] R. Manmatha, S. Ravela, and Y. Chitti. On computing local and global similarity in images. In *Proceedings of the SPIE conf. on Human Vision and Electronic Imaging III*, San Jose, CA, Jan 1998.

[7] R. Manmatha and N. Srimal. Scale space technique for word segmentation in handwritten manuscripts. In *In the Proc. of the Second International Conference on Scale-Space Theories in Computer Vision (Scale-Space'99)*, pages 22–33, Sep. 1999.

[8] R. Papka. Learning threshold parameters for event classification in broadcast news. Technical Report CIIR Technical Report No. IR-177, Center for Intelligent Information Retrieval, Computer Science Dept., University of Massachusetts, Amherst, 1999.

[9] S. Ravela and C. Luo. Appearance-based global similarity retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 267–303. Kluwer Academic Publishers, 2000.

[10] S. Ravela and R. Manmatha. On computing global similarity in images. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'98)*, pages 82–87, Princeton, NJ, Oct. 1998.

[11] V. Wu, R. Manmatha, and E. M. Riseman. Finding Text In Images. *Proc. of the Second ACM Intl. conf. on Digitial Libraries DL'97*, pages 1–10, July 1997.

[12] V. Wu, R. Manmatha, and E. M. Riseman. Text finder: An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1224–1129, Nov 1999.