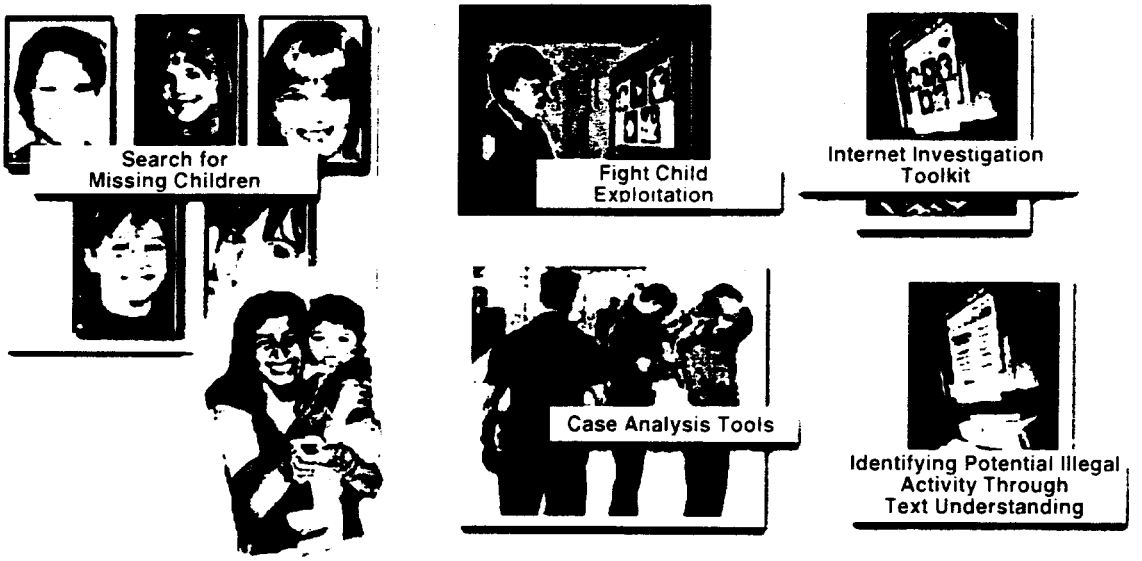
Final Report
for the National Institute of Justice

Technologies for Identifying Missing Children



28 September 2000
NIJ Contract 97-LB-VX-K025

Prepared by

ANSER Analytic Services Inc
Suite 202
1000 Marion Center
Fairmont, WV 26554

Table of Contents

Table of Exhibits

Executive Summary

An Analytic Services Inc. (ANSER) team has been conducting research and development (R&D) under National Institute of Justice (NIJ) Cooperative Agreement 97-LB-VX-K025. This effort has used several advanced technologies to support automation for law enforcement investigators as they search for and recognize relevant information—whether in images, text, or structured data. The initial focus of the effort, "Technologies for Identifying Missing Children," was soon expanded to more broadly support other topics in local, state, and Federal law enforcement. This Final Report covers ANSER's integrated response to selected law enforcement investigators' current and future operational environments (the user "pull") and tools produced by the software developers' technology environment (the technology "push"). The ANSER team has responded to current and future investigative requirements with automated and semiautomated software system solutions. This report examines several topics, including

- Search agents operating on large information sources
- Intelligent software agents (ISAs) as components
  - To collect data from the Internet and other sources
  - To recognize faces, files, text (unstructured data), and structured data
  - To develop, manage, and control ISAs
- Integrating components into prototype demonstrations
- Installation and alpha and beta testing of several systems in conjunction with three (and potentially four) pilot law enforcement agency partnerships

The ANSER team, throughout this grant effort, has developed a keener understanding of both the law enforcement environment and the developer's technology environment. Initially many of the problems seemed to be small. However, the amount of data available throughout the extended law enforcement community soon indicated a demand for highly scalable, robust solutions. This effort addresses many enterprise-level issues, including

- System speed and accuracy
- Recognition algorithms that discriminate face, file, text, and structured data very well
- High-speed data transfer among large databases

The Problem

Locating missing children and stopping the exploitation of children on the Internet—recently evolving, troubling problems in today's society—in general suffer from insufficient law enforcement resources (staff and funding). One emerging area of concern to law enforcement is the Internet and its growing impact on so many people. Today the Internet has many positive attributes, including both the numbers of people connected to it

worldwide and their increasing interaction with and reliance on it as a favored information medium. The many advantages of the Internet include interpersonal communications (e-mail), business transactions (e-commerce), information and learning (instantaneous access to current and historical information), entertainment (downloads of text, music, and video clips), and hobbies (downloads of catalogs and calendars). The growth of the Internet will continue far into the future.

However, the good derived from many of its positive attributes in certain areas has been overshadowed by criminal activity with negative societal applications, including increasing use of the Internet as a source of pornography and particularly a market channel for the child exploitation trade. Because of the Internet's size, relative anonymity, and borderless distribution channels, criminals who deal in child pornography have taken advantage of the Internet and made it the medium of choice to openly market, trade, and sell their illegal products. With the explosion in Internet use, vast amounts of newly available open-source data are known to potentially relate to missing children's cases (including "cold cases"). Open-source data now include home pages created by children, parents, clubs, and schools (including yearbooks published on the World Wide Web).

The overwhelming size of this data source poses an incredible challenge and strain on law enforcement resources, but also can simultaneously offer a tremendous opportunity for pursuing investigations. Although charged with locating and prosecuting cyber-crime, most law enforcement agencies are ill-equipped to fight these types of criminal activity. Law enforcement officers rarely have the staff, time, expertise, or tools for thorough manual searches of Internet data sources ("surfing the Web"). Each investigator needs a great deal more automation to be really effective in this increasingly digital world.

Law enforcement organizations are continually addressing new and expanding requirements with a variety of automated tools to help search, locate, track, and ultimately apprehend criminals; find and recover missing and exploited children; and combat criminal activities, child predation, and many other illicit activities. The required tools not only provide automated processing but also automated reporting of new and potentially relevant information.

The Internet and many other information sources can be leveraged with automated data collection and automated analysis tools to provide prescreened, relevant, and potentially critical information—previously never available—directly at the fingertips of law enforcement personnel.

In many cases today, technology developers are partnering with law enforcement organizations to provide enhanced and automated methods for collecting and analyzing information and producing a force-multiplier effect, freeing law enforcement officers to patrol the streets. Developers must consistently review, design, implement, and upgrade new tools. ISAs are particularly useful in searching for and analyzing data sources to provide key leads in solving crimes.

Software- and Research-Oriented Solutions

In 1997, Congress recognized the importance of solving these problems for law enforcement and ordered a search for promising emerging technologies. By the fall of 1997, the NIJ had awarded ANSER (a public-service research institute) a cooperative agreement for $3.1 million over a period of performance from August 1997 through July 1999 to conduct R&D on technologies for identifying missing children. ANSER formed a team that included commercial face-recognition software developers, the research community, law enforcement agencies, industry experts, and several West Virginia small businesses. This team brought together expertise in developing face-recognition and ISA technologies to apply to solving law enforcement problems.

The two primary goals of this project have been to

- Develop software systems that incorporate automated facial matching and ISA technologies into systems that can aid law enforcement and other organizations in locating missing children and fighting the exploitation of children, while leveraging scarce law enforcement staff resources.
- Demonstrate technologies to local, state, and Federal law enforcement officials to illustrate the technologies' usefulness and readiness.

The project emphasized

- Development of several technologies
  - A face-recognition tool (that also performs face age progression and age regression) to automatically match photographic resources, particularly in missing children's cases.
  - Software based on ISAs, to assist investigators in searching for, monitoring, and analyzing information on the Internet and in internal databases. Several artificial intelligence software systems have been developed as components in the areas of text categorization, data mining, and case analysis (based on expert systems).
  - Agent frameworks to develop, manage, and control lower-level ISAs.
- Integration of many of these complementary software components into several larger prototype systems for test and evaluation in pilot projects.
- Dissemination of a few initial and interim prototype demonstrations to encourage law enforcement agencies to participate as partners in a second NIJ cooperative agreement.

The Facial-Matching Tool

ANSER selected Visionics Corporation and Eyematic Interfaces, Inc., as the two face-recognition providers because of their reputations for excellent technical quality, based on the Face-Recognition Technology (FERET) tests. In addition, the companies were working on solution methods that were complementary to one another, and the companies were prepared to direct their R&D efforts toward automated face recognition

of children. Visionics, with its FaceIt™ technology, also had shown, over several years, a capability of delivering commercial software packages and providing tailored turnkey solutions to law enforcement and other organizations. At the start of this effort, Eyematic Interfaces was preparing its Eyematic Primary Library for delivery.

Throughout the 2-year project, Visionics and Eyematic Interfaces supported ANSER by focusing on R&D to improve face-recognition engines to

- Find single and multiple faces in images
- Tune existing algorithms to find faces of children
- Apply the technology to images with a wider range of variation in resolution and lighting
- Apply the technology to faces with a wider range of variations in age, pose, and facial expression
- Search large databases of faces
- Increase face-recognition functionality with higher performance (accuracy and speed)

Other members of the ANSER development team applied the Visionics and Eyematic Interfaces libraries to developing software components, which led to prototype systems. The first of the ANSER team's face-recognition prototypes, Missing Children Locator Agent (MCLA), implements an easy-to-use, dynamically generated graphical user interface based on a web browser. A case manager or other law enforcement investigator can use MCLA in two separate modes:

1. An investigator can search the Internet for pictures of children to compare against a gallery database of known missing children:
   - The investigator selects parts of the Internet to search for still photographs potentially containing faces of missing children
   - MCLA automatically finds facial probe images on the Internet or an intranet
   - MCLA automatically compares each probe image against gallery image databases of known missing children
   - MCLA automatically generates reports with thumbnail images of the closest gallery image matches to each known missing child
   - The investigator periodically reviews the MCLA reports
2. An investigator can submit a probe image directly, or submit it remotely, over the Internet as input against a gallery database of known missing children. The search tools compare each submitted probe image against gallery image databases of known missing children and then return a report to the submitted query. A user may conduct this search from any stationary or mobile location with an Internet connection from anywhere in the world.

Aging Algorithm Research

Children's faces can change between the time of the last known photograph before they were missing and any image made after they have been missing. Sketch artists have techniques of drawing age-progression and age-regression changes. Automated, computer-generated age progression is one R&D area addressed under this grant. The staffs of the Mount Sinai School of Medicine Laboratory of Applied Mathematics and of Rockefeller University have applied several mathematical methods originally developed for medical imaging problems to the age-progression technique. Several algorithms, including an interpolation method, two Eigenface approaches, and Supersnapshots (potentially a far more powerful technique), were investigated.

The lack of longitudinal (chronological) sequences of faces for each individual child initially limited the development of face age-progression algorithms and verification of their performance. Therefore, Mount Sinai and ANSER collected longitudinal data from yearbooks, relatives, friends, and acquaintances. The ANSER team also collected 87 yearbooks (1988 through 1999) from elementary, middle, and high schools in Marion County, WV, and scanned in and assembled longitudinal sequences of four to seven photographs for at least 500 pupils. These image sets were made available to Mount Sinai and Rockefeller University. Also, the ANSER team videotaped 2,500 elementary and middle school children at 17 schools in Marion County in April and May 1999, and the team plans to repeat videotaping at these same schools at intervals in the future.

Intelligent Software Agent Component Development

The ANSER team also developed six types of software agents that can operate automatically or semiautomatically to support law enforcement officers in investigative activities:

- Internet and database search agents
- Internet monitoring and mapping agents
- Text categorization agents
- Data-mining agents
- Knowledge management agents
- Agents for monitoring and controlling other lower-level agents

These software agents were developed on the basis of the initial requirements from several law enforcement agencies investigating missing and exploited children.

System Integration Toward Demonstration Prototypes

The ANSER team has begun to integrate these software agents to support law enforcement officers in

- Searching for missing and exploited children on the Internet
- Fighting the exploitation of children on the Internet
- Identifying suspects involved in criminal misconduct through surveillance with both live video and videotape
- Integrating pieces of information from one or more sources of both text and structured data

Ten prototype demonstration systems have been developed and integrated:

- **Missing Child Locator Agent, v. 1.0**—The ANSER team developed and integrated a system including Web spiders, face-recognition analysis components based on ISAs, and face-recognition search engines. ISAs continuously search the Web for images and match those they find against a gallery database of faces of missing children. MCLA generates reports for review by law enforcement investigators.
- **Criminal Booking System Search Agent (CBSSA)**—The ANSER team integrated the face-recognition components of the MCLA and IdentiFace systems with a digital booking system. A law enforcement officer can, in almost real time, query the criminal booking database system using images of suspects captured on film or through composite sketches as described by eyewitnesses. With the MCLA Web-based interface, the officer can even run these queries from cross-jurisdictional or intrastate organizations.
- **IdentiFace** (previously known as Video Watchdog)—The ANSER team implemented a video surveillance system that detects, locates, and captures faces of one or several individuals in live video or videotaped scenes; processes the facial image as probes; searches and compares the processed face templates to match against local or remote databases; identifies the individual; reports the results in real time to the investigating officer; and provides a user-friendly interface by which the law enforcement officer can query the database while adding filtering and receiving additional text information.
- **Child Online Pornographic Image Eradication System (COPIES), v. 1.0**—The ANSER team developed a system of ISAs integrated with digital signature utilities for autonomously finding instances of known child pornography and reporting the results specifying when and where COPIES has found a copy. COPIES generates reports for review by law enforcement officers to focus their investigations on websites known to disseminate child pornography. COPIES conducts autonomous online investigations, searching for previously confiscated or known child pornographic images. This system could help eradicate the open exchange of child pornography now available on the Internet and significantly reduce child exploitation.

- **Digital Signature Registrar (DSRegistrar), v. 1.1**—The ANSER team developed DSRegistrar as a complementary program to COPIES to convert large collections of known child pornographic image files to digital signatures.
- **Internet Investigator, v. 1.0a**—This desktop application includes a low bandwidth–intensive modem connection (as low as 28.8 kbaud) and interactive agents useful for interacting with Internet-based resources (for example, WhoIs, Finger, and DNS Lookup).
- **Case Analysis Software Agent (CASA)**—CASA is a knowledge management system that tracks data about entities, such as criminals and/or vehicles, in a highly flexible format. Database records can have differing numbers of fields, and fields can be added and removed from record formats. Values of some attributes can be inferred from the values of other attributes via rules stored in a knowledge base (KB).
- **NewsRetriever, v. 1.0**—The ANSER team developed NewsRetriever, which can be trained to return only NewsGroup articles containing pre-specified relevant information. Its design is sufficiently flexible for integration with almost any artificial intelligence filtering component for text categorization, image processing, and face recognition. It also includes a keyword-matching component. A news browser is being integrated with text-categorization agents.
- **Text Categorization**—The ANSER team developed several text-categorization agents using various algorithms including Category Discrimination Method (CDM), vector space, Naïve Bayes, and Document Clustering Agent. The CDM Agent and Naïve Bayes Agent strategies have been integrated with Internet NewsGroup search agents.
- **Disk Hound, v. 1.0**—Disk Hound supports search of a data storage device connected locally or shared over a network for embedded data and various file types. Disk Hound is a computer forensics tool that uncovers information and provides a documented chain of evidence.

Technology Transfer to Support Information Analysis by Law Enforcement

A major goal of the effort has been to disseminate information about this grant effort to local, state, and Federal law enforcement organizations. This information dissemination includes initial plans and expectations for developing potential technology solutions and the resulting demonstrations and, ultimately, delivering and operating prototype systems. The centerpiece of this strategy has been to identify and attract interested law enforcement agencies that could understand the benefit from this ANSER R&D effort and would be able to participate with the ANSER team as pilot project partners. Participants would serve as subject matter experts and test and evaluate systems in real-world operational settings.

Through demonstrations, presentations, and discussions, the ANSER team has attracted and developed pilot project partnerships with three, and potentially four, law enforcement organizations.

- ***West Virginia State Police (West Virginia Missing Children Clearinghouse)*** will focus on West Virginia's missing (abducted and runaway) children. The ANSER team

is integrating prototype software systems for Internet search, monitoring, and mapping; face recognition to match still images from the Internet to a known database; computer forensics analysis on hard drives from confiscated computers; and automated knowledge-based case analysis tools to integrate and fill gaps in consolidated and distributed data records. The West Virginia Missing Children Clearinghouse will pilot test and evaluate three prototype software systems: MCLA, CASA, and Disk Hound.

- ***U.S. Department of the Treasury***—The ANSER team will provide automated prototype software systems for fighting the exploitation of children and locating missing children. Treasury will pilot test and evaluate two prototype software systems: COPIES (including DSRegistrar) and NewsRetriever.
- ***South Florida High Intensity Drug Trafficking Area (HIDTA)***—The ANSER team will broaden face-recognition applications from missing children to adults for near–real time identification of suspects, photographed with video surveillance cameras. The South Florida HIDTA, a Drug Enforcement Administration organization, will pilot test and evaluate two prototype software systems: CBSSA and IdentiFace.
- ***Federal Bureau of Investigation (FBI) Criminal Investigative Division's Innocent Images Program***—Potentially a fourth pilot project partner, the FBI Innocent Images Program will apply automated software to fight the exploitation of children. The program would pilot test and evaluate a prototype of the COPIES software system with DSRegistrar.

For each pilot partnership, specific technology solutions were identified and then developed and integrated as prototype systems for demonstration to its Pilot Partners.

Summary of the Grant

The ANSER team successfully developed solutions that will substantially advance robust and scalable (potentially enterprise-level) investigative capabilities by

- Understanding law-enforcement needs and specific requirements
- Understanding the technology base
- Designing and implementing many individual components, applying face recognition and ISA technologies
- Merging face recognition, Web browsers, text categorization, file recognition, data mining, and knowledge management into integrated prototypes for automated collection, review, and analysis of very large law enforcement information holdings

When NIJ awarded this cooperative agreement to ANSER, there were no technologies with sufficient refinement for searching image databases compiled by law enforcement agencies and organizations for missing children. Related technologies lacked both real-world effectiveness and commercial support. Many of the available facial-recognition methods were being studied in academic laboratories and nearly ready to be transferred and integrated into software for public use. Visionics and Eyematic Interfaces improved

their FaceIt and Eyematic Primary Library technologies through this effort to meet the demanding requirements of fast and accurate searching of large facial databases. They have made these technologies commercially available to developers, including the ANSER team.

Solutions derived from these technologies are expected to be sufficiently robust and effective for law enforcement investigations of open-source information, particularly on the Internet and in other open-source holdings and video surveillance, for almost real-time identification of suspects.

In the past 2 years, the ANSER team has developed ISA and face-recognition components and integrated robust, scalable prototype systems. It has successfully demonstrated these systems to local, state, and Federal law enforcement organizations and sought evaluations of usefulness and readiness in law enforcement applications. Consequently, ANSER has established partnerships with three (and potentially four) law enforcement agencies and has begun to pilot test prototype software systems. These systems either have been or soon are to be delivered, installed, tested, and evaluated within law enforcement organizations to solve real-world law enforcement problems, especially in activities to benefit missing and exploited children and in Drug Enforcement Administration criminal investigations.

Implications for Law Enforcement

Without this NIJ cooperative agreement, law enforcement organizations would not be working with the current prototypes and awaiting others that are being integrated and are expected to be available shortly. Law enforcement and other public agencies are beginning to feel that computerized facial recognition can be an effective technology in searching for and identifying missing children. Facial-recognition technology promises to increase law enforcement effectiveness in tracking runaways and abductees. Regional and national agencies are now routinely digitizing and adding pictures to databases of missing children. ANSER's search and analysis software will soon make it possible to automatically share images of missing children and search for them across open sources such as the Internet and intranets. Computerized facial recognition will provide these agencies with an ability to search more images reliably and rapidly.

In the near future, law enforcement officers will also be able to use ANSER's prototype systems in patrol cars and surveillance vans equipped with digital cameras, laptop computers or digital terminals, and wireless communication equipment. Law enforcement officers soon will be able to capture and upload images of suspects as queries to a central digitized booking system and other related systems. The ANSER-developed system will automatically process the queries and return reports almost immediately.

Many of these current components and capabilities continue to be developed, integrated, and tested under a second NIJ cooperative agreement awarded in August 1998 that runs through September 2000.

1. The NIJ Grant

1.1 Background

The 1996 House Report 104-676 stressed the importance of locating missing and exploited children and suggested funding "for Facial Recognition Technology using aging algorithms."[1] In 1997 Congress recognized that law enforcement staffing was inadequate to locate missing children and to stop the exploitation of children on the Internet. Congress therefore directed the development of promising emerging technologies to solve this problem. It was believed that modern computer technology substantially enhanced investigative efforts for missing and exploited children and had extensive potential for a wide range of other law enforcement problems. In general, evolving computer technologies, such as automated face recognition and ISAs, were at the time quite immature and were known to require enhancement and integration to provide useful solutions.

The NIJ, in late summer 1997, awarded ANSER a 2-year $3.1 million cooperative agreement (97-LB-VX-K025) from August 1997 through July 1999 to research and develop "Technologies for Identifying Missing Children." The initial Statement of Work for the project included four primary tasks:

Task 1 Conduct system requirements, architecture and design analyses, system integration, installation, testing, and training

Task 2 Develop the face-recognition component

Task 3 Develop the intelligent agent component

Task 4 Research the transfer of this technology to aid law enforcement

The first three of these tasks have focused primarily on developing software tools and conducting research into several technologies with promising solutions for law enforcement needs and specific requirements. The fourth task has concentrated on finding ties to the law enforcement community, demonstrating the usefulness and readiness of several prototype solutions from these technologies, and initiating pilot projects with law enforcement organizations.

Based on the NIJ Statement of Work, ANSER formulated a total strategy for conducting R&D in several emerging technologies implementing software components, collecting data, customizing prototype development, and finally transferring the ANSER-developed technology solutions to several law enforcement organizations.

It should be noted that ANSER has been awarded a second NIJ cooperative agreement (98-LB-VX-K0021), entitled "Facial Recognition and Intelligent Software Agents for

[1] LECTAC Prioritized Requirements 25 March 1996—Biometric Facial Mapping.

Law Enforcement Applications," to enhance, extend, and promote many of the efforts described in this final report, including

- Extending efforts in face recognition from still images to video images with an application called IdentiFace
- Expanding the work in evolutionary agent societies (EASs) to more robust and scalable distributed systems and data sources
- Incorporating text-categorization methods (along with NewsRetriever) to support threat assessment
- Conducting technology transfer through pilot project partnerships with several law enforcement agencies

1.2 Goals and Objectives

The two principal goals of this cooperative agreement are

- To develop software systems incorporating technologies on automated facial matching and ISAs to aid law enforcement and other organizations by leveraging scarce human resources for locating missing children and fighting the exploitation of children.
- To demonstrate the usefulness and readiness of several integrated applications, based on new artificial intelligence and computer technologies for local, state, and Federal law enforcement applications.

This effort is aimed at helping NIJ reach its long-range goals "*to improve law enforcement and the criminal justice system* (Goal V) and *to develop new technology for law enforcement and the criminal justice system* (Goal VI)." Furthermore, it directly supports the Law Enforcement and Correction Technology Advisory Council prioritized requirement to "develop a system to image a human face and create a digital map for comparison and identification purposes. The system would compare the face with the name and the stored image map of the face."[2]

This effort initially focused on developing a system to locate missing children and to fight the exploitation of children. The same technologies seemed to also fit other law enforcement investigative requirements. The ANSER team developed several technologies and implemented many prototype systems that will substantially advance investigative capabilities. The solutions are expected to have a high potential for producing force multipliers, enhancing their ability to locate missing children and criminals, freeing law enforcement officers to patrol the streets or follow up on new leads, and in general reducing crime. For example, comparing live surveillance videos or videotapes to databases of digitized mug shots compiled by law enforcement agencies can identify criminals and terrorists.

---

[2] LECTAC Prioritized Requirements 25 March 1996—Biometric Facial Mapping.

R&D has been conducted in several technology areas leading to software tools and components, including

- A face-recognition tool that automates the process of finding and matching images to databases of missing children
- A similar face-recognition tool that automates the process of finding and matching video or images to photographic databases of criminal mug shots
- A photographic image enhancement tool
- Age-progression research
- ISAs to assist case investigators in searching, monitoring, and analyzing information on the Internet and within internal and external databases
- Automated collection, analysis, storage, and retrieval of very large information holdings, applying
  - Text categorization
  - Data mining
  - Case analysis, based on expert systems
- Integration of several prototypes for demonstration

1.3 Evolution of the Grant

ANSER began working with facial recognition and ISA technologies in conjunction with Lawrence Livermore National Laboratory (LLNL), a not-for-profit research and educational institution, and Science Applications International Corporation (SAIC).

- LLNL was to conduct R&D of
  - Face recognition with age progression
  - Computer-generated intelligent agents, such as the LLNL CAPTAIN system, to help Federal, state, and local law enforcement agencies locate missing persons, particularly missing and exploited children
  - System integration
- SAIC was initially responsible for research into how the Internet and other open sources could be searched to support investigations relating to the exploitation of children

ANSER believed that the National Center for Missing and Exploited Children (NCMEC) could be an end user of the new software systems, and ANSER and NCMEC have conducted a series of discussions on potential synergies.

1.4 ANSER Team Participants

The ANSER facility in Fairmont, WV, has been the primary location for the core research, development, and integration.

To leverage and capitalize on existing technologies and technical capabilities as well as to supplement ANSER's expertise and skills, through the course of the effort, ANSER has built a coordinated team of locally and nationally recognized subcontractors.

1.4.1 ANSER as the Prime Contractor

As the prime contractor, ANSER has been responsible for analyzing and managing the requirements and design, conducting R&D, collecting data, integrating multiple software components into system-level applications, and beginning to work with law enforcement and other agencies to transfer technology.

Specific technologies pursued by ANSER include face recognition, facial age progression, Internet search and retrieval, ISAs, and EASs.

The makeup of the ANSER team (including subcontractors) for this first NIJ program is shown in Exhibit 1.

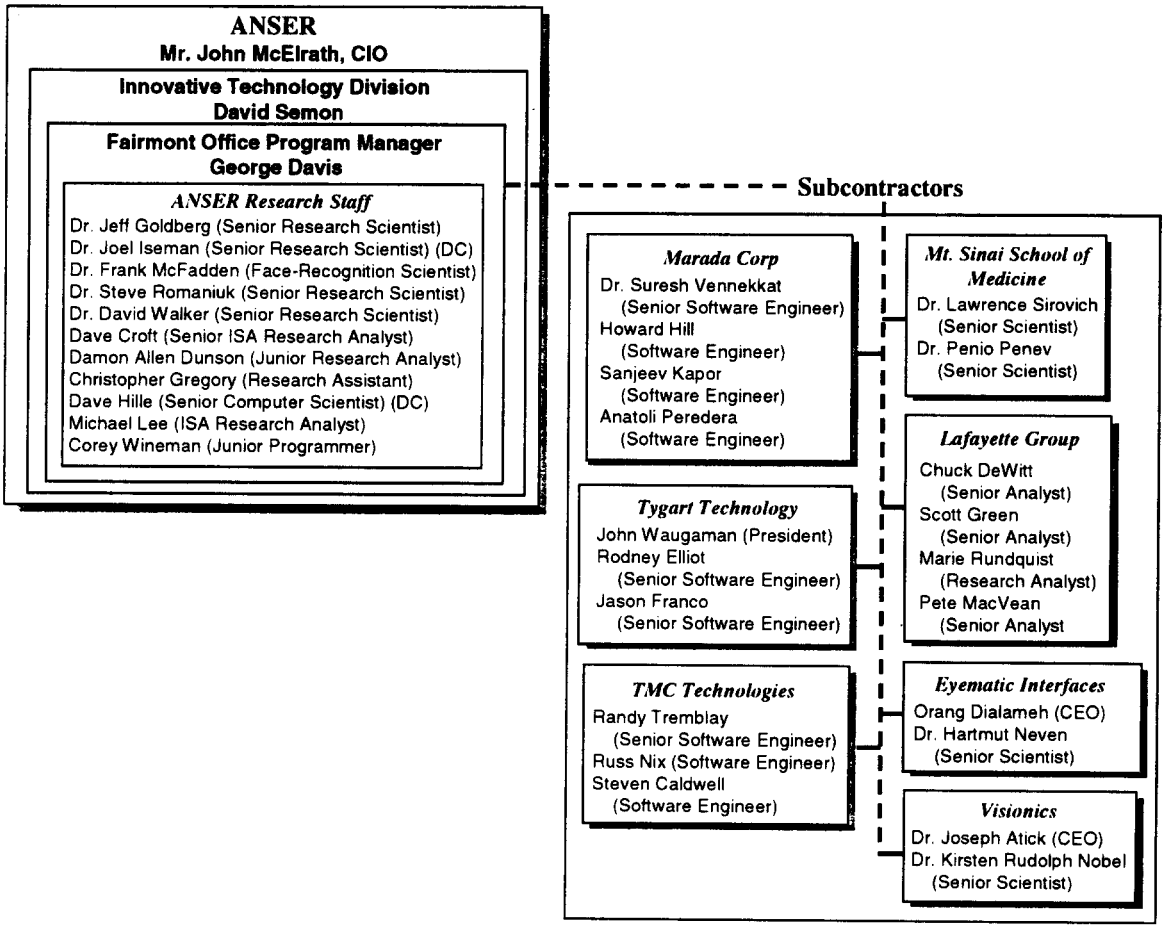


Exhibit 1—The ANSER Team

1.4.2 Visionics and Eyematic Interfaces

Both Visionics and Eyematic Interfaces have developed face-recognition software and marketed commercial face-recognition products.

- Visionics was the primary face-recognition provider because it excelled in technical quality in FERET tests and was selling and delivering commercial products.
- ANSER added Eyematic Interfaces because Eyematic also performed well in the FERET tests. The Eyematic Interfaces methods showed a strong promise for future improvements while complementing the Visionics approaches.

Visionics and Eyematic Interfaces have been conducting R&D to independently improve their own face-recognition capability—especially algorithm performance for images of children and for images in the uncontrolled environments that are typical of law enforcement applications. Because of the availability of Visionics software libraries—packaged as a software development kit (SDK)—the ANSER team has been able to implement face-recognition prototype solutions and to conduct successful demonstrations for law enforcement agencies. Each individual software package was evaluated, as well as an integrated model using a neural net, before the decision was made to use the Visionics FaceIt.

1.4.3 Mount Sinai School of Medicine Laboratory of Applied Mathematics and Rockefeller University

These two research institutions are conducting aging algorithm research that is expected to lead to software that will be incorporated into prototype solutions for finding missing children using age progression and regression.

1.4.4 Lafayette Group Inc.

This company is a group of law enforcement experts, consulting to the government and law enforcement agencies. Lafayette Group has provided support as subject matter experts for law enforcement software requirements, operational environments, constraints on automated software for conducting investigations, and plans for transferring technology to law enforcement agencies.

1.4.5 Marada Corporation, TMC Technologies, Inc., and Tygart Technology Inc.

These three West Virginia high-technology small businesses have provided assistance in the development of ISAs. Marada (now merged with TMC) also has been responsible for the development of image-enhancement and image-processing components. TMC has been conducting research on data-mining approaches. Tygart is involved with the text-categorization tools.

1.4.6 Pilot Project Partnerships

As the effort began to gain more focus, emphasis shifted and alternative end users were sought and selected as pilot project partners. At the conclusion of the initial NIJ contract, the ANSER team either was developing or trying to develop four pilot projects, each with one of a diverse group of law enforcement agencies. These law enforcement organizations have become central to the effort as pilot project partners (discussed in Section 7):

- West Virginia State Police and the West Virginia Missing Children Clearinghouse
- South Florida HIDTA, a Drug Enforcement Administration organization
- U.S. Department of the Treasury
- FBI Innocent Images Program (potentially)

Sections 2 through 7 of this Final Report provide a detailed summary of the major R&D efforts, software component implementation, prototype integration, and nascent pilot projects.

2. NIJ Grant Definition

This NIJ effort was designed to implement several proactive tools for use by investigators to recognize information on the Internet and in other data sources. The ANSER team reviewed available technologies in face recognition, Internet search, ISAs, database management systems, computer forensics, and case analysis.

ANSER's strategy in developing and conducting this grant has been to match law enforcement agency needs to technological potential and to work through six stages, incrementally building on the foundations of previous stages while revisiting them as the situation has warranted. The six stages of the ANSER team effort are

1. Understanding criminal activity on the Internet
2. Understanding law enforcement agency operations
3. Understanding data sources
4. Conducting R&D on software components
5. Implementing prototype systems for demonstration and gathering law enforcement agency feedback
6. Disseminating prototype systems to pilot project partners

Each of these six stages is discussed in greater detail in the rest of Section 2.

2.1 Understanding Criminal Activity on the Internet

To better understand the complex issues of child exploitation and pornography on the Internet, the ANSER team researched the psychology and modus operandi of pedophiles and those who traffic in illegal images of children. The team examined the prevalence of

child pornography sources on the Web and the method of operation for these sites. The team also examined how and why the Web has given rise to such a demand for this type of pornography.

The issue of unlawful sexual contact between adults and children via Internet chat rooms has also been examined. The Web offers a unique realm for a pedophile to make contact with a child and develop a relationship that may lead to abuse. The team examined the current Federal efforts employed to combat this activity and explored other possible ways to prevent this type of contact.

2.2 Understanding Law Enforcement Agency Operations

Initially, the ANSER team sought to identify and understand operational environments and operations of selected Federal, state, and local law enforcement agencies. ANSER selected a few law enforcement agencies that it believed could benefit from automated and scalable capabilities provided by the new biometrics and information recognition systems. The Internet has become an overwhelming challenge for law enforcement officials because it forms a vast, undisciplined source of information—good and bad. Retrieving overwhelmingly large amounts of irrelevant information frustrates many analysts who are searching for specific information. Law enforcement agencies, however, have neither the tools nor the time to search this vast and rapidly expanding data source for the important information and clues hidden in it. Child pornography circulates freely on the Internet. Law enforcement investigators search for specific Internet sites where they expect to find faces of missing and exploited children. However, their methods for monitoring the Internet are people-intensive and cannot keep pace with the rapid expansion and use of the Internet as a haven for criminal activities. Actually, static, indexed Internet data no longer provide significant help. The following technologies have potential to fit a variety of law enforcement agency needs and seemed to have widespread interest:

- Matching faces of children found in photographs on the Internet to gallery databases of missing children
- Automated age progression of children's photographs
- Matching images to known examples of child pornography
- Conducting case analysis
- Conducting computer forensic analysis on seized computers

Each of these topics is discussed briefly in the following sections from a law enforcement perspective.

2.2.1 Matching Faces in Photographs of Missing Children to Gallery Database Images

Photographs are extremely important in cases involving missing children. However, law enforcement agencies have not been equipped with the tools to compare photographic

images quickly, conveniently, and effectively against databases (photo galleries) of missing children. As the trail of clues for unsolved cases grows cold, law enforcement officers generally can react only after they are notified by people who recognize missing children's pictures that are distributed in mailers, on milk cartons, in flyers, or at kiosks. This reactive approach, although sometimes effective, requires people to see the distributed picture or photograph, to match it to a child's face, to contact the police or an organization for missing children, and to pass along important information. When this method does produce information, the task of manually comparing leads and photographs from outside sources with internal case records and original pictures of the missing children is time consuming and labor-intensive for investigators.

2.2.2 Automated Age Progression of Children's Photographs

Law enforcement organizations are successfully using highly skilled experts to manually draw or create computer-assisted, composite age-progressed sketches. Sometimes an artist's rendition is not made until many years after the child was last seen. The manual procedure requires extensive research into the background of the missing person—including old photographs of the child and pictures of parents and other relatives at similar ages—and knowledge of physiology. Age progression is required to keep databases of missing children up to date. ANSER's research will potentially provide a basis for automating this procedure.

2.2.3 Matching Images to Known Examples of Child Pornography

With the explosion of Internet use, immense amounts of open-source data containing information about almost any topic are available online. Child pornographers have taken advantage of the anonymity that the Internet provides in order to trade and sell child pornography. The primary problem law enforcement officers face is to weed through vast amounts of extraneous data to get to postings containing or related to child pornography.

Since 1995, the FBI's Innocent Images Program has been investigating child pornography on the Internet and the sexual exploitation of children. Part of the program focuses on investigating individuals who produce, distribute, or post child pornography via the Internet and online services. FBI agents and other Federal, state, and local investigators go online undercover, posing either as young children or as sexual predators, to identify individuals who are victimizing children. These investigations are labor-intensive and can cover only a very small percentage of the ever-growing Internet. ANSER is developing capabilities in the areas of text categorization and digital signature matching. These capabilities hold promise in aiding law enforcement investigations.

2.2.4 Conducting Case Analysis

Investigators analyze a large number of leads and tips about possible new information as well as existing case records and case histories. Diligent analysis requires a great deal of time in an environment where the availability of personnel is severely constrained. Law

enforcement agencies must allocate investigative resources to close cases effectively and efficiently. The limited number of law enforcement investigators makes it impossible to keep track of and follow more than a few associations or patterns among leads and cases of missing children. ANSER's knowledge-based artificial intelligence software will greatly assist these individuals.

2.2.5 Conducting Computer Forensic Analysis on Seized Computers

In ever increasing numbers, law enforcement investigators have been collecting personal computers and electronic media—either confiscated or submitted as potential evidence. Investigators have used generic commercial off-the-shelf products to develop a chain of evidence to recover and analyze electronic media such as hard drives, floppy disks, and CDs as part of an evidence chain.

Automated and potentially customized support in the form of computer forensics tools (see Section 3.3) could provide improved capabilities.

2.2.6 Learning From Law Enforcement Subject Matter Experts

As the ANSER team began working on the requirements definition, it recognized the need to

- Involve individual subject matter experts from representative law enforcement organizations in developing functional requirements for reusable software components and systems
- Attend seminars and conferences on law enforcement activities and processes

Specifically, the ANSER team conducted interviews with law enforcement domain experts (case managers) from various law enforcement agencies and were advised of the development of several types of software, focused on searching the Internet (the World Wide Web, NewsGroups, and eventually Internet Relay Channel chat rooms) for instances of known child exploitation. Work was begun on developing a concept of operations for each pilot site.

2.3 Understanding Data Sources

The ANSER team has been developing methods for finding, collecting, and analyzing open-source data resources that could successfully be applied (1) as background material for designing and implementing tailored algorithms, components, and prototypes for face, file, text, and data recognition or (2) as pointers to data resources for use in the prototypes, pilot demonstrations, and tests.

2.3.1 Required Data Resources

The ANSER team has worked with the following types of data resources:

- Collections of digital pictures of missing children (and their abductors)
  - From databases of scanned photographs and drawings provided by state clearinghouses
  - From downloaded state and private World Wide Web sites containing information and pictures of missing children
- Collections of images of known child pornography
  - Encoded images (digital signatures) of child pornography
- Galleries of mug shots from law enforcement bookings
  - From law enforcement booking databases of scanned photographs and digital images from organizations within the South Florida HIDTA
- Digital surveillance videotapes
  - From law enforcement digital video surveillance tapes provided by the Metropolitan Dade Medical Examiner Department

2.3.2 Legal Issues of Privacy and Freedom of Speech Affecting Searches of Internet Data Sources

The ANSER team researched the legality of the search engines for viewing and extracting information from various data sources. Current laws governing the Internet, along with general privacy and freedom-of-speech laws, have been examined to gain an understanding of how the Internet is governed. Because legislation pertaining to the Web is still emerging, recent *Law Review* articles and new congressional legislation have been examined to gain a sense of the direction of future legislation. U.S. Department of Justice staff members working on Internet privacy and security issues have provided further insight into laws and privacy boundaries on the Internet.

ANSER search engines are set to obey all legal restrictions and will not violate password protection and robot exclusions.

2.4 Conducting R&D on Software Components

Many law enforcement organizations have indicated to ANSER that there is a strong match between their requirements for automation and ANSER's new technologies and R&D initiatives in facial biometrics, Internet search, data mining, text categorization, knowledge management, and evolving ISAs.

Therefore, one of the principal aspects of this first NIJ grant has been the pursuit of R&D methods in several emerging technologies capable of developing enhanced, integrated, automated, prototypical, efficient, and effective tools and systems to assist the law enforcement community. Based on ANSER's initial observations and discussions with

law enforcement organizations and careful appraisals of available technologies, the R&D has focused on implementing software solutions to

- Recognize Internet sites with information leading to solving cases of missing and exploited children
- Recognize individual data and documents that support building knowledge for case analysis
- Support evidence chains with which to obtain warrants and to apprehend criminals

The ANSER team has focused on areas that include data search, content recognition, and ISA development and ISA management, as follows:

- Data search agents (Section 3) for use on the Internet and other open sources
  - Internet search agents for automating data monitoring, mapping, and collection on the Internet (Section 3.1)
  - Database search agents (Section 3.2)
  - Computer disk search (Section 3.3)
- Recognition agents (Section 4)
  - Automated biometrics face recognition (Section 4.1)
  - Digital signature agents to match pairs of image files (Section 4.2)
  - Text-categorization agents to match documents to pre-labeled exemplars (Section 4.3)
  - ISAs for data mining, inferencing, and knowledge management (Section 4.4)
- Agent infrastructure (Section 5) for developing and managing ISAs
  - Agent development environment—using the Symbolic Agent Development and Integration Environment (SADIE) (Section 5.1)
  - Agent monitoring and control environment—EAS (Section 5.2)

2.5 Implementing Prototype Systems for Demonstration and Law Enforcement Agency Feedback

The ANSER team developed both operational and functional requirements for prototypical systems in several law enforcement areas. Based on insights gained in discussions with law enforcement subject matter experts, the ANSER team is designing and developing ten demonstration prototypes:

- MCLA, v. 1.0
- CBSSA
- IdentiFace
- COPIES, v. 1.0
- DSRegistrar

- Internet Investigator
- CASA
- NewsRetriever
- Text Categorizer
- Disk Hound

The ANSER Team has demonstrated components and/or integrated prototypes of each of these systems to law enforcement agencies. Feedback from the law enforcement community has indicated that design concepts and implementations are generally sound. Many suggestions are being integrated into the software components and graphical user interfaces (GUIs).

2.6 Disseminating Prototype Systems to Pilot Project Partners

The culmination of this strategy has been to attract and sign up interested law enforcement agencies as pilot project partners and users for testing and evaluating prototype systems in real-world operational settings and providing substantive user feedback. To find interested partners, ANSER's efforts included

- Generating prospects and attracting future law enforcement participation and collaboration in pilot initiatives for ANSER's prototype systems in real-world environments by demonstrating prototype systems
- Implementing more complete pilot systems from suites of agents in support of program goals
- Installing a few pilot prototype systems for law enforcement use, test, and evaluation
- Reviewing substantive law enforcement feedback and evaluations about pilot prototype systems

ANSER is working in pilot project partnerships with the West Virginia State Police, the South Florida HIDTA, and the U.S. Department of the Treasury.

3. Data Search and Monitoring

The ANSER team surveyed the law enforcement community on the feasibility of tailoring automated software for searching, locating, and analyzing the Internet and other open sources (databases and computer hard drives). The surveys identified several automated approaches (continuous, periodic, or as directed by the user) for searching and locating sites that may contain useful information (images, text, or structured data) for further analysis. As a result, several tools and utilities were studied, designed, implemented, and integrated to facilitate Internet-based investigations, data mining, and computer forensic analyses of hard drives. ISAs can both automate and lend support to manual processes being conducted by investigators and case managers. These intelligent

software systems are expected to improve efficiency and effectiveness in law enforcement agencies' investigative methods.

3.1 Internet Search Agents

Most search and discovery on the Internet today is nonspecific, blind, and reliant on simple spider agents. The results of these searches populate large-index databases in which search engines seek keyword matches. The time required to spider the Internet is increasing along with Internet size. In 1998, the best coverage of the World Wide Web by any search engine was about 35 percent.[3] A year later, the best coverage stood at only 16 percent.[4] Many well-known directories provide only about 2 percent to 4 percent coverage. The Web has about 800 million static pages as well as billions of dynamic pages that are not covered by search engines.

The ANSER team reviewed, analyzed, and recommended a few available investigative support software products, including

- Primary analytical tools used by law enforcement
- Automated search engines and information retrieval and case-analysis tools in use by the law enforcement community in criminal investigations
- Commercial off-the-shelf Internet meta-search tools and technologies to support Internet image search and monitoring
- New technologies and methods for integrating the ANSER team's solutions for law enforcement

The ANSER team designed and implemented reusable Internet search agents with multiple missions of traversing the Internet and searching for images and other information. Search agents are being designed and developed to form the foundation as core components on which larger systems are being integrated. Four specific Internet protocols are Hypertext Transfer Protocol, Network News Transfer Protocol (NNTP), File Transfer Protocol (FTP), and Internet Relay Chat. The search agents

- Perform multiple missions, including those potentially relevant to cases of missing and exploited children
- Operate in three modes:
  - Executed as autonomous continuous background processes
  - Executed by the end user
  - Used interactively with other agents to control the order in which Internet sites are visited

---

[3] S. Lawrence and L. Giles, "Searching the World Wide Web," *Science*, Vol. 280, April 1998.

[4] S. Lawrence and L. Giles, "Accessibility of Information on the Web," *Nature*, Vol. 400, No. 6740, July 1999, pp. 107–109.

The following search agents have been developed under this first NIJ grant:

- **BloodHound, v. 1.0a1**—BloodHound is a reusable software search agent, a Web crawler or spider that can autonomously and interactively collect data from the World Wide Web. BloodHound is designed as a modular component to be integrated into a variety of higher-level applications to pass data to other intelligent agents for processing. BloodHound is used in prototype demonstration applications such as the MCLA, COPIES, SiteRetriever, Site Mapper, and SiteMonitor (see Section 6).
- **NewsHound or NNTP application program interface, v. 1.0**—NewsHound is a reusable software agent that autonomously and interactively collects data from NewsGroups and passes information to other intelligent agents for processing. It conforms to the NNTP standard (RFC 977) and contains useful extensions that simplify the development of NNTP-based applications. Higher-level end systems, such as the NewsFilter Agent, integrate this application program interface as a reusable component (a follow-on to NewsRetriever).
- **FileBrowser and FileHound, v. 1.0a1**—FileBrowser and FileHound are software search agents that search for information on FTP sites. FileBrowser is interactive; FileHound is not. The FTP application program interface implements the FTP standard (RFC 959).
- **ChatHound v.1.0a1**—ChatHound is a software agent that searches Internet Relay Chat rooms to find and retrieve specific types of sites, files, or information. The material retrieved can then be recognized by one of the recognition components. Face recognition, image file recognition, and text categorization would be used in most cases.

3.2 Database Search Agents

The ANSER team developed several types of ISA tools for conducting automated data discovery and case analysis for information in local databases. These ISAs automatically isolate the more important and nontrivial while extracting the apparently more valuable information for investigators to pursue.

Data-Mining Agents—Knowledge Discovery and Predictive Data Mining, v. 1.0 alpha: With the explosion in electronic data sources, law enforcement investigators face the problem of interactively searching and sorting through large numbers of very large and distributed databases to find information or patterns relevant to a case. The ANSER team developed autonomous data-mining agents to make this process more feasible, efficient, and effective. Data mining (also called "knowledge discovery in databases") is the process of automatically extracting valid, useful, previously unknown, and ultimately comprehensible information from large databases to support crucial decision making. Data mining is more valuable when human intervention is minimized.

The ANSER team designed fully autonomous data-mining agents capable of discovering data sources and their inherent relationships. Specifically, a system has been designed with three types of data-mining agents that cooperate and coordinate to recognize data:

- Surveyor agents to uncover or find potential data sources (databases) of interest
- Prospector agents to perform knowledge discovery and to preformat data of interest
- Miner agents to further process the data

Surveyor agents are discussed next, and prospector and miner agents are discussed in Section 4.

Surveyor Agents—In an initial implementation, surveyor agents locate all system databases that are accessible via open database connections. Surveyor agents identify data sources—both structured (for example, databases, KBs, and Extended Markup Language–enabled pages) and unstructured (for example, text documents and Hypertext Markup Language [HTML] pages). Relevant resources uncovered by the surveyor agents are forwarded to the prospector agents for further processing. In the future, enhanced surveyor agents will be capable of traversing both intranets and the Internet.

3.3 Computer Disk Search—Disk Hound as a Computer Forensics Tool

Disk Hound will search a data storage device (such as a hard drive) connected locally or shared over a network. Disk Hound will be capable of searching for both specific file types and embedded data. Disk Hound's software agents will traverse targeted storage devices and locate instances of specific data. The user can perform complex searches via a user-friendly GUI (see Exhibit 2). This application (developed using Java/Swing) assures multiplatform portability.

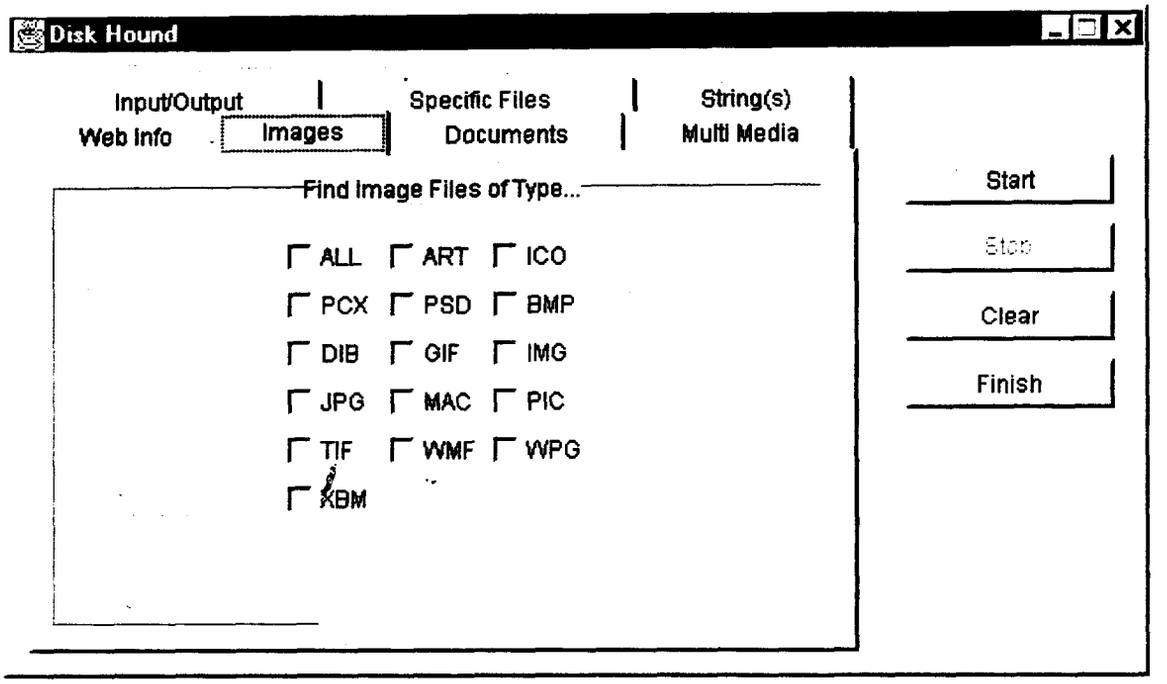


Exhibit 2—Disk Hound Has a Simple and Easy-to-Use Interface

4. Recognition Engines, Information Understanding, and Decision Support

Once potential Internet sites and data files have been selected either by data search engines or through human intervention, various recognition, understanding, and decision support tools can be used. This part of the grant effort has concentrated on four types of recognition components:

- Biometrics (facial) recognition
- File recognition
- Text recognition
- Data recognition

Each of these recognition systems is discussed in detail through the rest of Section 4.

4.1 Biometrics (Facial) Recognition

FERET has been the standard for face recognition, spurring great strides in recent years. However, improvements are still being made to perfect this technology. The ANSER team has been able to use software libraries from two of the best facial-recognition products on the market. While Visionics and Eyematic Interfaces commercial SDKs have been valuable tools for developing systems, the ANSER team has had very limited visibility into the inner workings of either Visionics or Eyematic Interfaces algorithms.

These companies protect their intellectual property and have not been willing to divulge their trade secrets (including the specific improvements to their software libraries).

The ANSER team conducted the following R&D activities to build components for automated face recognition for both still and video images:

- Applying face-recognition engines based on software libraries from more than one vendor to develop software that will automatically match probe images to galleries of known images of missing children or criminals
- Incorporating image enhancement methods from both ANSER team members and commercial vendors
- Combining face-recognition methods
- Comparing face-recognition methods
- Conducting research in extending the current state of the art, including
  - Age-progression or -regression algorithms and tools
  - Face profile recognition
  - Integrated face-recognition systems

Detailed descriptions of these face-recognition subtasks are provided in the areas of

- Enhancing commercial face-recognition engines
- Collecting face-recognition data
- ANSER team face-recognition components
- Combining face-recognition analyses
- Comparing face-recognition methods
- Enhancing face-recognition system performance
- Integrating face-recognition technologies into a usable tool with a common GUI

4.1.1 Enhancing Commercial Face-Recognition Engines

Visionics and Eyematic have developed and marketed commercial face-recognition products. Both companies participated as subcontractors on the ANSER team grant. Over the past 5 years, both have developed and improved face-recognition tools. Their efforts under this grant were twofold—improving and modifying Eigenface and Local Feature Analysis approaches and packaging C/C++ software library functions as SDKs. It should be noted that many other providers of face-recognition algorithms are working with similar techniques.

Once the subcontract between ANSER and Visionics was signed, Visionics made its commercial SDK available so that ANSER could apply it in developing an initial face-recognition prototype.

Throughout the grant effort, upgrades to the Visionics SDKs and Eyematic's Primary Library were provided to ANSER:

- Visionics provided ANSER with an upgraded SDK in August 1998
- Eyematic Interfaces worked on developing its SDK for vision-based computing and delivered a beta version to ANSER in November 1998
- Eyematic's Primary Library Version 1.0 was delivered in March 1999
- Eyematic Interfaces prepared a final upgrade, delivered in July 1999

Existing face-recognition engines are working well with ANSER's gallery databases, consisting of images collected under semicontrolled environments.

Some gallery databases of digital mug shots made available to the ANSER team are of high quality. However, recognition performance is usually reduced greatly by each variation from the ideal of the face in the image. Specific technical issues for face recognition include face resolution (size), pose (whether the face is directly toward the front to provide a straight-on shot of the face), occlusions, lighting, shading, and camera exposure.

A frontal pose (face toward the front for a straight-on view) is best for ANSER's current systems. Pose variations greater than 15 to 20 degrees from a frontal view cause recognition performance to drop significantly. For children, typically intervals of several years occur between test probe and gallery images. To successfully identify people—both children and adults—enhanced performance of face-recognition engines is necessary.

4.1.2 Collecting Face-Recognition Data

The ANSER team collected facial images—photographs of children and of adult criminals. In each case they sought two types of images: (1) many prototypical facial images from which to build the methodology and (2) the gallery databases of known people who might indeed be matched to the probe image.

4.1.2.1 Facial Data From Existing Sources

The ANSER team collected its current facial database from sources of data consisting of digital or scanned-in images on the Internet, on CDs, or on Zip disks. The sources include

- The NCMEC Web page
- West Virginia (and several adjacent states') Missing Children Clearinghouse Web pages
- Mug shots from
  - South Florida law enforcement booking systems
  - The U.S. Marshal Service Prisoner Tracking System
  - The U.S. Marshal Service Warrant Information Network system

4.1.2.2 New Image Collection Sources From Yearbooks and Videotaped Images

The ANSER team also acquired or photographed new image data, including

- The Mount Sinai data collection
- Faces videotaped in 17 Marion County, WV, elementary and middle schools
- Longitudinal sequences of still facial image data, scanned by the ANSER team from more than 80 Marion County elementary, middle, and high school yearbooks from 1988 to 1999

4.1.3 ANSER Team Face-Recognition Components

The ANSER team's biometrics face-recognition components have applications for both still and video images. A highly automated sequence of operations that are automatically executed and mainly transparent to the end user finds one face at a time in a visual scene. Enhancing and normalizing image background and face variations such as size, lighting, expression, and pose (rotation from frontal), these operations

- Calculate a probe template describing the face
- Rapidly compare the probe template to a large gallery database consisting of templates of known face images
- Report one or more of the best matches between the probe and gallery images
- Permit an end user to review, validate, and interact with the automatic face-recognition reports and decide whether a small subset of the best gallery matches actually contains a successful match to the probe image

4.1.4 Combining Face-Recognition Analyses

The ANSER team attempted to improve recognition rates by developing a neural network to combine (average) the rankings from the Visionics and Eyematic Interfaces face-recognition results. The results indicated that even though the integrated model scored high in some categories, Visionics scored highest overall.

4.1.5 Comparing Face-Recognition Methods

The ANSER team tested the various facial-recognition approaches: Visionics, Eyematic Interfaces, and a combined model that averages the rankings of the two commercial tools. The test benchmarked the performance by comparing probe images to a group of gallery images. Facial images from a total of 800 video sequences of students in various Marion County, WV, grade schools were used in this test. For each student, five still images were obtained with different poses, rotation angles, and facial expressions. One of the five stills was used to create a gallery of images, and the remaining four stills were used as probes.

Test results in the table (Exhibit 3) are the probability that a correct match is in the desired subset (the top 1, 5, 10, 20, etc.) for Visionics, Eyematic Interfaces, and the combined model. For example, using the Visionics software, there is a 58.83 percent

probability that a correct gallery match is in the top 10 scores with this particular gallery size (800 students). This table shows that Visionics outperformed Eyematic Interfaces. The combined model performed better than all models except the "top match" category against Visionics.

Currently, only single-vendor methods are being incorporated into ANSER prototype and demonstration systems. In the future, it is expected that ANSER will leverage the different techniques to create a face-recognition capability that is superior to using either vendor method in its standalone mode.

	Visionics	Eyematic Interfaces	Combined
Top Match	44.76%	41.64%	43.13%
In Top 5 Matches	53.67%	49.67%	53.96%
In Top 10 Matches	58.83%	53.54%	62.22%
In Top 20 Matches	63.86%	57.26%	69.70%
In Top 25 Matches	66.30%	58.68%	71.66%
In Top 30 Matches	68.05%	59.84%	73.18%
In Top 50 Matches	72.99%	63.31%	77.55%

Exhibit 3—Gallery Comparison Results

4.1.6 Enhancing Face-Recognition System Performance

Throughout this 2-year NIJ grant effort, ANSER, Visionics, and Eyematic Interfaces conducted research to improve performance, including recognition accuracy and speed.

Methods were developed, benchmarked, and validated to

- Improve automated face recognition of a single face in an image
- Improve single and multiple face finding
- Use large gallery databases
- Improve robustness to variation in age with varied lighting, poses, and facial expressions
- Improve robustness when working with images of children

4.1.6.1 Vendor Enhancements

Vendor enhancements included automation in preprocessing images to reduce face image variations created by the environment in which they were captured. The enhancements would

- Standardize and normalize images
- Simplify the complexities of automated processing for size and pose—including scaling, aligning (rotating and translating), and histogram equalization for lighting and expression
- Tune the facial identification template, specifically in recognizing children's faces
- Convert among GIF, JPG, TIF, and other file formats
- Improve search speed
- Improve head-outline detector, image quality evaluation, superalignment, and score normalization modules

4.1.6.1.1 Optimizing the Template to Search for Children's Faces

Visionics conducted R&D in a specialized children's face-recognition domain and created new face-recognition algorithms and new FaceIt facial identification templates. Included were cases where the child had aged significantly between the times at which the gallery database photo and the probe photo were taken. Work here included developing and testing new methods for improved invariant representation of faces with respect to lighting, pose, and facial expression. Tests conducted on sample images and a database of children created by Visionics validated the ANSER team's continuing progress in this domain.

4.1.6.1.2 Improving Face Alignment

Because accurate face alignment is critical for good matching performance, Visionics created two software tools that improve face alignment based on the eye location (found both automatically and manually).

4.1.6.1.3 Improving Search Speed

Rapid search speeds, particularly in large databases, are necessary for real-world computerized facial recognition searches to be successful. Visionics researched and developed methods for shrinking the size of the facial identification template from 3.6 kb to fewer than 90 encoded bytes. The facial identification template is stored in the database and retrieved for matching. In addition, Visionics retooled various functions to encode and decode this new facial identification template.

4.1.6.1.4 Improving Head-Outline Detector, Image Quality Evaluation, Superalignment, and Score Normalization Modules

Head-Outline Detector Module. This module is useful for input images with a uniform background and only one head and provides simplified face recognition with increased

accuracy and speed. The algorithm in this module computes a rectangle to enclose just the face in the image. Face recognition is simplified so that continuing processing can concentrate on only the facial image in the rectangle.

Image Quality Evaluation Module. Newer versions of Visionics FaceIt give the user an opportunity to evaluate whether processing the given image is expected to provide adequate results on the basis of brightness, darkness, spatial resolution, and the two-dimensional projection of the subject's head shape. If the user judges that the image quality can provide adequate results, the image can be added to the gallery database.

Superalignment Module. This module provides increased accuracy by better aligning the image to the pupils of the eyes in high-resolution images.

Score Normalization Module. This module allows a user to input the original matching ("confidence") score and calculate a normalized or scaled score to take into account within-person versus across-person sources of facial variability.

4.1.6.2 ANSER Development Team Face-Recognition Enhancements

The ANSER development team has worked in three types of enhancements—image enhancement, facial aging, and profile (ear) recognition, which are discussed in the next three sections.

4.1.6.2.1 Image Enhancements

The ANSER team developed two post-processing routines to facilitate human visual comparison of images. Both a Gamma Corrector and an Image Standardization Module have been developed and tested. The Gamma Corrector adjusts the light intensity of images to compensate for poor lighting conditions. The Image Standardization Module converts the size and intensity of the probe image to match the gallery image. A standard histogram further enhances the quality of the displayed images. These processes were embedded both in a standalone application as reusable C++ DLLs and as Java Native Interface wrappers. The Java Native Interface wrappers have been integrated into other Java-based applications (such as face recognition in MCLA). The screen shots in Exhibit 4 show improved visual comparison of two images as a result of applying the Image Standardization Module to MCLA.

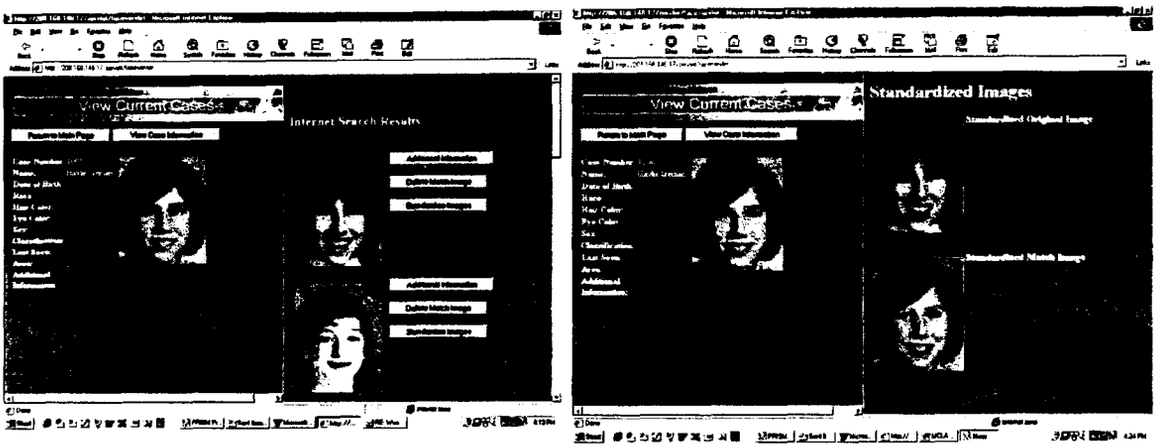


Exhibit 4—Visual Comparison Before and After Image Standardization

4.1.6.2.2 Face-Aging Algorithms for Age Progression and Regression

At the beginning of this project, facial-recognition algorithms were based on models that had been developed for searching adult faces, rather than those of the less angular and smaller-scaled children's faces. The baseline for face aging initially resided in manually drawing and aging faces according to rules known to the artist. Moreover, there was no notion of developing a computerized algorithm capable of taking into account the changes in facial geometry that occur over time because of physical growth. No effort had been undertaken to collect the data necessary to support age-progression research.

Part of this NIJ grant effort has been to develop age-progression methods to convert from a manual to an automated suite of age-progression algorithms.

The ANSER team included the Mount Sinai School of Medicine Laboratory of Applied Mathematics and Rockefeller University. Mount Sinai conducted R&D on automated mathematical age-progression algorithms, including an interpolation method, two Eigenface approaches, and a potentially far more powerful technique called Supersnapshots.

In addition, Mount Sinai developed chronological databases to support their analytic studies. This database initially was populated with 600 images from IDNET, a commercial provider of children's identification cards. A few images were also downloaded from the Internet. These two sources provided children's images, represented by both sexes and several ethnicities with sequences extending over no more than 2 years. Mount Sinai normalized these images based on distance between eyes, lighting conditions, and a cropping template. Testing has not begun with Mount Sinai's algorithms or databases.

4.1.6.3 Profile and Ear Recognition

ANSER also conducted R&D of face profile-recognition algorithms to potentially improve the robustness of a biometrics identification system when only profile images are available. Analysis of profile views, with emphasis on the image of the ear, is potentially useful for identification, although profile views are less frequently

encountered in images. Development (including code conversion) and testing of ear-analysis software has been completed, including development and testing of a neural network model.

No test results are yet available.

4.1.7 Face-Recognition Applications

The emergence of the Internet as a data source of unprecedented magnitude creates an opportunity to proactively search for information related to missing and exploited children. ANSER's automated face-recognition tools seem to have tremendous capabilities for automating an unmanageable manual process. In July 1998, ANSER's facial recognition components were integrated into an initial demonstration system (Section 6.1) that proved to be a precursor to the MCLA.

The ANSER team is developing prototypes that have applications in three pilot systems for face recognition from

- Searching the Web
- Submitting photographs (missing children or criminal mug shots)
- Video (real-time and tape) surveillance

4.2 File Recognition

A second recognition area is file recognition, in which information can be determined about a digital data file, including

- Whether a file (or a portion of a file) matches another digital file
- Whether the content of a file or website has changed since it was last checked
- Links, both internal and external, to the website
- The name, address, and telephone number of the organization or person registered as the owner of the Internet domain name
- The geographic location of the Internet server

4.2.1 Digital Signatures

The ANSER team implemented a National Institute of Standards and Technology Secure Hashing Algorithm–1 (SHA-1) to generate unique 160-bit digital signatures for each image file. SHA-1 is a one-way encryption technique. The resulting encrypted signature is also highly likely to be unique for every image. Since SHA-1 is a one-way, nonlinear technique, an actual file of a child pornographic image cannot be recreated from the 160-bit digital signature string.

This file encryption process is built into the DSRegistrar as part of the COPIES demonstration prototype discussed in Section 6.4.

It is noteworthy that SHA-1 is used to build and match image files with child pornography; it is equally valuable for encrypting and comparing any collection of files. For example, SHA-1 could be applied in checking for and fighting copyright piracy.

4.2.2 Internet Monitoring and Mapping Agents

The ANSER team developed several tools to monitor, map, and download websites that are suspected to contain specific topical information relevant to child pornographic materials or missing children's cases.

- **Site Monitor**—This servlet autonomously polls websites to monitor whether the contents have changed. If a change has occurred, the user can be notified by e-mail with a copy of the old Universal Resource Locator (URL) content for visual comparison. For efficiency, the polling period is adaptive.
- **Site Mapper**—This servlet uses the BloodHound Web crawler to graphically map a website. It reports results to the user as a dynamically generated HTML page. The website will potentially contain a website mapping and analysis tool (see Exhibit 5), two tools (DNS Lookup and WhoIs) for investigating who the registered owner of a website is, and a utility (Finger) for investigating who owns a particular e-mail address. Site Mapper makes an analysis and produces a report, which can be reviewed in an easy-to-use format that reveals hyperlink hierarchy, broken links, unauthorized (password-protected) links, domain names, and e-mail addresses. Site Mapper recursively follows and lists every link at a website so that the user can methodically document and review all of them. The hierarchical structure of the report is designed to make reviewing a site easier than with a standard browser. Site Mapper has been integrated into the Internet Investigator.

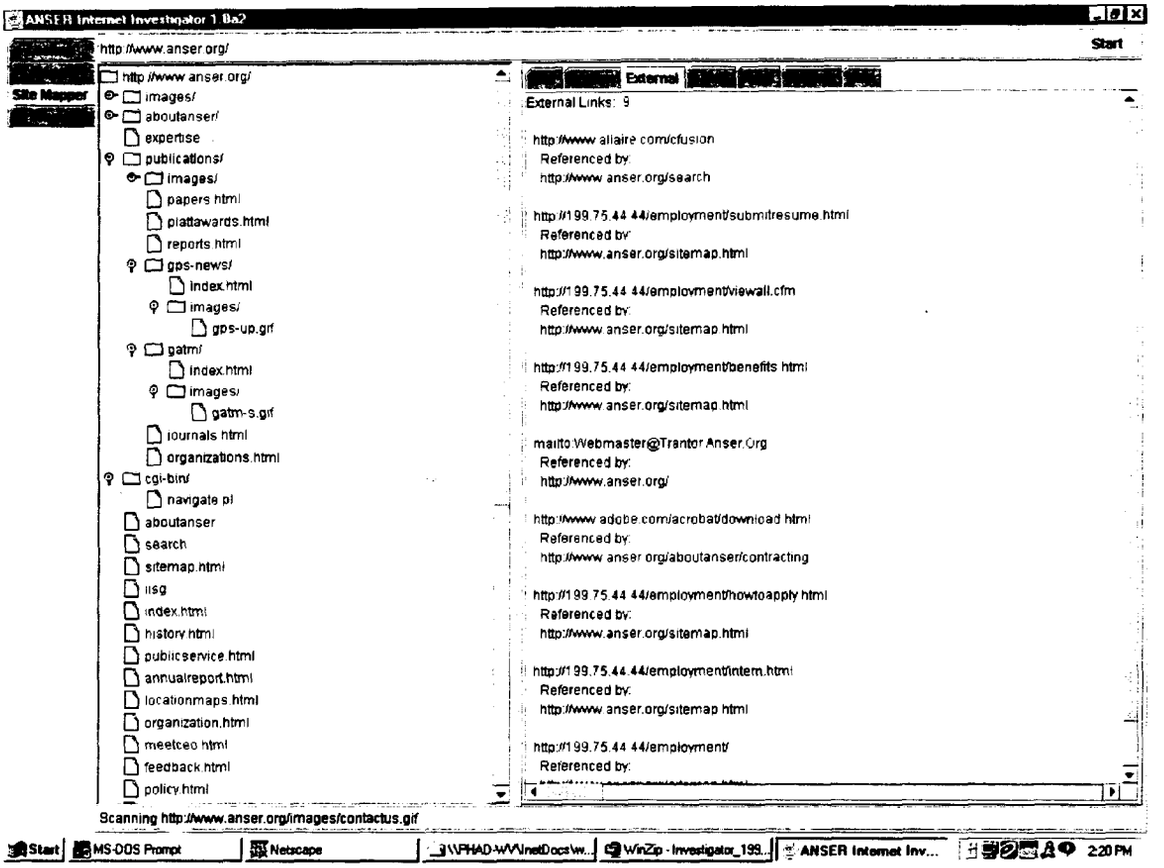


Exhibit 5—Site Mapper Creates a Hierarchical Structure of a Website and Reports Other Site Information Such as Links and E-Mail Addresses

- **SiteRetriever**—This servlet provides the capability to download a complete or partial copy of a website to a local hard disk. This tool enables law enforcement personnel to collect and archive websites for evidence—especially sites that typically are transient.
- **WhoIs**—Internet utilities such as http://www.whois.com/ provide information about the name, address, and telephone number of the organization or person registered as the owner of the domain name on which the Web page is hosted.

4.3 Text Recognition

A third recognition area is automatic text recognition, in which the content of a sample of text can be categorized as matching or not matching text categories of known text entities (that is, documents).

Automatic text categorization applies machine-learning techniques to text-based documents. A generic text categorizer determines the topic or the disposition of those documents according to their content alone. This method can be used in various applications, including

- Organizing textual entities into categories of direct interest to law enforcement users—for example, in this grant effort, classifying whether postings to child pornographic Usenet NewsGroups pertain to individuals trading child pornographic material.
- Supporting Internet search agents by classifying documents into categories indicating their relevance for further processing by various content-specific agents. For example, Web pages can be routed on the basis of textual content for further recognition processing by other ISAs. In this grant effort the ISAs may perform data-mining or facial recognition to filter out irrelevant Web pages.

The ANSER team has conducted research on several text-categorization agents (text categorizers) with various strategies (including CDM, Vector Space Agent, Bayesian Agent, and Document Clustering Agent). The ANSER team then implemented and integrated the strategies with Internet search agents. Certain algorithms have been determined to perform better than others for various categories. Overall system performance will improve if several algorithms and strategies are available.

4.3.1 Category Discrimination Method Agent

The ANSER team enhanced and ported the full CDM algorithm from a very early alpha version running in UNIX to the Windows NT environment. The algorithm was also substantially refined in both operation and performance.

A second version of the CDM—the Incremental CDM (iCDM)—has been designed. Unlike the full CDM, which requires hundreds of examples of each pre-labeled category to learn a categorizer, the iCDM can start to learn from a single example. The iCDM may be used either as an alternative to the full CDM or to refine categorizers previously learned by the CDM. The iCDM algorithm can adjust the evolving features of a categorizer to account for changes in the labeled categories due to concept drift or changes in lingo. A prototype iCDM algorithm was manually tested, but an automated version has not yet been implemented.

4.3.2 Vector Space Agents

The ANSER team developed two text-categorization agents to implement well-established pattern-recognition approaches. One agent is based on the vector space algorithm; the other is based on a Bayesian algorithm.

The Vector Space Agent is based on a vector space model and is trainable for specific categories. The vector space text-categorization algorithm has been tested using computer hacking–related sites.

4.3.3 Bayesian Agents

A Bayesian classifier, based on a pattern-recognition approach, was developed to function as both a categorizer and a retrieval agent. E-mail messages belonging to different categories were used to train and test this agent. As a precursor, an information extractor agent was developed to

- Retrieve HTML-based Web pages or documents
- Process Web pages to filter HTML tags and irrelevant words
- Implement algorithms for processing recursively stemming words
- Feed processed words to the text-categorization agents

4.3.4 Document Clustering Agent

A prototype version of the Document Clustering Agent was developed for automatically classifying URLs and text documents. The Document Clustering Agent groups similar URLs and similar documents. This agent has potential for use in author clustering or identification and in Web page clustering or classification.

4.3.5 Comparison of Text Categorizer Performance

The ANSER team conducted a performance test on six text-categorization agents, including CDM and Naïve Bayes using NewsGroup postings related to the topic "Traders of Child Pornography." One thousand Usenet news postings were collected from alt.binaries.pictures.erotica.pre-teen, a NewsGroup that includes many child pornographic images. A sampling of other Usenet topics in other NewsGroups (rec.arts.movies, sci.physics, and talk.politics.guns) was also included. The rationale was to train the algorithms on data related to child pornography and the topic "Traders of Child Pornography" but also to include data related to other typical Usenet topics.

The three primary factors contributing to the ease or difficulty of a category are

- Whether there is a direct relationship between the features in the text and the output category
- Whether there are a sufficient number of positive and near-positive examples
- Whether there is a sufficient amount of text (more than one full sentence) in each of the positive examples

The collection contained approximately 500 postings from the child pornography group and 500 from the other groups. The collection was formatted, and the postings were labeled either "junk" or "cporn." The category "junk" denoted messages that were off-topic for those posted to a NewsGroup. Junk messages are often commercially oriented and represent the bulk of messages posted to a NewsGroup. The category "cporn" (for child pornography) denotes the category of individual traders in child pornographic

materials. The corpus of these postings was then divided into two groups: training data and performance test data.

After the news postings were labeled, the agents were operated:

1. Training against the training data was performed so that text categorizers could learn the categories of "junk" and "cporn"
2. Categorizers were run on the new news stories
3. Performance of these categorizers was evaluated

A text categorizer produces a set of binary decisions ("yes" or "no"), identifying whether each document is about the category. The set of decisions made by the categorizer can then be compared to the set of decisions made by a human. The results are recorded in a contingency table (Exhibit 6).

	Human Decides Yes	**Human Decides No**	
System Decides Yes	*p*	*fp*	*p* + *fp*
System Decides No	*fn*	*N*	*fn* + *n*
	p + *fn*	*fp* + *n*	*p* + *fp* + *fn* + *n* = *N*

Exhibit 6—Contingency Table for a Set of *n* Binary Decisions

Recall and precision, the traditional measures of performance used in the field of information retrieval, were computed for each text-categorization algorithm.

Recall:

$$\frac{\text{Total correct category assignments made by the program}}{\text{Total category assignments made by the human}} = \frac{p}{p + fn}$$

Precision:

$$\frac{\text{Total correct category assignments made by the program}}{\text{Total category assignments made by the program}} = \frac{p}{p + fp}$$

By varying a parameter that affects the tendency for an algorithm to assign categories, the system can be made more willing, for example, to say "yes," and recall will increase. At the same time, precision normally will decrease. For a given algorithm and set of documents, the value of a breakeven point is found by equating the recall and the precision. Exhibit 7 shows values of the breakeven point for two categories—"Traders in Child Pornography" and "Junk."

Category 1—Traders in Child Pornography

Algorithm	Breakeven Point
CDM	40%
Naïve Bayes	31%
Remaining 6 methods	<10%

Category 2—Junk

Algorithm	Breakeven Point
Naïve Bayes	97%
CDM	88%
Remaining 6 methods	<70%

Exhibit 7—Results of Text-Categorization Test

The breakeven point can be considered a single summary figure indicating the accuracy of a text-categorization system. If different text-categorization systems have been applied to the same data under the same experimental conditions, then the values of the overall breakeven point can be used to compare the relative effectiveness of the different categorizers. As shown in Exhibit 7, Naïve Bayes and CDM outperformed the other six methods. Naïve Bayes and CDM performed best on the "Junk" and "Traders in Child Pornography" categories, respectively.

All algorithms performed significantly better on the "Junk" category. For this category in each of the three cases presented in Exhibit 7, all three questions relative to (1) relationships between features in text and output category, (2) sufficient positive and near-positive examples, and (3) sufficient amount of text are answered "yes," while the answers for all three questions for the "Traders in Child Pornography" category are "no." Thus all three factors make this particular text-categorization example for "Traders in Child Pornography" category very difficult.

4.4 Data Recognition (Data Mining and Knowledge Discovery)

A fourth area, data recognition (data mining and knowledge discovery), focuses on developing data-mining approaches and identifying U.S. and international information resources that could prove valuable in locating missing children and their abductors. The ANSER team developed a baseline capability to interface either directly or with limited customization to generic and specific databases used by law enforcement and to identify the steps necessary to access the data and to improve operational effectiveness.

Questions arose throughout the effort regarding politics, security, and privacy. ANSER made serious attempts to evaluate the validity of these issues and to explore ways of overcoming them to obtain access to the data.

An evaluation of data-mining tools identified several commercially available products, which were compiled by the primary technique used, such as clustering, classification, prediction, or discovery and other parameters such as cost, hardware platforms supported, data sources accepted, the ability to generate a source code, and the availability of demonstration software.

The ANSER team also designed and implemented several ISAs. Because data come in a variety of forms, some ISAs attempt to ascertain the usefulness of data obtained. As new data sources are discovered, ISAs determine the data formats, compatibility, and conversion requirements necessary to make the data usable. Agents developed by the ANSER team are discussed below.

- **Prospector Agents**—Prospector agents are invoked by surveyor agents (Section 3.2) and pointed toward promising data sources. Prospector agents perform knowledge discovery and preformat the encountered data for further processing by the miner agents. Similar to conventional data-mining tools, prospector agents attempt to find interesting relationships among groups of database tables and fields and construct standard symbolic data-definition formats. Prospector agents can run in fully autonomous mode or guided mode. In guided mode, a prospector agent investigates hints supplied by the KB search engine. One or more knowledge engineers previously have developed the KB.
- **Miner Agents**—Miner agents can receive their data from prospector agents and identify interesting relations contained in the data, then relay them in either human- or machine-readable format (for example, rules for expert system shells). Miner agents can come in different flavors, exploiting various biases in multistrategy learning approaches. Two such strategies were implemented:
  - **IBLMiner** uses instance-based learning (IBL) to mine potential relationships discovered by the prospector agents. IBL agents form rules that can be fed into an expert system shell or the KB browser or search engine.
  - **TDLMiner** uses transdimensional learning (TDL) to form a neural network–based pattern-recognition system. TDL agents automatically grow a neural network from raw training data and are capable of cooperating and sharing their learning experience. TDL agents communicate among themselves via either Windows Messaging or a shared open database connectivity source. The latter mode allows TDL agents to communicate with other ANSER-developed agents (such as prospector agents and miner agents). TDLMiners work off the same data as the IBLMiner to mine interesting relations.
- **Collector Agents**—These agents collect and fuse neural networks created by the TDLMiner into a single neural network. The user can ask the network to predict new patterns for any of the learned contexts.
- **KB Browser**—The KB browser allows a knowledge engineer to browse a KB of rules that the IBLMiners generate.

- **KB Search Engine**—The KB Search Engine allows a knowledge engineer to provide hints to the Knowledge Discovery and Predictive Data Mining system regarding which databases, tables, and fields may contain interesting relations. Once hints are supplied, the system attempts to derive rule bases describing these relations. These rule bases will be reported back to the KB search engine.

5. Agent Management Structures

5.1 Symbolic Agent Development and Integration Environment

In response to the initial requirement to develop "expert systems for case investigations," the ANSER team developed SADIE, a system that can create knowledge-based intelligent agents and supports ISA planning and problem solving. SADIE modules include

- **Knowledge Base Manager**—The KB manager component controls the creation and modification of KBs (for example, rules, concepts, instances, goals, and relationships). It includes file kernel, input and output, insert, modify, delete, and query modules for managing the KB. The KB manager also permits some degree of agent introspection (with knowledge about its knowledge) and planning (with goal, event, plan, step, and resource assignment KB elements). This module also includes a function for importing objects, relations, and facts from a database into the KB. The user interface includes a dictionary editor browser, concept or instance editor, association browser, concept or instance browser, rule editor, event editor, and goal editor, as well as an object-definition screen for use with the import module.
- **Inference Engines**—Two inference mechanisms were developed: (1) a crisp inference mechanism based on constraint posting and backward chaining inference and (2) a fuzzy inference mechanism based on fuzzy sets and fuzzy logic.
- **Planning System**—A decision process was developed to project state changes through time and to make decisions to maximize agent goal satisfaction using Bayesian decision trees.
- **Server**—A Lisp-based server was developed to exchange sockets with a Java interface.

The ANSER team also developed a Java interface to the SADIE Lisp-based server. A Java GUI serves as the layer that sits between the user and the Lisp-based server. This interface allows the KB manager to open, create, save, or delete KBs and to browse, edit, and import the individual objects (rules, concepts, instances, goals, and relationships) that make up the KB. Development consisted of implementing and debugging the basic functionality mentioned above and depended on a series of "wizards" to guide the user through the process of creating or editing some of the more complex KB elements, such as fuzzy sets and rules.

The ANSER team used SADIE and the data-mining tools to develop a KB using the National Incidence Studies of Missing, Abducted, Runaway, and Throwaway Children

1988 database developed by the Office of Juvenile Justice and Delinquency Prevention. These sample data demonstrated the capabilities of data-mining and knowledge-based agents to uncover patterns and relationships hidden in the data. The team used the data to develop concepts and rules related to cases of missing children. A demonstration system was also developed implementing SADIE. A sample KB and user interface were developed to determine to what extent a person matches a given criminal profile, such as that of a child molester or an abductor of newborns or infants.

5.2 Evolutionary Agent Societies of Intelligent Software Agents

These systems are starting to be applied to semi-autonomous and autonomous decision-making and various artificial intelligence methods (for example, expert systems, neural networks, genetic algorithms, fuzzy logic, and reinforced learning) to communicate and coordinate activities as they continue to adapt and learn to

- Search, monitor, collect, and analyze data and information from public sources, particularly the Internet—for applications such as data mining
- Provide support for autonomous and semi-autonomous planning, scheduling, management, and coordination of agent activities—such as in Web crawling and information retrieval

ANSER is developing a suite of ISAs for law enforcement that conduct autonomous activities such as Internet and intranet search and monitoring, data mining of semi-structured data sources, and content identification via text categorization. Additional agents capable of advanced data mining and threat assessment agents are being developed. However, these agents currently operate as independent processes with limited multi-agent communication, cooperation, or control. ISA effectiveness, efficiency, resource consumption, and reliability can be improved by implementing an EAS—a society of individual ISAs that can learn to cooperate over time and exchange information with one another based on past cooperation. To accomplish these challenges, the ANSER team (1) endowed individual agents with persona, which define the agent's basic behavioral and problem-solving skills, and (2) employed evolutionary principles (such as genetic algorithms) and reinforced learning techniques (such as multilayered feedback channels) to evolve individual populations of agents. It is anticipated that a rich pool of agent behavioral traits and capabilities will give rise to the emergence of agent societies.

Automatic recognition will require that specialized processing be developed from the viewpoint of a framework that is available to support adaptation and learning. The intelligence of the overall systems and applications rests on the framework's ability to support many ISA components and their communication. For example, multiple search agents will explore, locate, and retrieve files from Internet and other sources and will communicate what they have found (and where they have found it) with text-categorization agents. These agents will quickly prescreen and simultaneously examine the textual content of the site using machine-learning algorithms for text categorization.

This preprocessing will aid in the efficient selection of the most effective recognition processing options based on the content of the information and on experience.

EASs of ISAs expand current agent-based solutions for recognition by further emphasizing reliability, scalability, and adaptability. EASs can be applied to the following areas:

- Investigation and exploitation of data sources found on public intranets and the Internet
- Monitoring, control, and modification of their own behavior

As ISAs become more prolific and consume computer resources at an ever-increasing rate, a mechanism must be developed for monitoring, metering, controlling, and modifying their own behavior. To achieve this agent behavior modification process, it is necessary to

- Develop and implement an integrated agent framework that incorporates in-process metering of the consumption of computational resources.
- Provide a sample library of optimization algorithms from the generic to the application-specific, which will allow the agent developer to embed adaptive intelligence in both the design process and the run-time strategies. This library would include
  - A genetic algorithm to evolve mutants
  - A selection function, trained using a neural network learning technique
  - A continuously adaptive attention (sampling) heuristic

The ANSER team developed several agents that monitor and control the execution and distribution of other software agents:

- **Monitor Agent**—This agent monitors system resources (such as central processing unit workload and communications overhead) of a particular machine and communicates and coordinates with other monitor agents running on other systems within a network. The monitor agents can distribute agents onto other networked machines and can block new agents from being executed if the machine they are monitoring is overcrowded.
- **Scavenger Agent**—The scavenger agent helps to keep hard disk resources available by cleaning up the refuse (deleting old files that are no longer in use) left behind by agents that have terminated prematurely.

6. Prototype Demonstration Systems

The ANSER team integrated several face-recognition and ISA components (discussed in Section 4) into several prototype demonstration systems, including

- MCLA, v. 1.0
- CBSSA
- IdentiFace
- COPIES, v. 1.0
- DSRegistrar
- Internet Investigator
- CASA
- NewsRetriever
- Text Categorizer
- Disk Hound

6.1 Missing Children Locator Agent, v. 1.0

The ANSER team developed MCLA, a system composed of ISAs that continuously search the Web for images and match them against a database of images of missing children. The MCLA system includes the integration of the BloodHound Web spider with face recognition. A relational database management system was implemented for storing the database images and templates of missing children, the results of Internet searches, and the ISA search logs.

6.1.1 Internet Search—Integration of BloodHound With Face-Recognition Search and Match Components

For each gallery image, the MCLA system builds a list of the best matches and updates this as the search continues. A case manager or investigator can periodically review the list of best matches for a particular missing child (see Exhibit 8). The results are available on the Internet or an intranet through a dynamically generated Web-based user interface. Links to the pages where images were found are also displayed. Thus, investigators can quickly search the website for additional clues.

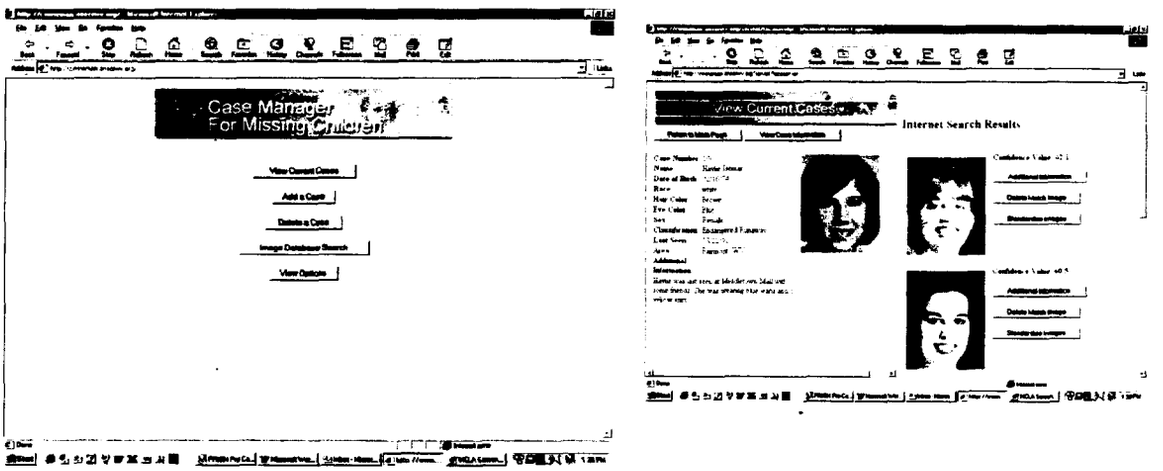


Exhibit 8—Investigators Periodically Check for Matches Using a Browser-Based GUI

6.1.2 Database Search Based on Submitted Digital Images

In addition, an image submission feature in MCLA, CBSSA, and IdentiFace can take an input image as a query, which may be entered remotely over the Internet. For example, a law enforcement agency may have arrested a juvenile whom they suspect is a missing child. The agency can photograph the juvenile, log onto the MCLA site via the Internet, and upload the juvenile's photo to be compared against the database of missing children. The top 10 to 20 closest matches (user selectable) are displayed (see Exhibit 9). This type of search could be conducted from any location (stationary or mobile) in the world with an Internet connection (and permission to use the system if it is not an open system).

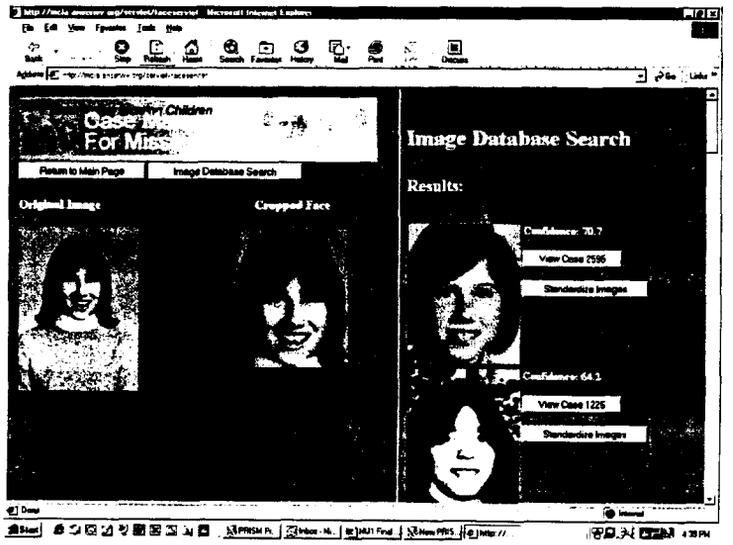


Exhibit 9—Results From a Submitted Image

ANSER has formed a relationship with the West Virginia Missing Children Clearinghouse and the West Virginia State Police to set requirements for and to test and evaluate MCLA. Many sources of images of known missing children have been

downloaded into the ANSER team's missing child gallery database. An operational version of MCLA is being run now for the West Virginia Missing Children Clearinghouse. The West Virginia State Police periodically review the results.

6.2 Criminal Booking System Search Agent—Digital Mug Shot Comparison and Match

By integrating the face-recognition components of the MCLA or IdentiFace systems with a digital booking system, a law enforcement officer can, in almost real time, query the criminal booking database system using images of suspects captured on film or through composite sketches as described by eyewitnesses. With the MCLA Web-based interface, the officer can even run these queries from cross-jurisdictional or intrastate organizations.

For example, a law enforcement officer may take a suspect into custody and

- Photograph the suspect
- Log onto the CBSSA site via the Internet
- Submit a digital image to form a facial query
- Wait for the system to compare the submitted digital image to a gallery database of images of known criminals and rank the potential closest matches
- Upload the resulting potential closest matches of suspects
- Inspect the top 10 to 20 (user selectable) closest matches displayed as thumbnail photographs

6.3 IdentiFace—Integration of Facial Capture From Live Feed or Videotape With a Face Search Engine

ANSER and its South Florida HIDTA pilot partner have developed requirements for testing and evaluating IdentiFace, which will capture images (mainly live video-camera surveillance scenes but potentially still photographs) of suspected criminals. Using automated face-recognition technology, IdentiFace will match these probe images to a gallery database of booking photographs and other stored digital images. A short list of the highest probable matches will be generated and ranked to aid identification of narcotics traffickers and related criminals. As the program is currently envisioned, these searches for matches would be carried out in near–real time. ANSER's version of this system will include the following modules:

- Components to capture a single face (or a sequence of faces) for persons under surveillance
- A component to convert the images into an appropriate facial representation format for facial recognition
- An automated face-recognition engine that can operate into a gallery database of booking photographs and other stored digital images

- A gallery database that can be used at a centralized site or taken into the field for processing video
- A reporting system that interactively displays the probe and matching gallery database images to the surveillance officers

IdentiFace uses live real-time surveillance video, videotapes, and stills as sources and matches faces in the images to gallery databases of booking photographs. Governmental sources of images in these booking databases and the videotape files include these in south Florida:

- Monroe County
- Broward County
- Metropolitan Dade Medical Examiner Department

6.4 Child Online Pornographic Image Eradication System, v. 1.0

The ANSER team developed COPIES as a system of ISAs integrated with digital signature utilities to autonomously find image files that match known child pornography and report the results to law enforcement agencies without further exploiting children (by not continually viewing the children's images). BloodHound continuously searches the Web for images. Once an image is found, its digital signature is computed as a probe digital signature that is compared to a gallery database of digital signatures of known child pornography. A match of digital signatures indicates that a copy of known child pornography has been located on the Internet. A report is generated to specify when and where the copy was found. Law enforcement officers can use this report to focus their investigations on Internet sites where known child pornography is being disseminated (see Exhibit 10).

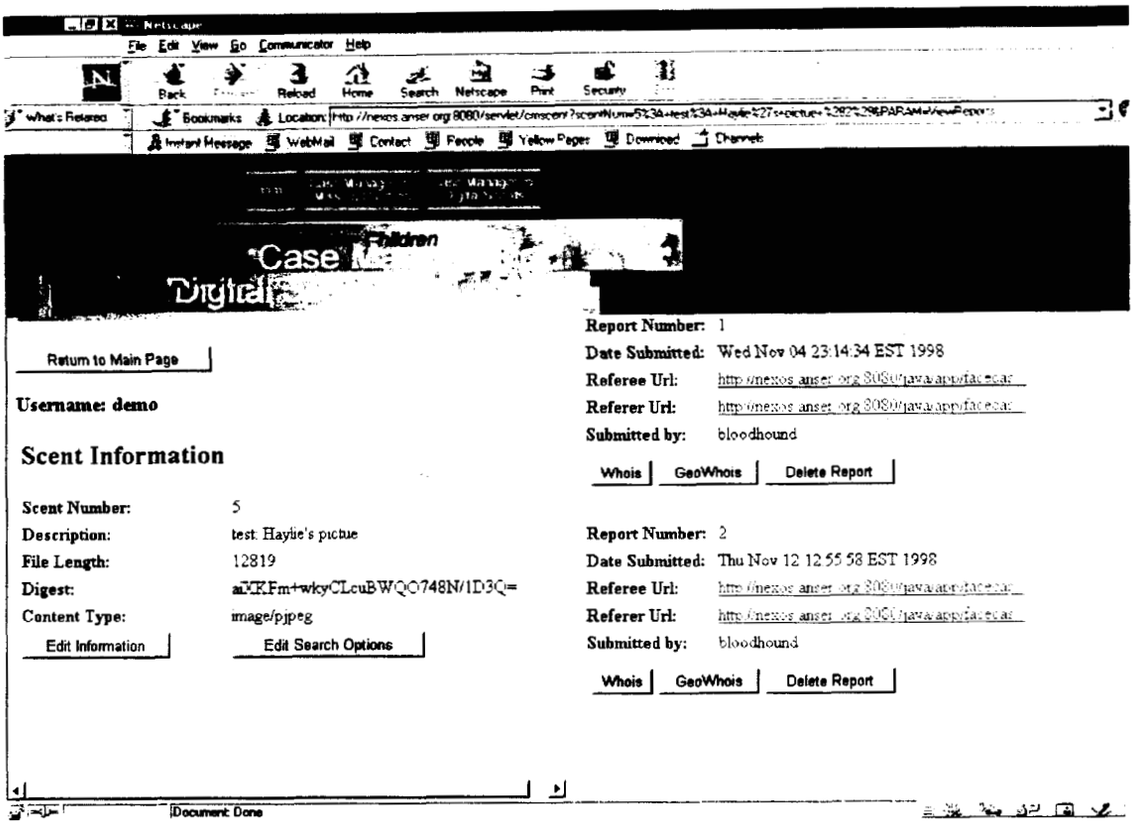


Exhibit 10—COPIES Reports When and Where a Copy of an Image Was Found

The U.S. Department of the Treasury and potentially the FBI are the initial targeted pilot project partners for deploying and using COPIES (see Sections 7.3 and 7.4). Both command-line and GUI versions of COPIES have been developed in Java (see Exhibit 11).

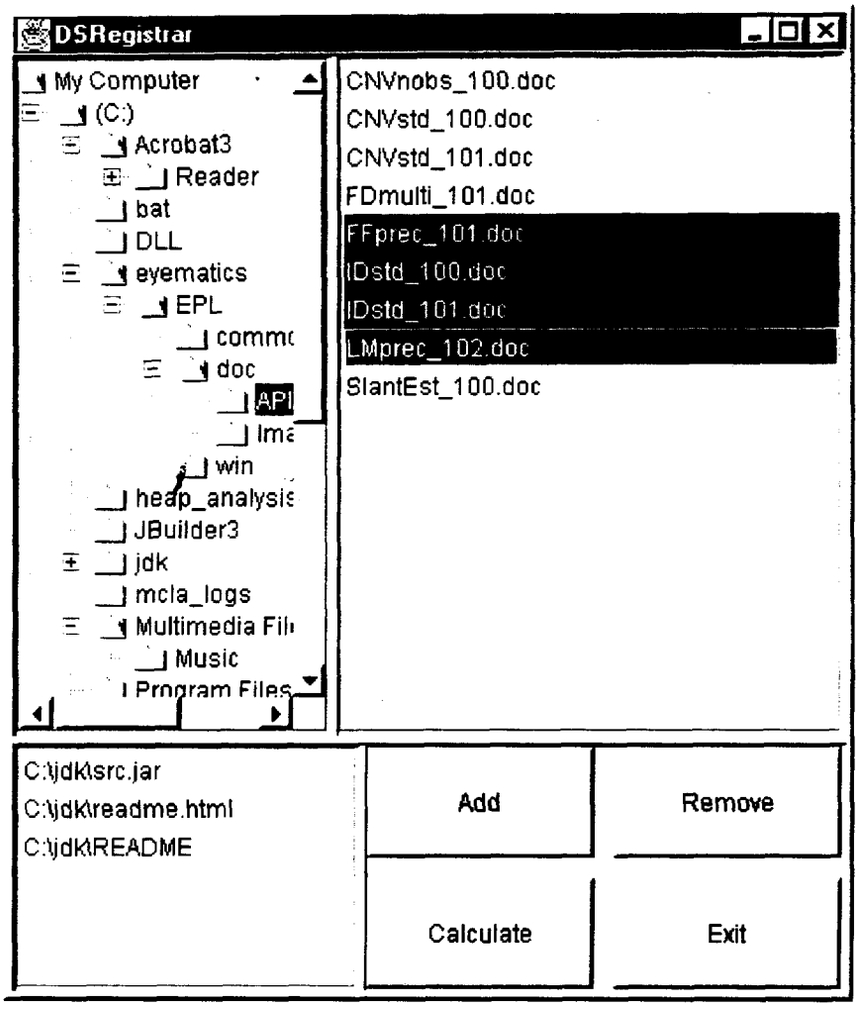


Exhibit 11—DSRegistrar GUI

6.5 DSRegistrar, v. 1.1

The ANSER team also developed the DSRegistrar utility to automatically encrypt large collections of known image files as digital signature strings. This utility also automatically populates gallery databases with a Digital Signature of the image. Automated comparisons can be made between the unknown probe Digital Signature and the gallery database. DSRegistrar is applied as an adjunct to COPIES on files of known child pornography.

6.6 Internet Investigator, v. 1.0a

Internet Investigator is a low-bandwidth-intensive application (able to run with a 28.8-Kbaud modem connection) plus interactive agents that are useful in investigating Internet-based resources.

Internet Investigator has been designed to be modular and have user interfaces with a standard look and feel. Internet Investigator will therefore permit new features to be easily added in the future.

6.7 Case Analysis Software Agent

The ANSER team has been applying core components of SADIE to develop a CASA, which will manage information about entities, such as criminals and vehicles, in a highly flexible record format. Information in CASA records can be customized to reflect the information needed for a specific case. This is in contrast to a database system where some fields do not apply to a given case and other fields do not exist to store relevant case information.

6.8 NewsRetriever

The ANSER team developed NewsRetriever to search NewsGroups for specific articles. NewsRetriever can be trained to return only NewsGroup articles that contain relevant information. Its design is sufficiently flexible for integration of almost any artificial intelligence filtering component, including text categorization, image processing, and face recognition.

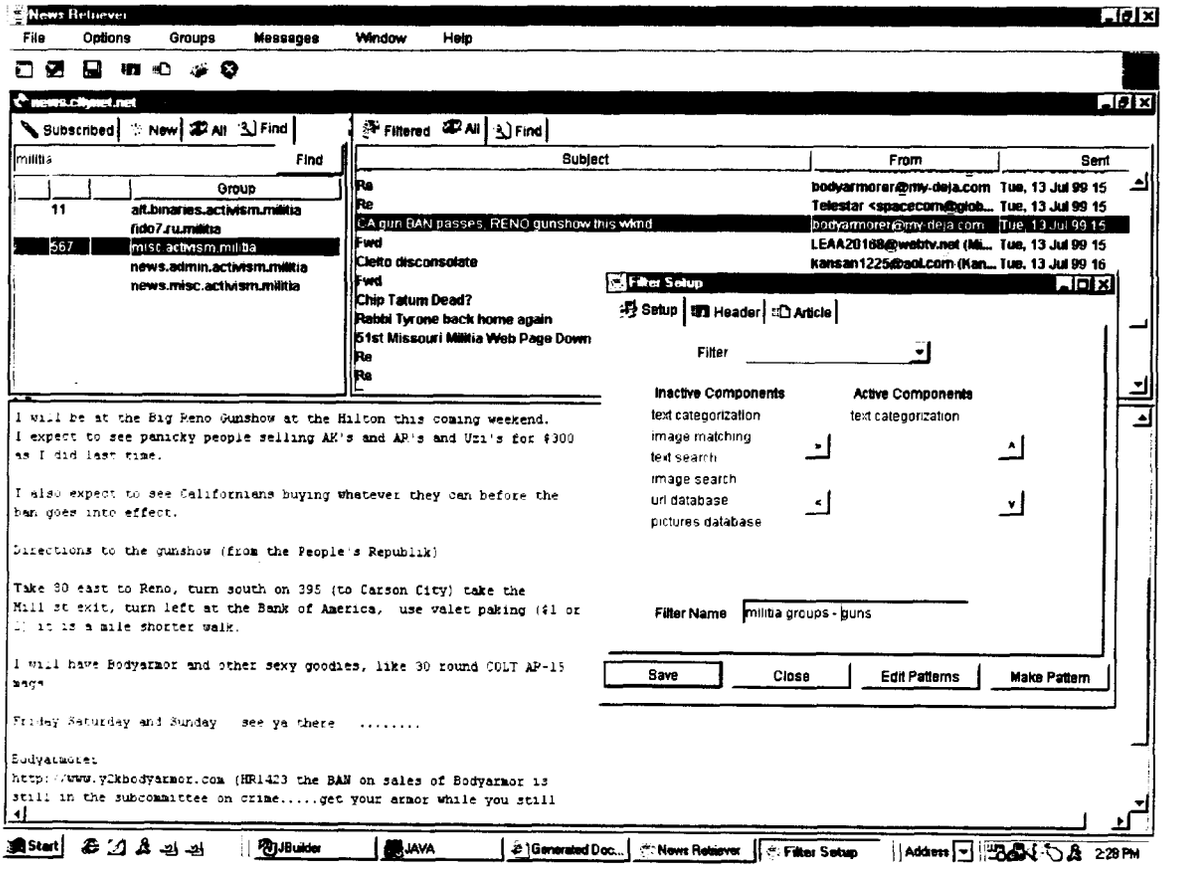


Exhibit 12—NewsRetriever Agent With Filter Setup Screen Enabled

6.9 Text Categorizer

The text-categorization agent is designed to enable a law enforcement officer

- To easily develop a specific text categorizer for each text-categorization profile, based on holdings of documents—for example, NewsGroup postings within the topic area "traders in child pornography"
- To use this text categorizer and other filters to preprocess new NewsGroup postings and report only articles that match the developed document profile

This capability will provide huge savings in law enforcement personnel staffing requirements.

The ANSER team is developing this system first with the U.S. Department of the Treasury, where it will pilot test the application. In addition, the ANSER team may develop a pilot project for text categorization with the FBI Innocent Images Program.

6.10 Disk Hound, v. 1.0

While working with the West Virginia Missing Children Clearinghouse, the ANSER team learned that the home personal computer is becoming more prevalent in cases of missing and runaway children. In one case, a runaway teenager was suspected of befriending an individual via the Internet. It was also suspected that the teenager ran away to be with that individual.

The ANSER team conducted manual data mining to help the West Virginia State Police recover and analyze data from the hard drive of the computer that the runaway had been using.

Clues about the child's location were left on the computer. This information corroborated other leads in the case, which ultimately led to finding the runaway. The ANSER team assisted the West Virginia State Police in locating the runaway child.

The ANSER team performed these manual computer forensic activities:

- Mirrored the hard drive of the missing child's home computer and ran all investigations using the mirrored drive. This process avoided corrupting or destroying part of the potential chain of evidence.
- Undeleted files using various disk utilities.
- Searched the drive (at the file and sector levels) for the first names and aliases of known Internet acquaintances.
- Uncovered and identified several items of e-mail correspondence between the missing child and a circle of Internet acquaintances.
- Uncovered correspondence indicating plans to run away.

- Uncovered a list of Internet acquaintances' aliases, phone numbers, and, in some cases, addresses.
- Searched America Online® for user profiles related to a user's screen name or alias.
- Produced a report for the West Virginia State Police summarizing and correlating this information.

This manual process also helped the ANSER team develop the requirements and capabilities necessary for an automated Disk Hound computer forensics tool (Section 3.3) tailored to cases of missing children. In Disk Hound, ISAs will ultimately automate the collection and analysis of information from personal computers.

Disk Hound is designed to find files with both specific and user-defined embedded data, such as e-mail addresses, URLs, documents, textual strings, and keywords or phrases. The user will be able to choose specific search tools to locate any number of file types corresponding to images, documents, multimedia, or user-defined file extensions (see Exhibit 13).

Image Files	Document Files	Multimedia Files
ART	WPS	AVI
ICO	WPD	MPG
PCX	DOC	MPEG
PSD	RTF	MOV
BMP	TXT	WAV
DIB	HTM	RAM
GIF	HTML	MIDI
IMG	VRML	RMI
JPG	ASP	
MAC		
PIC		
TIF		
WMF		
WPG		
XBM		

Exhibit 13—Disk Hound Will Find and Report All Files With the Selected Extensions

Disk Hound meets the West Virginia Missing Children Clearinghouse requirements.

7. Pilot Projects Building Upon Prototypes

As the ANSER team started to develop the individual components and integrate them into prototype demonstrations, system requirements were learned from subject matter experts in the pilot project law enforcement organizations. By more fully developing and analyzing these requirements, the ANSER team was able to define system architectures and system designs and to develop adequate requirements documentation. The requirements analyses were to ensure that

- An effective system could be developed
- The software being developed met the defined system operational and functional requirements
- Systems could be built using existing components rather than requiring creation of new ones
- Data resources could be identified, collected, and integrated into each application
- A system-level test plan could be generated
- Software components could be seamlessly integrated within the user environment
- Plans could be developed for system installation and onsite training of personnel

ANSER coordinated with several Federal and state law enforcement organizations to assess their needs for face-recognition and ISA software. To better understand the software capabilities that law enforcement officers require, ANSER met with and demonstrated software to the West Virginia State Police, the South Florida HIDTA, the U.S. Department of the Treasury, the FBI, the National White Collar Crime Center, and the U.S. Secret Service. The ANSER team also met with developers of various commercially available link and case analysis tools and text-retrieval products to evaluate related products (for example, Autonomy Systems and I2).

As a result of the demonstrations, presentations, and discussions, ANSER successfully attracted three (potentially four) law enforcement organizations interested in becoming pilot partners and providing sites for testing and evaluating the technologies under development. Support from the pilot project partners will be crucial in the test-and-evaluation phase.

Two (and potentially three) of these pilots are oriented toward fighting the exploitation of children and locating missing children. The fourth pilot broadens the application of these technologies into the domain of near–real time identification of suspects from surveillance video cameras. These technologies in near–real time surveillance and identification can similarly be used to identify suspects in the review of videotapes from surveillance operations.

Short descriptions of each pilot project partner and the envisioned pilot project are provided below.

7.1 West Virginia State Police

The West Virginia State Police, headquartered in Charleston, WV, are responsible for operating the West Virginia Missing Children Clearinghouse and therefore conducting investigations to locate West Virginia's missing children. This pilot project is designed to test and evaluate several integrated automated ISA tools, such as MCLA, CASA, and Disk Hound.

The goal of this pilot project is for the West Virginia State Police to apply new tools and methods for recovering abducted and runaway children and combating sexual exploitation of children.

7.1.1 Missing Child Locator Agent

MCLA integrates autonomous Internet search agents coupled with face-recognition agents to locate missing persons, especially missing children. During this project, a gallery database of photographs of missing children is being collected from the West Virginia State Police. Search agents such as BloodHound (and maybe later, NewsHound and FileHound) will continuously search the Internet for information (checking, for example, sites with yearbook photos, pornography, and names) that may be related to a missing child case. When a photograph or video segment is discovered, automated face-recognition agents will compare it to the gallery database of missing children and report high-probability matches. This system will give law enforcement agents interactive capabilities with photographs. Version 1.0 of MCLA was completed and will be deployed in October 1999.

7.1.2 Case Analysis Software Agent

CASA is an ISA that uses a sophisticated KB manager to store information in a flexible way, learn new information through the application of data-mining techniques, and retrieve information using a powerful query mechanism. This will be deployed at the end of 2000.

7.1.3 Disk Hound

Disk Hound will integrate numerous commercial off-the-shelf forensic tools to find, search, and analyze data files. The ANSER team recognized the viability of the integrated Disk Hound system as it assisted the West Virginia State Police in recovering and analyzing data from a computer's hard drive. The West Virginia State Police recognized ANSER's involvement by writing that ANSER recovered leads that, combined with other information, seem to show "that the child is still alive and identified the town in Illinois he lives in." This information corroborated other leads in the case, which ultimately led to finding the runaway.

This work also helped the ANSER team develop the requirements and capabilities necessary for the Disk Hound computer forensics tool (Section 6.10), tailored for missing children's cases.

The ANSER team will aid the West Virginia State Police in establishing a state-of-the-art computer forensic capability with Disk Hound.

7.2 South Florida High Intensity Drug Trafficking Area

The South Florida HIDTA is a Drug Enforcement Administration organization that fights drug trafficking. As one of two dozen Drug Enforcement Administration–operated HIDTAs throughout the United States, the South Florida HIDTA is made up of more than 50 law enforcement and social service agencies. The South Florida HIDTA employs fixed and mobile surveillance units to monitor the movements of suspected drug traffickers and must rapidly identify individual suspects under surveillance. The ANSER team has primarily interfaced with only a few of these organizations, including

- Broward County Sheriff's Office
- HIDTA office
- Metropolitan Dade Medical Examiner Department
- Miami Police Department
- Monroe County Sheriff's Office
- U.S. Marshals Service

The goal of this pilot project is to work with selected organizations in the South Florida HIDTA and apply ANSER's IdentiFace to successfully identify suspects in video surveillance (both live and taped). This is a collaborative effort to tailor a solution based on the ANSER team's capabilities to design and implement a facial-recognition system and on the South Florida HIDTA law enforcement officers, conducting criminal investigations, to test and evaluate capabilities of the system in real-world law enforcement environments.

IdentiFace will capture images (live video-camera and videotaped surveillance scenes and still images) of suspected criminals. Using automated facial recognition, IdentiFace will match these probe images with a gallery database of booking photographs and other stored digital images. A short list of the highest probable matches will be generated to aid identification of narcotics traffickers and related criminals. As the program is envisioned, these searches for matches would be carried out in near–real time. ANSER provided the initial version of this system to the South Florida HIDTA (with in-person demonstrations, user training, and documentation) in early 2000.

Mobile and eventually mobile cellular face-recognition systems will enable Federal and state narcotics task forces to quickly make positive identifications of offenders recorded by surveillance cameras and associated equipment.

7.3 U.S. Department of the Treasury

The Department of the Treasury is one Federal law enforcement agency that is responsible for the investigation of electronic trafficking in and distribution of child

pornography. ANSER has established a relationship with the Treasury office in Fairfax, VA. The goal of this pilot project is to test and evaluate ANSER-developed systems in combating the sexual exploitation of children. The plan for this pilot project is to develop, test, and evaluate technologies that will increase the effectiveness of Treasury Department investigation of Internet sites or other sources of information. These systems will be used.

- To locate and identify relevant information
- To develop cases for prosecuting distributors of child pornography
- To identify children victimized by child pornographers

The ANSER team will develop operational software system prototypes specifically to meet their needs. These prototypes will integrate ISAs and face recognition to aid them. in investigating child pornography on the Internet. ANSER will provide them with a single integrated software system (along with in-person demonstrations, user training, and documentation) to use and evaluate. This system will include the COPIES and DSRegistrar modules.

7.3.1 DSRegistrar

DSRegistrar computes and stores digital signatures of known child pornographic images. The ANSER DSRegistrar agent is being used to develop a database of hundreds of thousands of digital signatures of "certified" instances of child pornography. This large database of digital signatures of "certified" child pornographic images forms the baseline for operating COPIES.

7.3.2 COPIES

COPIES operates on any image file to calculate a digital signature, which it compares to digital signatures of known pornographic images. COPIES will automatically

- Accept the search specifications of the user (to search one-to-one or one-to-many)
- Carry out directions automatically and continuously within Internet sites
- Find and capture Internet still and video images, as well as ancillary information, of potential child pornography
- Process and match these returned images to a database of digital signatures of known child pornography
- Generate and provide a report for review of the matching images and their URL locations

COPIES performs tasks that are virtually impossible for law enforcement officers to conduct manually. Within 48 hours from the beginning of an initial test using 29,000 digital signatures of known child pornography, COPIES found its first three instances of child pornography on the Web. COPIES reported the time the image was found and the

URL of the site posting the illegal image. Although the site was outside the United States, this test quickly proved that the concept of operations is valid.

This partnership between ANSER and the Department of the Treasury is expected to emplace some of the ANSER prototypical systems that will provide increasingly automated—efficient and effective—support for investigators, saving time in browsing and monitoring Internet sites for extended periods to find, isolate, and mitigate child exploitation.

7.4 FBI

The goal of this potential pilot project is for the ANSER team to develop, integrate, and provide versions of COPIES and DSRegistrar tailored to the FBI Innocent Images Program needs and the concept of operations at the Maryland Metropolitan Office at Calverton, MD.

This FBI pilot partnership is expected to support the goals and objectives of the Innocent Images Program and to benefit the program by implementing ISA capabilities according to FBI guidelines. These systems will support investigative personnel at the FBI by

- Providing increasingly automated support
- Increasing efficiency and effectiveness
- Reducing the time required to extensively browse and monitor Internet sites
- Helping to identify child exploitation and child exploiters

This customized COPIES will automatically

- Accept the search specifications from the investigator
- Carry out specified searches continuously within public and predicated sites
- Find and capture Internet still and video images and ancillary information about potential child pornography
- Process and match these returned images to a database of digital signatures of known child pornography
- Generate and provide a report for review of the matching images and their URL locations

If the existing digital image database of known child pornography has not been converted into digital signatures, then, at the discretion of the FBI, the ANSER team will provide DSRegistrar software to automatically generate and store digital signatures in requisite files.

8. Dissemination

ANSER developed a strategy to disseminate its software systems to law enforcement through

- Presentations to individual law enforcement agencies
- Demonstrations at law enforcement agency conferences
- Focused meetings with law enforcement investigators and technologists
- Installation of CD-ROMs
- Users' manuals and concepts of operations

ANSER initially met with the following organizations:

- West Virginia State Police
- Federal law enforcement agencies, including the FBI, the Department of the Treasury, and the U.S. Secret Service
- Several organizations making up the South Florida HIDTA
- The National White Collar Crime Center, the National Law Enforcement and Corrections Technology Center–Western Region, and the California Department of Justice (Megan's Law and Calphoto initiatives)

ANSER also pursued transferring these technologies to support counterterrorism and counterdrug initiatives within the Department of Defense, including the Defense Intelligence Agency Center, the Naval Surface Warfare Center, and the U.S. Army Intelligence and Security Command.

ANSER also held discussions with the Office of Law Enforcement Technology Commercialization to assist ANSER in determining the potential

- For marketing various ANSER systems to law enforcement agencies
- For fostering partnerships with commercial companies

In promoting its efforts, ANSER participated in several conferences and seminars, including

- The 1998 NIJ Technology Fair, Washington, DC
- The first National Commercialization Conference, sponsored by the Office of Law Enforcement Technology Commercialization, Orlando, FL
- The 1999 conference on Technology and Tools for Public Safety in the 21st Century, sponsored by the Center for Technology Commercialization, Orlando, FL
- The 1999 Law Enforcement Information Management Training Conference and Exposition of the International Association of Chiefs of Police, Ft. Lauderdale, FL

To automate software installations of the MCLA face-matching tool, ANSER created installation routines using Install Anywhere. The MCLA system is now available on CD and can be easily installed on any computer that has a CD-ROM drive.

9. Results

Results of this NIJ grant are broad-based, covering a wide range of concepts, components, and prototypes in the areas of face recognition and ISAs as well as a great deal of depth in many areas. Our project has focused on technological solutions to satisfy a number of critical law enforcement requirements for automated information search, processing, and reporting.

The ANSER team, in the course of this grant effort during the past 2 years, has successfully developed software solutions applicable to law enforcement requirements and expects to continue to provide advanced, robust, and scalable (potentially enterprise-level) software systems.

The ANSER team has been using the following process in specifying, designing, implementing, testing, and evaluating several software systems:

- Understand law enforcement needs and specific requirements
- Understand the available technology base for face recognition, search agents, other ISAs, and databases
- Design and implement many individual components for face recognition, Web browsing, text categorization, file recognition, data mining, and knowledge management
- Understand identification, collection, and integration of data resources for each application
- Validate the robustness and scalability of these components relative to input, retrieval, and storage capacities
- Integrate the components into robust, scalable prototype systems, tailored to automated collection, analysis, and decision support for very large information holdings
- Demonstrate the prototypes to local, state, and Federal law enforcement organizations
- Establish partnerships with three (and potentially four) law enforcement agencies
- Initiate pilot testing of prototype software systems
- Receive feedback and evaluations from law enforcement personnel on the utility and readiness of the software prototypes for designated applications

The ANSER team has investigated and applied several significant technologies in developing components and integrating the prototype systems for demonstration and evaluation. Multiple software-intensive technologies have been explored and incorporated into face recognition, automated Internet searching, automated text understanding, data mining, knowledge discovery, knowledge management, and evolutionary agents. A summary of the results of these component technologies, the prototype systems, the pilot projects, and the impact on law enforcement follows.

9.1 Component Technologies

9.1.1 Face Recognition

At the time the NIJ awarded this cooperative agreement to the ANSER team, no technology with adequate refinement was available for searching image databases constructed by local and Federal law enforcement (including missing children clearinghouses). Potentially relevant technologies lacked real-world robustness and commercial support. Moreover, many of the available facial-recognition methods were mainly embedded in academic laboratories and had to be transferred to commercial applications for public use. The ANSER team made progress in several other areas, including face-recognition performance and automated age-progression algorithms. As a part of this grant, Visionics improved the FaceIt SDK, and Eyematic Interfaces improved Eyematic Primary Library technologies to meet the demanding requirements of fast and accurate searches over large databases. Both companies now have made these technologies commercially available to developers. Face-recognition technologies also have become more robust and available in broader law enforcement applications, including near–real time identification of suspects under video surveillance.

Research is ongoing at Mount Sinai School of Medicine and Rockefeller University to develop automated age-progression algorithms. The ANSER team is also developing a database of more than 500 school-age student images, spanning a time series of at least 4 years. For some individuals, longitudinal sequences of up to 7 consecutive years have been created. A data set of this nature is very difficult to collect and is extremely valuable in the successful continuation of the research.

9.1.2 Internet Search

The ANSER team developed a set of Internet search and retrieval "hounds" for navigating open sources on the Internet and recognizing and collecting interesting image files. Law enforcement solutions can be especially effective when taking advantage of open-source information, particularly the information available on the Internet.

9.1.3 Face Recognition Prototypes for Finding Missing Children—Integrating Internet Search and Face Recognition

Computerized facial recognition is becoming a desirable method for identifying persons. Use of this technology by law enforcement and public agencies that search for missing children promises to provide value and increase the effectiveness of current systems for tracking runaways and abductees. Images of missing children are now routinely digitized and added to databases that can be accessed and shared by regional as well as national agencies. Using photographs, law enforcement agencies can also query databases of missing children and other individuals through the Internet from any location in the world. With computerized facial recognition, these images can be searched reliably and rapidly.

9.1.4 Text Categorization

The ANSER team conducted research on several text-categorization agents with various strategies (including CDM, Vector Space Agent, Bayesian Agent, and Document Clustering Agent) in order to implement and integrate them with Internet search agents. Certain algorithms are known to perform better than others for various categories; hence, overall system performance will improve if several algorithms and strategies are available.

9.1.5 Case Analysis Software Agent

The ANSER team will continue to develop the CASA ISA as both a generic tool and a component in specific applications. In the second NIJ cooperative agreement, CASA will be refined further for two specific law enforcement applications: one at the West Virginia Missing Children Clearinghouse involving information related to missing children, and one at the U.S. Department of the Treasury related to exploited children.

9.2 Prototypes for Law Enforcement Pilot Projects

This NIJ project has developed and integrated prototype software systems with computerized facial recognition and ISAs. These systems either have been or soon will be delivered, installed, tested, and evaluated in law enforcement organizations to solve real-world law enforcement problems, especially those of finding missing and exploited children and supporting drug enforcement investigations.

Capabilities now exist to autonomously search the Internet for information relating to cases of missing and exploited children. Law enforcement agencies can also query databases of missing children using photographic resources.

Facial recognition technology promises to increase the effectiveness of law enforcement investigators in searching for and finding missing children and their abductors and tracking down runaways. Regional and national law enforcement agencies can search and share digitized images of missing children from multiple local databases. With computerized facial recognition, these agencies can search reliably and rapidly for these images.

Law enforcement officers also could use this system in the field by equipping patrol cars with laptop computers or digital terminals, wireless communication equipment, and digital cameras. Law enforcement officers soon will be able to capture and transmit images of suspects back to the station, where the officers can automatically query the booking system or related systems and identify suspects almost immediately.

Without support from this cooperative agreement, it would not have been possible to make such progress in the same amount of time. These tools would not be in the hands of law enforcement personnel, with more tools following in the near future.

10. Future Directions

The results (see Section 9) of this NIJ grant are not only providing significant milestones but also steppingstones to real, automated systems that can be used by law enforcement in finding missing children and combating a host of criminal activity. Information needed for these tasks can be pieced together from the Internet and other open sources as well as various law enforcement structured and unstructured data holdings. These tools will support scalable solutions in automated information recognition and automated decision support.

This section describes the expected transfer of capabilities that have accrued from this effort (including specific technologies, R&D, systems, and lessons learned in the initial pilot projects) and are expected to be selected and integrated under a second NIJ cooperative agreement, 98-LB-VX-K021.

10.1 Technologies

10.1.1 Face Recognition

The methods for comparing still images are being improved for the MCLA prototype and extended into video imaging for a video surveillance prototype—IdentiFace. In addition, age-progression research is being continued. The large yearbook database (including new sets of long chronological or longitudinal data) from Marion County, WV, will be adequate to complete the facial aging research. Additional work is being undertaken on an interpolation method, two Eigenface approaches, and a potentially far more powerful technique called Supersnapshots. Completion, including verification by testing, will be accomplished under a second NIJ cooperative agreement.

10.1.2 Internet Search

The growth of information systems, particularly the Internet, will doubtless continue. Many people worldwide, including criminals, will use this new medium increasingly for communication, commerce, and entertainment. The overwhelming size of this data source poses an incredible challenge and strain on law enforcement resources, but it also offers a tremendous opportunity.

Future work under a second NIJ cooperative agreement will focus on integrating NewsHound, FileHound, and Internet Relay Channel chat room search agents so that NewsGroups, FTP sites, and chat channels can also be searched.

10.1.3 Internet and Database Search

Leveraged properly by means of automated data collection and analysis tools (particularly ISAs), critical information that was never previously available can be put at the fingertips of law enforcement personnel. ISAs can provide key leads related to unsolved crimes.

ISAs can also be deployed to uncover and report criminal activity, such as child pornography trafficking.

10.1.4 Text Categorization

Under the second NIJ cooperative agreement, the ANSER team will continue to integrate text categorizers into systems with more extensive and more useful capabilities, including Internet-based monitoring and investigative systems, such as a new NewsFilter Agent. A prototype iCDM algorithm has been designed and manually tested under the current NIJ effort. However, a fully operational automated iCDM algorithm will be implemented as part of a second NIJ cooperative agreement.

10.1.5 Data Mining

An alpha version of a Knowledge Discovery and Predictive Data Mining system was completed under this initiative, with continued R&D being funded under a second NIJ cooperative agreement.

10.2 Pilot Projects

Over the past 2 years, the ANSER team has successfully developed and demonstrated ISA and face-recognition systems to local, state, and Federal law enforcement organizations. These agencies evaluated the usefulness and readiness of many technologies for law enforcement applications. As a result, several law enforcement agencies will be applying and pilot testing these new systems under a second NIJ cooperative agreement.

10.2.1 West Virginia State Police

The West Virginia Missing Children Clearinghouse, which the state police administer and operate, has begun searching for missing children from West Virginia using MCLA. The pilot has found no matches in the short time it has run, but finding a missing child on the Internet will perhaps become more likely as other states join West Virginia to provide image information in case files, as more content becomes available on the Internet, and as intelligent search strategies emerge.

Under the second NIJ cooperative agreement, CASA is being developed for the West Virginia Missing Children Clearinghouse using SADIE as a development tool.

10.2.2 South Florida High Intensity Drug Trafficking Area

A video surveillance system is being developed and will be improved for pilot testing with the South Florida HIDTA to identify individuals engaged in the drug trade. Similar systems could also be developed to monitor transportation portals for known terrorists or illegal aliens. By integrating the face-recognition components of the IdentiFace system with digital booking systems, a law enforcement officer will soon be able to quickly query the booking system using images of suspects that have been captured on film or

through eyewitness composite sketches. By implementing a Web-based interface, cross-jurisdictional (intrastate) as well as out-of-state organizations can run these queries. This system could also be used in the field by equipping patrol cars with laptop computers or digital terminals, wireless communication equipment, and digital cameras. Using this equipment, law enforcement officers will soon be able to capture images of suspects and transmit these images back to the station, where the officers can automatically query the booking system or other related systems with images. The top image matches would be returned in near–real time to the officer in the field. Thus, the officer could make a positive identification and determine whether any outstanding warrants exist.

These systems can also be extended to provide access monitoring and control functions for physical security systems.

10.2.3 Department of the Treasury and the FBI Innocent Images Program

Both the Department of the Treasury and the FBI Innocent Images Program will be implementing COPIES, which conducts autonomous online investigations, searching for instances of previously confiscated or known child pornographic images. With the availability of large bandwidth, these agents could rapidly search every linked page on the Internet. This system can greatly reduce open exchange of child pornographic materials available via public URLs on the Internet. When COPIES is fully operational in the Department of the Treasury (and potentially in the FBI Innocent Images Program), it is expected to significantly reduce open child exploitation on the Internet and to cause child pornographers to go back underground—achieving a victory for law enforcement.

In the future, it is expected that this pilot project will integrate the following:

- Text-categorization tools for recognizing the context in textual documents
- Threat assessment software tools for application in Internet sites that are disseminating pornographic information (images or text)
- ISA tools that will assist Department of the Treasury investigators with improved automation for researching and analyzing case information (leads, photographs, etc.), including pinpointing Internet service providers and Internet protocol addresses
- A face-recognition tool for identifying victimized children

Several other law enforcement agency applications could be spun off using the underlying technologies of COPIES. Web spidering can be used to locate many digital files, images, video and audio recordings, and software executables disseminated on the Internet. Thus, organizations fighting the piracy of copyrighted materials or the illegal exportation of protected software technologies could use this system, which can be modified to support alternative uses simply by developing a tailored database of digital signatures.

The ANSER team will develop and conduct pilot tests of COPIES with the U.S. Department of the Treasury under a second NIJ cooperative agreement.

Appendix: Glossary

ANSER	Analytic Services Inc.
CASA	Case Analysis Software Agent
CBSSA	Criminal Booking System Search Agent
CD	Compact disk
CDM	Category Discrimination Method
CEO	Chief Executive Officer
CIO	Chief Information Officer
COPIES	Child Online Pornography Image Eradication System
DC	District of Columbia
DNS	Domain Name Service
DSRegistrar	Digital Signature Registrar
EAS	Evolutionary Agent Society
FBI	Federal Bureau of Investigation
FERET	Face-Recognition Technology
FTP	File Transfer Protocol
GUI	Graphical User Interface
HIDTA	High Intensity Drug Trafficking Area
HTML	Hypertext Markup Language
IBL	Instance-based learning
iCDM	Incremental Category Discrimination Method
ISA	Intelligent software agent
KB	Knowledge base
LLNL	Lawrence Livermore National Laboratory
MCLA	Missing Children Locator Agent
NCMEC	National Center for Missing and Exploited Children
NIJ	National Institute of Justice
NNTP	Network News Transfer Protocol
R&D	Research and development

RFC	Request for comment
SADIE	Symbolic Agent Development and Integration Environment
SHA-1	Secure Hashing Algorithm–1 (from the National Institute of Standards and Technology)
SAIC	Science Applications International Corporation
SDK	Software development kit
TDL	Transdimensional learning
URL	Universal Resource Locator