DEFINING A MOTION IMAGERY RESEARCH AND DEVELOPMENT PROGRAM

Organization: National Center for Supercomputing Applications
Automated Learning Group
152 Computing Applications Building, MC-476
605 East Springfield Avenue
Champaign, IL 61820

Technical Point of Contact: Peter Bajcsy, PhD
Phone: 217-256-5387
Fax: 217-265-8022
Email: pbajcsy@ncsa.uiuc.edu
Url: **www.ncsa.uiuc.edu/STI/ALG**

Project Title: White Paper on Metadata Information Generation

Abstract

The proposed goal of this project is to generate metadata information from motion imagery and other supplemental data for content retrieval. In this effort, we address most of the following issues that we believe will be critical in generating metadata from motion imagery in the near future. The critical issues include (1) developing a unified software framework, (2) establishing computational benchmarks for metadata generation, (3) exploring methods for computer learning from annotated imagery, (4) building probabilistic hierarchical organization of metadata, (5) engaging tools for processing heterogeneous multi-modal data sets, (6) generating metadata from compressed imagery and (7) defining a standard format for metadata generation. A proposed project of Seismic Hazard Mapping illustrates how some of the aforementioned issues could be resolved.

1. Introduction

In general, metadata is understood as more comprehensive (more highly organized or specialized form) of information obtained from a sensor, for example, motion imagery sensor. According to [Ref 7], motion imagery is defined as imaging sensor/ systems that generate sequential or continuous streaming images at specified temporal rates (normally expressed as frames per second), within a common field of regard beginning at frame rate 1 frame per second or higher. In this paper, we will focus on metadata generation from (a) motion imagery and (b) from any other supplementary data that bolsters our confidence in extracted metadata information from imagery. We will consider input data to be represented by still images (one frame in a temporal sequence), a temporal sequence

of frames (video or images separated by more than one frame per second), hand-annotated images (a frame with annotations of an image analyst), a set of multi-sensor images (electro-optical, synthetic aperture radar, hyperspectral, multi-spectral, infrared, laser radar), compressed imagery and other supplementary data (e.g., sensor specifications, terrain information, surface geology information, ground observations, weather data, intelligence data).

In terms of application domains for metadata generation, metadata can be used for information browsing and retrieval (digital libraries), data compression (lossy or lossless communication with a finite bandwidth), reconnaissance (change detection in areas of interest), surveillance and monitoring (identification and tracking of entities), or high level data mining (e.g., detecting activities or anomalies for intelligence purposes). There are several industries that are dealing with digital libraries, document archival, telecommunication, data compression, publishing, visual media and audio-visual tools. Among all industries, the National Imagery and Mapping Agency (NIMA), defense departments, intelligence agencies, law enforcement organizations and the national security community have had a growing need for systems that generate metadata from imagery and retrieve information based on its content.

There is a vast amount of literature on many related topics relevant to metadata generation from imagery, for example, publications about content-based retrieval from video, digital libraries, document analysis, video tracking, multi-sensor fusion and video compression. A list of most recent publications can be obtained from (a) conference proceedings (e.g., Message Understanding Conference (MUS), Text Retrieval Conference (TREC), ACM Special Interest Group on Information Retrieval (SIGIR) Conference), and (b) active and past programs and projects (e.g., TIPSTER Text Program sponsored by DARPA and NIST; VisualSEEk and VideoQ at Columbia University [Ref 3]; Informedia Digital Video Library at CMU; LAMP at University of Maryland, Video Analysis and Content Extraction (VACE) by ARDA; National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) by NSF [Ref 1]; ViPER Video Performance Evaluation Resource at NIST; and projects by American Society for Information Science (ASIS)). There are also available documents about video standards (e.g., MPEG-7: "Multimedia Content Description Interface" [Ref 4], Motion Imagery Standards Profile [Ref 7]) and limited information about commercial products Virage (Virage Inc., San Mateo, CA) [Ref 9] and Google (Google Inc., Mountain View, CA).

In this paper, we will address several issues related to research and development of a set of general-purpose and domain-specific analysis tools for generating information metadata from imagery. The issues and our motivation to address these issues are summarized next. We believe that the following issues will be critical in generating metadata from motion imagery in future: (1) unified software framework, (2) computational resources for metadata generation, (3) learning from annotated imagery, (4) probabilistic hierarchical organization of metadata, (5) processing heterogeneous multi-modal data sets, (6) metadata generation from compressed imagery and (7) standard formats for metadata representation. Next, we overview the above issues and present a proposed project of Seismic Hazard Mapping. The proposed project is described to illustrate how some of the aforementioned issues could be resolved.

2. Unified Software Framework

First, it is well understood that indexing motion imagery data collections often requires domain-specific code either to extract metadata from complex formats, or to extract features from the data itself using analytical codes and represent those features as metadata. Specifically, the national security community deals with a variety of imagery (multi-sensor imagery and video data), large amounts of data (terra bytes per sensor modality) and data with little internal structure (outdoor scenes), which require domain specific code to extract metadata. Thus, there is a need for a software framework that would enable sharing common metadata extraction algorithms, as well as using domain specific extraction methods.

NCSA Automated Learning Group (ALG) has developed a visual programming environment called Data To Knowledge (D2K) [Ref 11] allowing users to easily connect software modules together in "itineraries" analogous to the flow of data through an analysis process (see Figure 1). The ALG has been developing a suite of tools for image analysis called Image To Knowledge (I2K) inside of the D2K environment, including tools for automated and semi-automated processing of data [Ref 12] from agricultural engineering, cartography, microscopy, clinical medicine, genomics and life sciences. The tool set addresses the problems of camera noise, image calibration, image registration, computational reduction [Ref 17], unsupervised classification [Ref 15], [Ref 16], and statistical image analysis and synthesis [Ref 14]. In addition, the tools offer basic visualization capabilities for two-dimensional multivariate imagery (multi-band imagery), motion imagery and for resulting label imagery obtained from classification (see Figure 1 and I2K documentation [Ref 12]). As part of the visualization, numerical results can be visualized using a scatter plot tool or a text area dialog. We continue developing generic and domain specific tools that can be shared by users from multiple application domains.

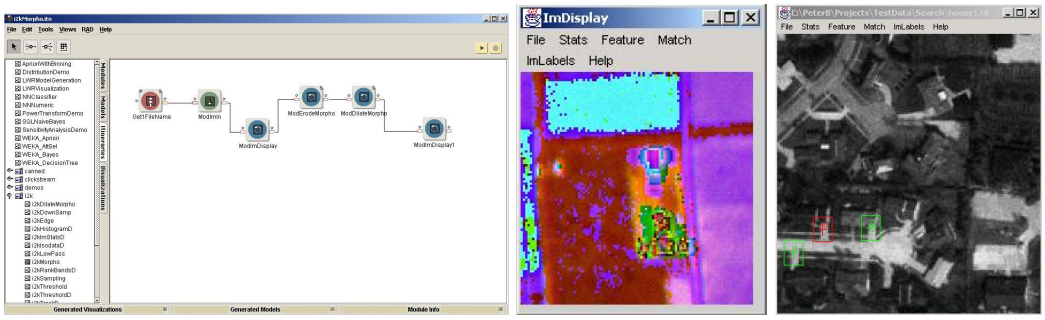

Figure 1: D2K visual programming environment (left) and visualization of hyperspectral imagery (middle) and results of data mining (right).

3. Computational Resources For Metadata Generation

Computational requirements of motion imagery processing are enormous. For example, one color camera would record about 2.6 terra-bytes of data in 24 hours and thus any efficient information retrieval requires a succinct representation of the data. It is very desirable to experiment and obtain computational benchmarks for processing using supercomputers. The need for computational benchmarks comes from the fact that metadata generation can be used not only for archival analysis but also for real-time

screening of high-throughput data streams. Thus, the time and memory requirements become critical in selecting appropriate extraction methods.

The NCSA supercomputing facility will provide the computational resources for establishing benchmarks for processing large and small amounts of imagery with methods of multiple degrees of computational complexity. The NCSA institution maintains a broad array of high-performance computing, storage, and communication systems. The NCSA computing and storage environment consists of four primary components: (1) Two teraflop of Linux clusters, integrated by IBM. (2) The Silicon Graphics Origin2000 (1,520 processors and 618 GB RAM). (3) NCSA's Supercluster, running both Linux and Microsoft Windows NT. (4) A 250 TB storage archive, powered by an eight-processor Origin2000 running UniTreeCFM 2.1.

4. Learning From Image Annotations

Although image annotation has been a part of image analysis, annotations have not been used for learning and future automatic and semi-automatic metadata generation. When the problem of automatic metadata generation seems very difficult and thus the extracted metadata is of little confidence, supervised methods for metadata generations are appropriate. A prior knowledge about features in imagery can be learned by supervised methods. The knowledge of image analysts can be represented by a set of computer generated or hand drawn symbols (see Figure 2) that represent the image annotations and are overlaid on the original image. Techniques and tools for automated and semi-automated feature extraction and computer-assisted metadata generation from annotated imagery can be widely used in all application domains.

We will develop a generic set of techniques and tools for extracting features from a variety of annotated imagery in several application domains. The extracted features will be mined for appropriate models with supervised and unsupervised approaches. The developed models and the extracted features will provide a hierarchical organization of the extracted information that will be converted into metadata.. In the scope of this problem, we will address the following three fundamental computational problems occurring across multiple application domains: (1) analysis of image annotations (2) supervised development of a nonlinear classification model and (3) unsupervised evolution of a generative classification model. The details are presented in Section 9.

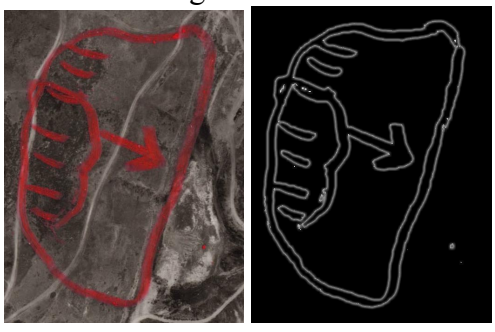

Figure 2: An example of annotated aerial photography (left) and the contours of the annotation to be analyzed (right).

5. Probabilistic Hierarchical Organization of Metadata

In order to perform intelligent and efficient information retrieval, it is desirable to create probabilistic hierarchical organization of metadata. The hierarchy of metadata reflects the level of details that one might be interest in. For instance, searching for a white car or for a moving convoy of vehicles requires analyzing motion imagery at multiple levels. The novel aspect of the metadata generation is its probabilistic framework. Every extracted feature has its own associated probability of detection. The probability will be part of metadata. The probability of feature detection will be combined with the probability of a feature matching user's query match. Thus the probabilistic information in metadata will serve during information retrieval to improve standard recall and precision measures of any browsing and retrieval system.

We will extract metadata information at multiple hierarchical levels such as, a level of each video frame (entities and their spatial and photometric attributes), a level described by a common event (entities and their temporal and spatial changes) and a level described by a common theme (a set of spatio-temporal changes of entities and their attributes that form a pattern defined by theme or scenario). Extracted information at the frame level includes (a) global and local statistics of color, e.g., parametric probability distribution functions, (b) textural properties, e.g., co-occurrence features, (c) geometrical primitives, e.g., edges, corners and lines, and (d) aggregations of geometrical primitives that infer man-made objects, e.g., parallel lines, dashed lines and symmetrical shapes. At the event level we focus on the motion of frame features that show persistence over a large set of consecutive frames (see Figure 3). The motion information will be derived based on an optical flow analysis combined with a correlation-based tracking. At the theme level we use clustering of events and objects to aggregate spatially and temporally related video pieces. Furthermore, generated metadata will contain probabilities of its entries derived during feature extraction.

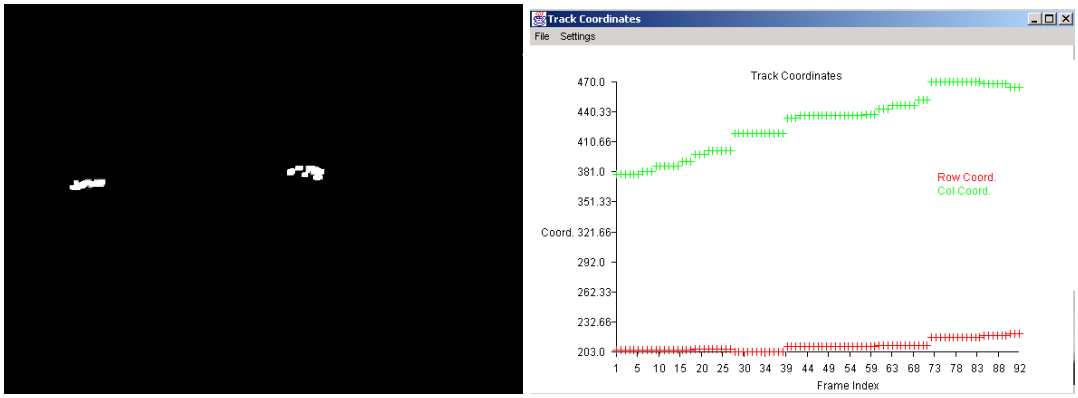

Figure 3: Example of analysis at the event level. Top row shows two frames from a video sequence with two moving vehicles entering a parking lot. Bottom row shows detected vehicles as white blobs (left) and the plot of tracked pixel coordinates of the white pickup truck (the vehicle on the right side) as a function of frame index (time).

6. Processing Heterogeneous Multi-Modal Data Sets

In many cases of metadata generation, multiple data sources should be considered simultaneously. The commonality of multiple data sources is usually in the information content describing the same event in space and time. However, the differences among these data sets come from (a) sensor modalities (e.g., infrared, electro-optical, radar, multi-spectral, hyperspectral), (b) data resolution (e.g., high-resolution electro-optical imagery and low-resolution synthetic aperture radar imagery), (c) arrangement of measurements (e.g., regular grid in images and irregular grid of weather measurements), (d) acquisition geometry, (e) time of data collection, (f) number of attributes of each datum, (g) dynamic range of attribute values,, and so on. In general, processing heterogeneous multi-modal data sets requires several pre-processing steps, such as sensor distortion correction, calibration, geo-registration, before metadata can be extracted from multiple data sources. After pre-processing, each sensor modality should be evaluated for its strengths and weaknesses so that the confidence in extracted information from the fused data can be established accordingly.

The NCSA ALG has been developing a suite of tools for the problems of camera noise, image calibration, image registration, computational reduction, unsupervised classification, and statistical image analysis and synthesis [Ref 12]. These tools have been applied to satellite and hyperspectral imagery [Ref 17]. We will continue developing algorithms for processing heterogeneous multi-modal data depending on any supplementary data available for supporting metadata generation from motion imagery.

7. Metadata Generation From Compressed Imagery

It is apparent that the volume of motion imagery is enormous and thus compression algorithms are a necessity. As a consequence, extracting metadata from compressed imagery can significantly reduce computational resources (memory and time) for metadata generation. The core of current compression standards for still and moving imagery is the discrete cosine transform (DCT). DCT is used by JPEG, MPEG, H.261

and HDTV standards. In general, the algorithms for metadata extraction have to recover low level image features from discrete cosine transform coefficients without recovering explicit pixel values. Preliminary experiments found in literature [Ref 21], [Ref 22] report about 20 times faster edge information recovery and 5 times faster object recognition.

Although this problem is very important to the metadata generation process, we will not explore the methods in a great detail under this project.

8. Metadata Format Standards

Metadata is extracted from a variety of data sources and for a large number of applications. Each application defines its own dictionary of desired information to be extracted from motion imagery. As of today, applications driven by commercial industry have defined industrial standards (e.g., Resource Description Framework (RDF) [Ref 2], Extensible Markup Language (XML), Joint Photographic Expert's Group (JPEG-2000) [Ref 6], Society of Motion Picture and Television Engineers standards [Ref 8] or Moving Picture Experts Group (MPEG-7) [Ref 4] with Description Definition Language (DDL)). Applications driven by national security (Department of Defense/Intelligence Community/United States Imagery and Geospatial Information Service (DoD/IC/USIGS)) have defined government standards (Core Motion Imagery Metadata Format, National Imagery Transmission Format (NITF) specified by Motion Imagery Standard Board (MISB)). There is a significant overlap in both sets of standards and we foresee the format evolution driven by currently used metadata descriptions and by the state-of-the-art metadata extraction methods applied in application domains.

NCSA has been actively participating in development of metadata interoperability standards for use with scientific data collections on the Grid [Ref 10]. Under this effort, we do not intend to work on metadata standards although we might provide suggestions based on the algorithmic work of metadata generation.

9. Experimental Project

We identified a project that would cover some of the aforementioned issues. The project, Seismic Hazard Mapping, focuses on metadata generation from hand annotated aerial imagery and supplementary data of digital elevations and surface geology. The historical data with and without annotations will provide information about temporal changes of hazard zones that will be captured by metadata.

Losses from Earthquake damage have prompted state governments to develop zoning legislation based on estimates of seismic hazard risks [Ref 18]. These estimates are developed by geotechnical experts on the basis of a variety of historical data, including boring logs, ground measurements, digital elevation maps (DEM), and aerial and satellite photography. By integrating all of this data, geotechnical experts produce a variety of products, the most significant of which are geo-referenced inventories of hazards such as landslides and faults. As the pace of land development and the availability of new sources of digital data increase, it becomes increasingly difficult to manage the heterogeneous data required to develop and maintain these inventories. Furthermore, the users of this data are diverse, including engineers, legislators, and insurance companies, all of which have different information needs. Our system will demonstrate that by

learning from annotated imagery at a few time instances, the extracted metadata from incoming imagery can reliably assist in detection of landslide and fault hazards.

Figure 4: Left: aerial photograph of region. Right: same photograph, annotated to show locations and characteristics of landslide hazards. Annotations like these, which are currently done manually by hazard mapping agencies, will be used as a training set for our machine-learning-based metadata extraction system. (Source: California Department of Conservation, Division of Mines and Geology).

To address the detection of landslide and earthquake fault zones, we will focus on three technical approaches: (1) analysis of hand-drawn image annotations, (2) supervised development of a nonlinear classification model, (3) unsupervised evolution of a generative classification model.

During the course of investigating the three scientific problems, we will develop generic tools for fusing photography imagery with digital elevation maps and surface geology maps; extracting hand-drawn image annotations and recognizing annotation symbols, extracting training data based on annotation symbols; classifying unknown heterogeneous data with a supervised classification model using dynamic Bayesian networks [Ref 20]; identifying class rarely distinguished humans with an unsupervised transformed hidden Markov model [Ref 19]; and generating structured metadata from the results of the classification. The proposed tools will be a part of the metadata generation of hazard zones according to the flow diagram in Figure 5. The functionality of each module/ block in the flow diagram is summarized in Table 1. Table 1 also presents scientific issues that will be novel to our research and development.

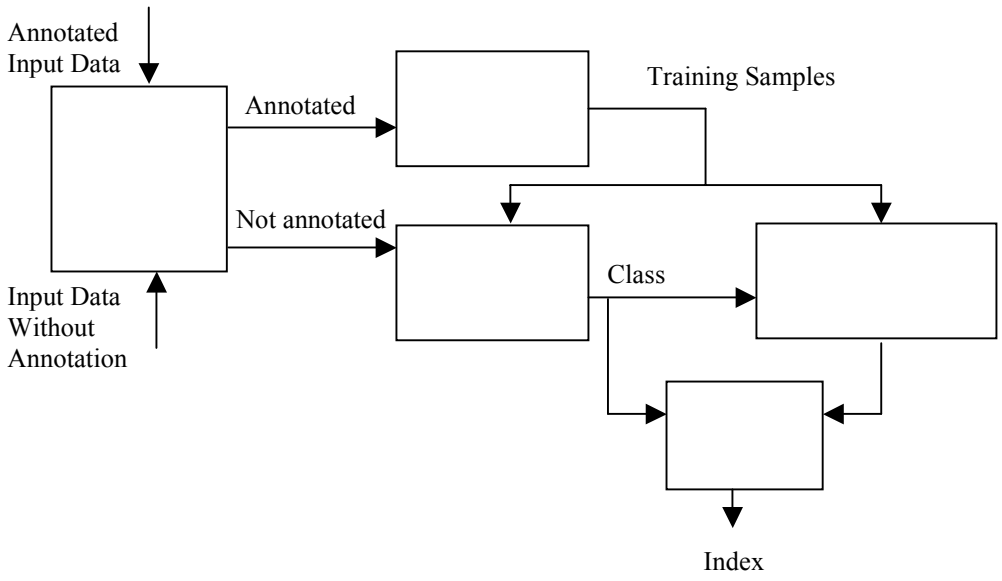

Figure 5: Dataflow for feature-based metadata extraction from heterogeneous scientific datasets.

Table 1: Description of the modules in Figure 5.

Processing Module	Description of Functionality	Scientific Issues
Data fusion	Registration, interpolation	Interpolate given *a priori* knowledge (e.g. known plate tectonics)
Sample extraction and annotation analysis	Detection, connectivity analysis, skeletonization, feature selection, classification into symbols, extraction of training samples based on symbols	Characterize symbols, evaluate robustness of classification with respect to variations of symbols in annotations
Supervised classification	Build supervised classification model using dynamic Bayesian network model, classify un-annotated data based on the model	Integrate multiple data sources and infer complex variable dependencies, incorporate feature variations (position, scale and orientation) into the model, select similarity metric
Unsupervised classification	Build transformed hidden Markov model for sub-classification, sub-classify un-annotated data based on the model	Incorporate *a priori* information (e.g. DEM and surface geology) into the model, integrate multiple data sources and incorporate feature variations (position, scale and orientation) into the model, select similarity metric
Metadata extraction	Extract and organize information into structured metadata records	Describe classes and subclasses by metadata (e.g., keywords, spatial and temporal information,

		orientation and scale) associated with its likelihood values

The critical issues of metadata generation will be addressed in the following way. First, the software development will be use the Data To Knowledge (D2K) visual programming environment and it will be an add-on to the current suite of tools called Image To Knowledge (I2K). Second, the NCSA supercomputers will be used for benchmarking metadata extraction algorithms to evaluate memory and speed requirements of algorithms for a given application. Third, the analysis of hand of hand-drawn image annotations will be a part of supervised learning of a nonlinear classification model and also a part of unsupervised evolution of a generative classification model. Fourth, while each image will be described by its color, texture and a set of geometrical primitives, a set of historical data sets will be analyzed for its events (changes in landslide geometry over time) and themes (co-occurrence of several landslide events in space and time). Fifth, heterogeneous data sets (imagery, digital elevation maps and surface geology) will be pre-processed in order to maximize the resulting confidence in generated metadata. The issues of metadata generation from compressed imagery and metadata format standards will not be addressed.

10. Conclusions

We have examined several critical issues of metadata generation including (1) a unified software framework, (2) computational resources for metadata generation, (3) computer learning from annotated imagery, (4) probabilistic hierarchical organization of metadata, (5) processing heterogeneous multi-modal data sets, (6) generating metadata from compressed imagery and (7) defining a standard format for metadata generation. The proposed project of Seismic Hazard Mapping was described to illustrate how some of the aforementioned issues of metadata generation would be resolved in a concrete application.

Our proposed work will be built on our previous work with information extraction, and data mining in NCSA's Data To Knowledge (D2K) effort. D2K software will provide a framework of metadata extraction components for handling broad classes of motion imagery data objects. To adapt this framework to particular domains, a "plug-in" architecture will be used, minimizing the redundant effort required to integrate new types of data into the system. The development of techniques and tools for automated and semi-automated feature extraction and computer-assisted metadata generation will be a collaborative effort of the NCSA ALG and the Image Formation and Processing Research Group (IFP) at the University of Illinois.

We believe that hand-drawn annotations of the sort produced by geotechnical experts, or image analysts in general, are extremely valuable sources of domain knowledge, since they correlate particular image regions with expert judgments about what domain features those regions represent. Automatically interpreting hand-drawn annotations is difficult, however, because it requires disambiguating individual variations among annotators as well as recognizing symbols in all their positional, rotational, scale- and shear-related variations. Nonetheless, the tools developed for learning from annotated aerial photography can be easily used on motion imagery.

We also proposed to extract metadata information at multiple hierarchical levels such as, a level of each video frame, a level described by a common event and a level described by a common theme. This effort lays down the groundwork for probabilistic hierarchical metadata organization that will be the final product of metadata generation process after all pre-processing, benchmarking, learning, recognition and metadata creation stages are completed.

References

[Ref 1] "National Science, Math, Engineering and Technical Education Digital Library (NSDL) Program," National Science Foundation, http://www.dli2.nsf.gov/

[Ref 2] "Resource Description Framework (RDF)," World Wide Web Consortium, http://www.w3.org/RDF/

[Ref 3] "VisualSEEk," http://queen.sungshin.ac.kr/~hkim/VisualSEEk/manual.html.

[Ref 4] "MPEG-7," http://mpeg.telecomitalialab.com/

[Ref 5] "Universal Description, Discovery and Integration (UDDI) project ," http://www.uddi.org/about.html

[Ref 6] "JPEG 2000," http://www.jpeg.org/JPEG2000.htm

[Ref 7] "Motion Imagery Standards Profiles," Department of Defense/Intelligence Community/ United States Imagery and Geospatial Information Service (DoD/IC/USIGS), Version 1.7, March 2001, http://164.214.2.51/vwg/announce/MISP_17_010301.pdf

[Ref 8] "Society of Motion Picture and Television Engineers, " http://www.smpte.org/smpte_store/standards/

[Ref 9] "Virage, Inc.," http://www.virage.com/

[Ref 10] "Alliance Metadata Standards Working Group," http://metadata.ncsa.uiuc.edu/.

[Ref 11] "Data To Knowledge (D2K)," NCSA ALG Projects, **http://www.ncsa.uiuc.edu/Divisions/DMV/ALG/activities/index.html**

[Ref 12] "Image To Knowledge (I2K)," Software Overview and Documentation, http://www.ncsa.uiuc.edu/Divisions/DMV/ALG/activities/projects/i2k/documentation/index.html

[Ref 13] Bajcsy, P. and Ahuja, N. "Hierarchical Clustering of Points Using Similarity Analysis" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, NO. 9, pp. 1011-1015,September 1998.

[Ref 14] Bajcsy, P., Sullivan, D. and Ryan, P. "Estimating Weibull Distribution Parameters," *Intelligent Engineering Systems Through Artificial Neural Networks*, Editors C. Dagli et al., ASME Press, New York, vol. 10, pp. 677-682, November 2000.

[Ref 15] Bajcsy, P. and Ahuja, N. "Hierarchical Texture Segmentation Using Dictionaries", ACCV '98, Hong Kong, January 1998.

[Ref 16] Bajcsy, P. and Ahuja, N. "A New Framework for Hierarchical Segmentation Using Homogeneity Analysis", First International Conference on Scale-Space Theory in Computer Vision, the Netherlands, July 1997.

[Ref 17] Bajwa, S. G., Bajcsy, P., Groves, P., and Tian, T., "Methods for Hyperspectral Band Selection Applied to Precision Farming," Submitted to *IEEE transactions on Geoscience and Remote Sensing, 2001*.

[Ref 18] Real, C. R. "California's Seismic Hazards Mapping Act: Geoscience and Public Policy". California Department of Conservation, Division of Mines and Geology, Sacramento, CA, 2001.

[Ref 19] Jojic N., Petrovic N., Frey B. and Huang T., "Transformed Hidden Markov Models: Estimating Mixture Models and Inferring Spatial Transformations in Video Sequences", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Hilton Head Island, South Carolina, 2000.

[Ref 20] Garg A., Pavlovic V., Rehg J., Huang, T. S.,"Audio-Visual Speaker Detection using Dynamic Bayesian Networks", accepted for publication at FG'2000.

[Ref 21] Shen, B. and Sethi, I. K., "Direct Feature Extraction From Compressed Images," SPIE Vol. 2670, Storage & Retrieval for Image and Video Databases IV, pp. 1-12, 1996.

[Ref 22] Seales, W. B., Yuan, C. J., Cutts, M. D., "Object Recognition in Compressed Imagery," Image and Vision Computing, Vol. 16, pp. 337-352, 1998.