ROHINI K. SRIHARI

Using Speech Input for Image Interpretation, Annotation, and Retrieval*

This research explores the interaction of textual and photographic information in an integrated text/image database environment. Specifically, three different applications involving the exploitation of linguistic context in vision are presented. Linguistic context is qualitative in nature and is obtained dynamically. By understanding text accompanying images or video, we are able to extract information useful in retrieving the picture and directing an image interpretation system to identify relevant objects (e.g., faces) in the picture. The latter constitutes a powerful technique for automatically indexing images.

A multistage system, PICTION, which uses captions to identify human faces in an accompanying photograph, has been developed. We discuss the use of PICTION's output in content-based retrieval of images to satisfy focus of attention in queries. The design and implementation of a system called Show&Tell*—a multimedia system for semi-automated image annotation—is discussed. This system, which combines advances in speech recognition, natural language processing (NLP), and image understanding (IU), is designed to assist in image annotation and to enhance image retrieval capabilities. An extension of this work to video annotation and retrieval is also presented.*

INTRODUCTION

This discussion explores the interaction of textual and photographic information in an integrated text/image database environment. In designing a pictorial database system, some of the major issues to be addressed are the amount and type of processing required when inserting new pictures into the database and efficient retrieval schemes for query processing. Searching captions for keywords and names will not necessarily yield the correct information, as objects mentioned in the caption are not always in the picture, and often objects in the picture are not explicitly mentioned in the caption. Performing a visual search for objects of interest (e.g., faces) at query time would be computationally too expensive, not to mention time-consuming. It is clear that selective processing of the text and picture at data entry time is required.

Whereas most techniques for automatic content-based retrieval of images and video have focused exclusively on statistical classification

* This work was supported in part by ARPA Contract 93-F148900-000.

techniques, the approach presented here is based on object recognition. By exploiting multimodal content accompanying an image or video, object recognition (which otherwise would be considered a futile approach) is enabled. By integrating robust statistical techniques with object-recognition techniques (where possible), one obtains true semantic content-based retrieval.

The need for exploiting domain-specific and scene-specific context in vision has been acknowledged; Strat and Fischler (1991) discuss such an approach. The core research here has centered on the use of linguistic context in image understanding. There are several issues which make this problem interesting and challenging on both the image understanding and natural language processing fronts. First, information obtained from language is qualitative. Since most vision algorithms rely on quantitative information (e.g., geometric site models), considerable effort has been expended in designing (or redesigning) vision algorithms to exploit qualitative context. Second, the task of extracting useful information from language and converting it to a suitable representation for use by an image understanding system has also been investigated. Finally, the design of a robust system combining image understanding and natural language processing within the framework of a convenient multimedia user interface has posed the greatest challenge.

Significant progress has been made in the design of a system which exploits linguistic context in the task of image interpretation. A theoretical model for using natural language text as collateral information in image understanding has been formulated; collateral information has been represented as a set of *constraints*. Preliminary natural language processing techniques for extracting collateral information from text have been formulated. A control stucture has been designed which efficiently exploits the above constraints in the process of image interpretation.

Finally, this research has led to three prototype systems: (1) PICTION (Srihari et al., 1994; Srihari, 1995c; Srihari & Burhans, 1994) which, when given a captioned newspaper photograph, identifies human faces in the photograph; (2) *Show&Tell*, a semi-automated system for image annotation and retrieval; and (3) *MMVAR*, a Multimedia system for Video Annotation and Retrieval. It should be noted that the last system is still under construction. This discussion presents an overview of all three systems with an emphasis on their use in content-based retrieval.

PICTION: A CAPTION-BASED FACE IDENTIFICATION SYSTEM

PICTION (Srihari, 1995c) is a system that identifies human faces in newspaper photographs based on information contained in the associated caption. More specifically, when given a text file corresponding to a

newspaper caption and a digitized version of the associated photograph, the system is able to locate, label, and give information about people mentioned in the caption. PICTION is noteworthy since it provides a computationally less expensive alternative to traditional methods of face recognition in situations where pictures are accompanied by descriptive text. Traditional methods employ model-matching techniques and thus require that face models be present for all people to be identified by the system; our system does not require this. Furthermore, most current face recognition systems (Chellappa et al., 1995) use "mugshots" (posed pictures) of people as input; due to standardized location and homogeneous scale, detection of facial features is facilitated. Recognizing faces which have been automatically segmented out of an image is a much more difficult problem.

A significant amount of work in face location and gender discrimination has been developed and employed in the above mentioned system. PICTION has served as the basis for research in content-based retrieval from image databases (Srihari, 1995a)—retrieval is based on information from the image (face identification) as well as information gleaned from the caption.

We now discuss our work on content-based retrieval of images which takes into account both the information content from the caption as well as the information content from the picture. There are four distinct sources of information which we have identified in computing the similarity between a query and a captioned image. These are:

- text-based objective term similarity (exact match);
- text-based content term similarity (inexact match);
- image-based objective term similarity (exact match); and
- image similarity (inexact match).

Text-based objective terms include manually assigned keywords or other *keys* which have been assigned values using manual techniques. Examples of such keys are: (1) names of people in the picture, (2) who (or what) is actually depicted in the picture (not necessarily the names of the people as in item 1), (3) the event type, (4) the location, (5) the time, (6) the general mood of the picture (happy, somber, serious, etc.). More recently, Chakravarthy (1994) discusses methods of automatically assigning values to such keys. Although it is possible to derive values for some predefined keys, other robust methods of measuring content-term similarity between a query and captioned image should also be considered. The availability of large-scale lexical resources such as machine-readable dictionaries and WordNet enable such methods. For example, for each content word W_q in the query, one could count the number of words in a caption with the same context by following pointers from W_q; each pointer

represents a different type of relationship. The scores are weighted by the distance (path length) from the original word. Other methods of capturing context include computing dictionary definition overlap.

Any positive object/people identification made by PICTION is represented in the database by the image coordinates. Similarly, any characteristic information that has been visually verified (e.g., gender or color of hair) is also noted. Image-based information useful in determining the presence of an individual can be quantified based on: (1) whether the face was identified; (2) the size, orientation, etc. of the face; and (3) the method used to identify faces. The last measure of similarity concerns purely image-based techniques which have been discussed extensively in the image processing literature. Examples of such measures include texture similarity.

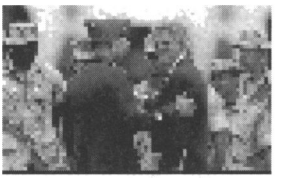
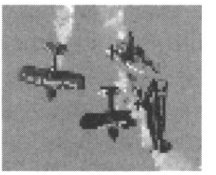
Based on the above measures of similarity, we can compute a combined similarity measure between a query and a captioned image as:

$$Sim(CapImage, Query) = \hat{\alpha} \{text\ objective\_term\ similarity\} + \beta \{text\ content\_term\ similarity\} + \gamma \{image\ objective\_term\ similarity\} + \delta \{image\ similarity\}$$

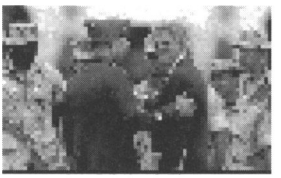
We are in the process of empirically attempting the values for $\hat{\alpha}$, ß, γ, and δ. Intuitively, we can see that higher emphasis should be placed on the exact match components, especially the image-based exact match component. In the next section, we describe the dynamic assignment of weights in order to satisfy the focus of attention in users queries. The order of presentation of images to the user will depend on the value of the above metric.

Dynamic Satisfaction of Emphasis in Image Retrieval

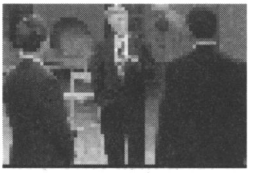
We have performed experiments where text and image information are dynamically combined to best satisfy a query. In such cases, a user specifies not only the context of the picture he is seeking but also indicates whether the emphasis should be on image contents or text content. An example of this is illustrated in Figure 1. This illustrates the top two "hits" on the following queries: (1) *find pictures of military personnel with Clinton*, (2) *find pictures of Clinton with military personnel*, and (3) *find pictures of Clinton.* In the first query, the emphasis is more on satisfying the context of "military personnel." In the second, the emphasis is more on Clinton, and in the final query, there is no context provided; therefore in the last query, the user presumably is only seeking good pictures of Clinton. We have already described methods of computing contextual similarity based on the text. To measure how well the picture contents satisfy the query, we have considered the following factors: (1) whether the required

Left: President Clinton, right, talks with Colin Powell, left, during a ceremony at the White House marking the return of soldiers from Somalia on May 4.
Right: Four aircraft performing daredevil stunts on U.S. Armed Forces Day open house. President Bill Clinton took part in the celebrations and gave away awards to the best Cadets from the U.S. military and armed forces.




Left: President Clinton, right, talks with Colin Powell, left, during a ceremony at the White House marking the return of soldiers from Somalia on May 4.
Right: President Bill Clinton and Vice-President Al Gore walk back to the White House after they welcomed back U.S. troops returning from Somalia at the White House.

Left: President Bill Clinton gives a speech to a group of eleventh graders at Lincoln High School on his visit there April 2.
Right: President Bill Clinton, center, responds to questions put forth by interrogators.

Figure 1. Top 2 (left, right) responses to the following three queries. Row 1: Find pictures of military personnel with Clinton; Row 2: Find Pictures of Clinton with military personnel; Row 3: Find pictures of Clinton.

face was actually identified (by PICTION), (2) the size and orientation of the face, and (3) the centrality of the face in the image. The last factor is given a very low weight compared to the first two.

As the results illustrate, the first query is weighted more toward similarity of text context; notice that the second hit does not even contain any people, let alone Clinton. The words "Armed Forces" (which are part of

a larger title) caused a strong contextual match; we are attempting to refine our measures of context to overcome such problems. The second query results in pictures with Clinton for the most part; the picture with the airplanes is ranked very low. The last query produces the best pictures of Clinton with disregard for context.

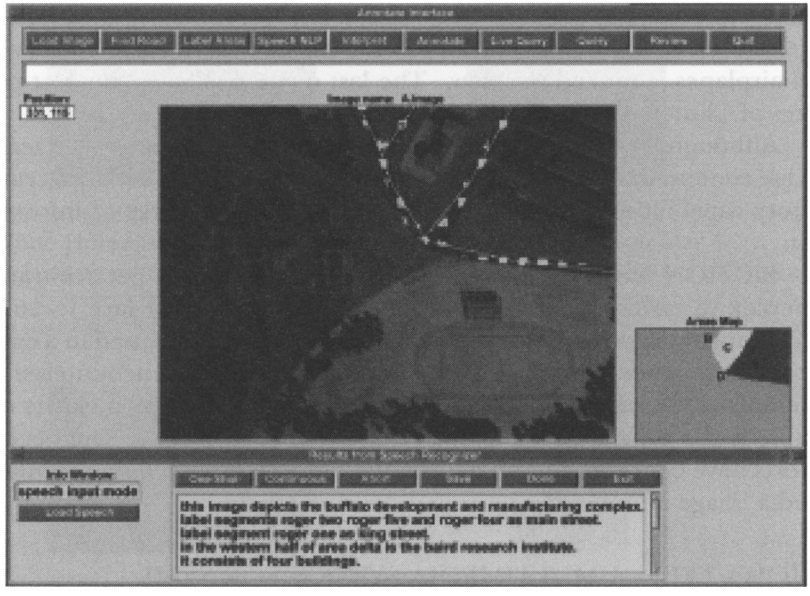
Although this is only a preliminary foray into a truly integrated text/image content-based retrieval system, it illustrates the additional discriminatory capabilities obtained by combining the two sources of information.

PICTION has its limitations since: (1) captions can get arbitrarily complex in terms of extracting reliable information, (2) face location and characterization tools are not sufficiently robust to be used in a completely automated system, and (3) there are limitations encountered in attempts to derive 3D information from a static 2D image; a classic example is the processing of spatial relations such as *behind*. Our recent efforts have concentrated on developing an interactive system for multimedia image annotation.

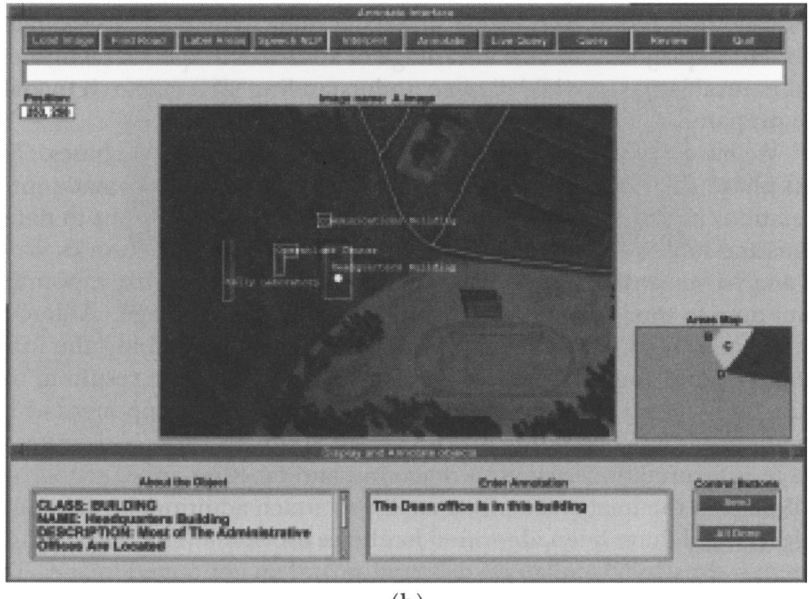
SHOW&TELL: A MULTIMEDIA SYSTEM FOR SEMI-AUTOMATED IMAGE ANNOTATION

Show&Tell is a system which combines speech recognition, natural language processing, and image understanding. The goal of this image annotation project is to take advantage of advances in speech technology and natural language (NL)/image understanding (IU) research to make the preparation of image-related collateral more efficient.

We have developed a prototype system consisting of two phases. The first phase (illustrated in figures 1a and 1b) consists of automatic interpretation/indexing of images. It begins by using mouse input to detect roads and subsequently partition the image based on road networks. Next, an analyst views the image and describes it in spoken language, pointing from time to time to indicate objects or regions in the image. A state-of-the-art speech recognition system is employed in transcribing the input and synchronizing the speech with the mouse input. The resulting narrative is processed by a natural language understanding component which generates visual constraints on the image. This in turn, is used by an image interpretation system in detecting and labeling areas, roads, and buildings in the image. Finally, a facility to attach additional collateral to objects which have been identified has been provided. The output of the system is thus an NL collateral description and an annotated image. The annotated image consists of: (1) a semantic representation of the text, and (2) locations of objects or regions identified in the image. Information is represented such that spatial reasoning, temporal reasoning, and other contextual reasoning is enabled.



(a)



(b)

Figure 2. (a) result of road detection; (b) results of image interpretation

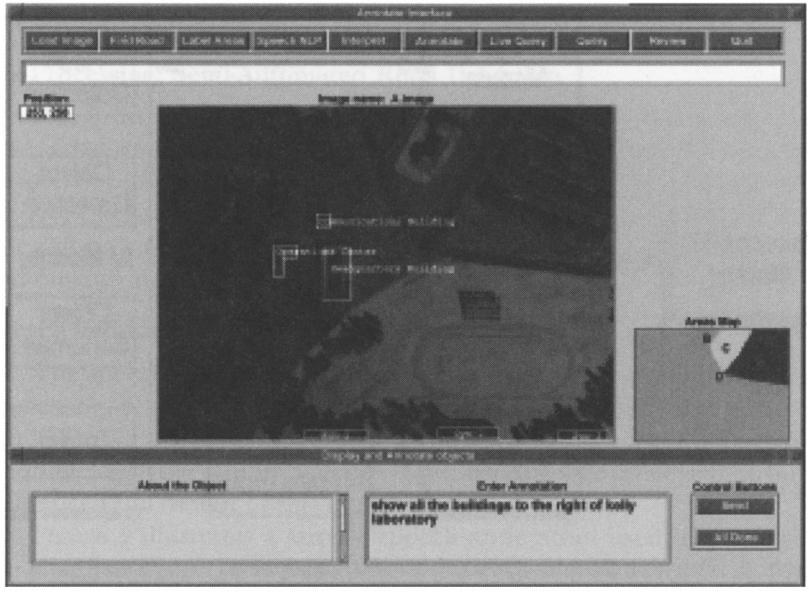


Figure 3. Results of querying.

In the second phase, we provide point and click querying synchronized with speech on the annotated images. For example, the query *Show all man-made structures to the west of this <click> forest* would cause the appropriate areas in the image to be highlighted; each of these could be further queried for corresponding textual information. This is illustrated in figure 2.

There are several assumptions we are making in specifying the task and our proposed solution. From the perspective of an image analyst, this approach constitutes a healthy compromise between: (1) tedious manual annotation, even when tools such as *snakes* are provided, and (2) completely automated (image-based) interpretation. Since the task involves *co-referencing* image areas with textual descriptions, our system uses the text for dual purposes: co-referencing as well as for assisting image interpretation.

The second assumption concerns the use of preconstructed geometric site models. These have been used effectively in the RADIUS community for registering new images of a known site and subsequently for change detection. The initial version of *Show&Tell* does not use site models since the objective was investigation of linguistic context alone. The version in development takes a different approach by utilizing site models.

Finally, at this point we assume that an approximate shape representation for objects is sufficient for reasoning purposes. We represent objects

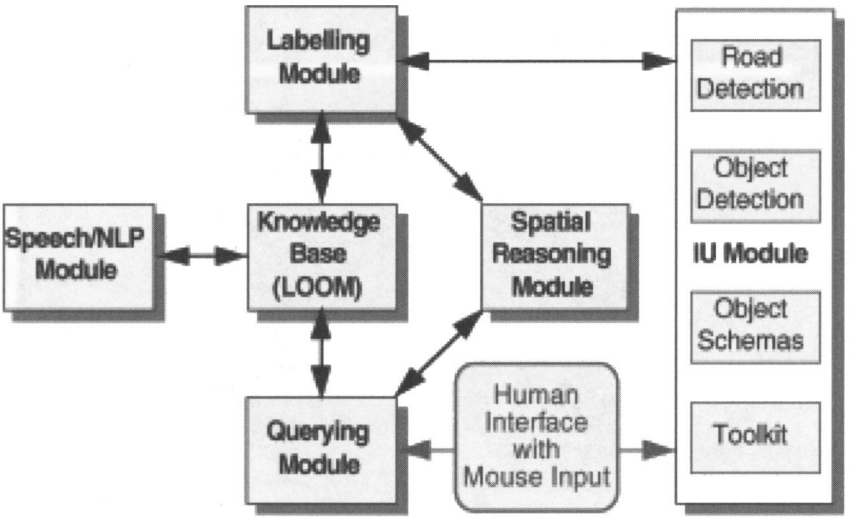


Figure 4. Functional architecture of *Show & Tell* System.

such as buildings or areas by simple bounding boxes with orientation; polylines are used for roads, rivers. Our system does allow for more exact representations for display purposes.

Figure 4 illustrates the functional architecture of *Show&Tell.* Techniques from several subdisciplines of artificial intelligence have been assimilated into this system, including computer vision, natural language understanding, spatial reasoning, knowledge representation, and information retrieval. The system is implemented in C and LISP and uses LOOM (ISI/USC) for knowledge representation; BBN's HARK speech recognition software is employed. It has been tested using images from the RADIUS model board series; these aerial images depict buildings, roads, foliage, power plants, etc. The system is now described in detail.

Knowledge Representation

We are currently using a description logic language, LOOM (ISX Corporation), to construct an object-oriented model of our world. Multiple levels of representation are employed for objects in our world. Apart from the standard composition and specialization relations between entities, we also define a *concrete* relation to relate entities at different levels of abstraction (Niemann et al., 1990). Consider the entity *building* for example; a three-level representation is used for this: (1) building *concept* where functional information is stored, (2) building *objects* which represent instances of buildings in an image, and (3) polygons which represent the shape (geometry) of a building object. This representation

scheme permits visual and conceptual information to be stored in a distinct, yet shareable, manner.

Pre-Processing: Semi-Automated Road Detection

Roads form natural partitions in aerial scenes. They serve as useful landmarks upon which the analyst may base his/her description of the image. In our algorithm, it is assumed that the analyst provides an initial seed for the detection. Every connected network requires a separate seed. The algorithm is based on controlled continuity splines (Kess et al., 1987) and entropy minimization (Geman, 1994).

Speech and Language Processing

The speech annotation facility is comprised of two parts: (1) speech processing, resulting in transcribed ASCII text; and (2) natural language processing of the ASCII text. The limitations of the speech recognizer, combined with the requirement for real-time language processing strongly influenced the design of this interface.

Figure 5 illustrates a sample speech annotation used in processing the image in figure 1. In designing the speech annotation facility, there were several issues that had to be addressed. These included:

- Constraining the vocabulary and syntax of the utterances to ensure robust speech recognition; the active vocabulary is limited to 2,000 words.
- Avoiding starting utterances with words such as *this, the*; such words promote ambiguities resulting in poor recognition performance.
- Synchronizing the speech input with mouse input (e.g., *this is Kelly Laboratory*). Currently, we assume only one mouse click per utterance; the constrained syntax allows unambiguous synchronization.
- Providing an editing tool to permit correction of speech transcription errors.

In this first prototype of *Show&Tell*, the text (resulting from speech recognition) was processed as a complete narrative unit (i.e., "batch mode") rather than a sentence at a time. The justification for this is that it leads to more efficient search strategies; if all information is known a priori, the system can select an interpretation strategy which best exploits the information. Such a scenario is also reasonable if speech annotations are to be recorded and processed off-line. Finally, collateral information residing in site folders typically consists of narratives; thus any progress made in processing such narratives has broad applicability.

Processing narratives has intrinsic difficulties. People may refer to the same entity in several ways—e.g., *Baldy Tower, the tall building, the skyscraper.* Anaphoric references *it* and *them* are ubiquitous in narratives and

```
This image depicts the buffalo development and manufacturing
     complex.
Label segments roger two roger four and roger five as main
     street.
Label segment roger three as king street.
In the western half of area delta is the baird research institute.
It consists of four buildings.
Of these the leftmost is a long rectangular building.
Label this as the kelly laboratory.
This is used for developing and testing new products.
Label the l-shaped building as the operations center.
Label the large two storied building as the headquarters building.
This is where most of the administrative offices are located.
Label the small square building as the communications building.
```

Figure 5. Sample speech annotation for image in figure 2a.

require maintaining previous history. An important element in language processing is the construction of domain-specific ontologies. For example, it is important to know that a gym is a building and is associated with athletic facilities. Construction of large-scale ontologies such as this remains an open problem. With the proliferation of machine-readable lexical resources, working ontologies may be constructed which are sufficient for restricted domains.

Understanding spatial language in context (e.g., *the buildings on the river*) can get arbitrarily complex. To ensure robustness, we curtail idiomatic use of prepositions. Identifying and classifying proper noun sequences is a problem which also needs to be addressed. It is recognized that natural language understanding is itself a complex area. Since we are using language to simplify the task of vision, constraints have been imposed on the syntax and semantics of utterances to simplify processing. Although the IA cannot use unrestricted "natural language," there is sufficient flexibility to render it a convenient interface.

The output of language processing is a set of constraints on the image. These can be: (1) spatial, (2) characteristic (i.e., describe properties of an object or area), or (3) contextual. Constraints may be associated with single objects or a collection of objects (e.g., *the set of ten buildings*). The set of LOOM assertions output falls into two categories: (1) area, building, and aggregate concepts, etc.; and (2) relations between concepts indicating spatial and other relationships.

Using Scene-Specific Context for Image Interpretation

Recently, there has been a lot of academic interest in the *Integration of Natural Language and Vision* (INLV) (McKevitt, 1994; Srihari, 1995b). One of the objectives of this research effort is to use the interpretation of data in one modality to drive the interpretation of data in the other. We highlight the fact that collateral-based vision exploits a reliable hypothesis

of scene contents. We obtain this hypothesis from sources other than bottom-up vision to aid the vision processing. However in the INLV community, there is sufficient interest in using bottom-up vision to generate natural language descriptions of scenes; to use vision techniques in natural language understanding; and to model deep representation and semantics of language and perception. We choose a far more conservative goal of examining how image interpretation can benefit from collateral information.

The use of collateral information is extended in this domain by: (1) devising a uniform representation for domain- and picture-specific constraints, (2) employing spatial reasoning to use partial interpretation results in guiding the application of the vision processing tools, and (3) generalize the problem representation in an object-oriented manner to deal with multiple object types (not just faces).

The idea of using partial interpretation results to localize the search for related objects has been used in several vision systems. Earlier, we classified constraints as contextual, spatial, and characteristic. When the control algorithm employs any of these constraints to disambiguate candidates, we call them *verification* constraints and when it employs them to locate candidates for an object, we term them *locative* constraints. We assume cost and reliability measures for our object locators and attribute verifiers are available. The control algorithm loops over three stages: (1) decision stage, (2) labeling stage, and (3) propogation stage.

The input to the interpretation module consists of a constraint graph. The nodes represent objects such as buildings, roads, logical areas, and aggregates of these. The arcs denote the spatial and characteristic constraints on these objects. A *working* constraint graph is created incrementally by adding nodes chosen by the decision module. Partial labeling is attempted, and the results are used for spatial prediction (Chopra & Srihari, 1995).

Annotation and Querying

Figures 2 and 3 illustrate the annotation and querying functions provided in *Show&Tell.* These are the facilities which eventually are most valuable to an IA—all the processing described to this point enables these functions. Using the annotation tool, an IA may point to an object (which has already been segmented) and type in any further information which may be relevant. Such information would be available for querying at a later point.

Querying is with respect to a single image or (eventually) across an image database. Currently, we have focused on spatial and ontological queries—e.g., *Show all buildings to the right of Kelly Laboratory* or *Show all athletic facilities.* Future plans include temporal query processing as well as a speech interface.

MMVAR: A MULTIMODAL SYSTEM FOR VIDEO ANNOTATION AND RETRIEVAL

The objective of this project is to create a multimodal system for video annotation and retrieval. Although the system is general enough to be applicable for any video, it is most suited for video such as intelligence surveillance; such video is characterized by long panoramic sequences (aerial or ground-level) of natural scenes, activities involving people and vehicles, building complexes, etc. The objective is to automatically derive a semantic segmentation of the video such that it can be efficiently queried based on its contents. Since automatic statistical segmentation techniques based on video alone are not useful for retrieval purposes, the speech annotation is used as a guideline for determining logical discontinuities in the video. Features such as long pauses in annotation and the presence of keywords are combined with video-level segmentation in order to arrive at a semantic segmentation of the video. Statistical natural language processing techniques are subsequently invoked which attempt to classfiy the resulting shots based on topic.

The goal is to produce a robust scalable system demonstrable by late August 1996. This will serve as a sound infrastructure for further research in combining linguistic and visual information for intelligent content-based retrieval.

The system we are working on has the following functionality: (1) populating (entering new video into) the database, (2) annotation of existing video using a speech interface, and (3) retrieval of video based on exact and inexact (full text) techniques. All this is being provided in a single graphical user interface.

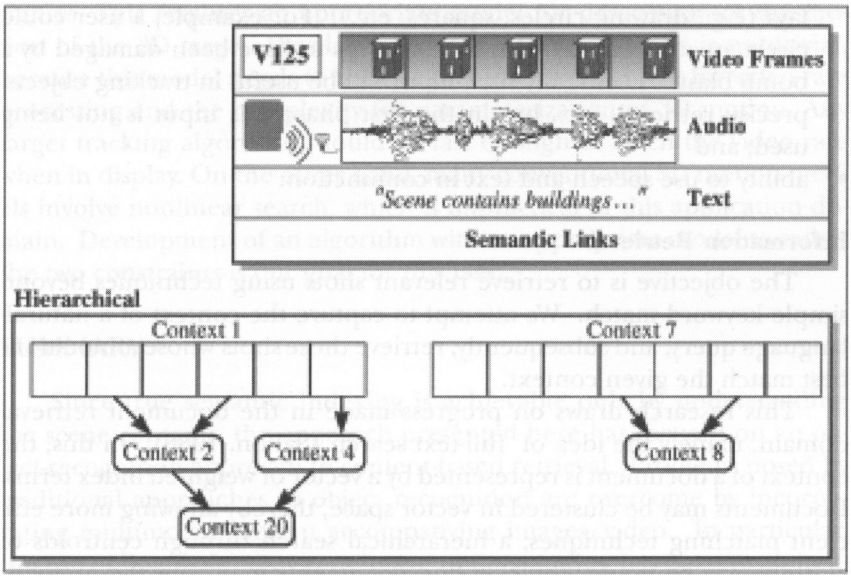
Populating Database

This feature permits new video to be entered into the video database. We allow either digital video to be entered or a video capture facility to be used. Video is segmented using statistical segmentation techniques and stored in the database as video objects. These constitute the atomic units of the database.

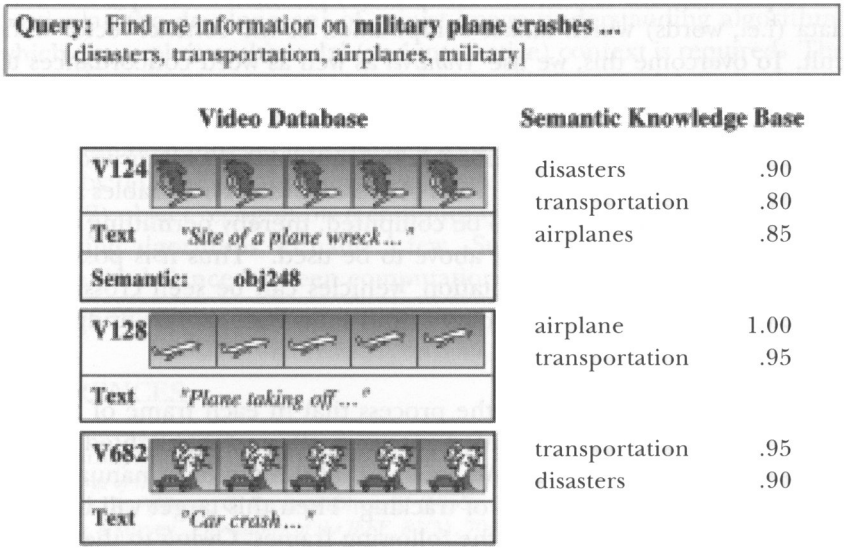
Video Annotation

This feature permits the creation of nonvideo objects by augmenting video with audio, text, graphical overlays, etc. Users employ a speech recording device to attach video annotation to a video. This is subsequently processed offline by a speech recognition algorithm. The output is video synchronized with corresponding audio, text, and any graphical overlays input by the user. The following features are permitted:

- multiple annotations per video;
- ability to pause video and continue speech annotation;



(a)



(b)

Figure 6. (a) hierarchical classification of video objects; (b) combining information sources in retrieval.

- ability to do single frame annotation using speech and graphical overlays (e.g., drawing circles, squares, etc.). For example, a user could circle an area and say "this area appears to have been damaged by a bomb blast." Eventually such input will be useful in tracking objects, precise retrieval, etc., but in the first phase this input is not being used; and
- ability to use speech and text in conjunction.

Information Retrieval

The objective is to retrieve relevant shots using techniques beyond simple keyword match. We attempt to capture the context of a natural-language query, and subsequently, retrieve those shots whose annotations best match the given context.

This research draws on progress made in the document retrieval domain, namely the idea of full-text search (Salton, 1988). In this, the context of a document is represented by a vector of weighted index terms. Documents may be clustered in vector space, thereby allowing more efficient matching techniques; a hierarchical search through centroids of clusters may be used to quickly find documents matching the query of the context thereby avoiding sequential search.

This method needs adaptation if it is to be used with text annotations corresponding to video shots. The primary problem is the sparseness of data (i.e., words) which makes computation of the context vector difficult. To overcome this, we use *WordNet* as well as word concordances to generate a neighborhood of words related to a given word; *WordNet* includes relationships such as synonymy, hypernomy, hyponomy, etc. Using statistical procedures, it is possible to compute the co-occurence probabilities of the given word with each of these words. This enables a larger set of weighted index terms to be computed, thereby permitting classification techniques such as the above to be used. Thus it is possible to retrieve a shot with the annotation "vehicles can be seen crossing the causeway...." as a result of the query "find shots with cars on bridges."

Vision Module

Target tracking refers to the process that in each frame of a video sequence, the target is segmented automatically and highlighted until it disappears in the sequence. We assume that the target is manually segmented out at the first frame of tracking. Then this target will be automatically tracked down in all the following frames. Owing to the nature of this problem, its solution has great application interest in intelligence surveillance and monitoring.

Due to the camera focal change (e.g., zoom in/out), camera motion, and/or the target motion, the 2D images of the target in different frames may have different shapes and intensity/color values. An automatic

tracking algorithm should be able to handle these changes in different frames. This requires an appropriate model for description of the motion of the 3D target. Development of this motion model is nontrivial, because this model needs to balance the time requirement for video-rate processing and the complexity for parameterization of 3D motion. Any target tracking algorithm should be fast enough to catch the video rate when in display. On the other hand, many conventional 3D motion models involve nonlinear search, which is impractical in this application domain. Development of an algorithm with an appropriate model to satisfy the two constraints is our goal for this task.

SUMMARY

Since true semantic indexing is achievable only by understanding the scene contents, the approach presented here has focused on an object-recognition approach to content-based retrieval. Problems posed by traditional approaches to object recognition are overcome by incorporating multimodal content accompanying images/video. In particular, the use of linguistic context in image understanding has been exploited. Both automated (PICTION) and semi-automated applications (*Show&Tell*, MMVAR) have been presented.

Much work remains in making this technique completely viable; in particular, the development of robust image understanding algorithms which can exploit multimodal (and interactive) context is required. The work presented here is only a start. Issues relating to scalability and computational effort must also be examined. Finally, techniques for integrating multimodal indexes must be developed. These indexes may be generated by: (1) statistical pattern recognition techniques (e.g., color, texture), (2) object recognition techniques (e.g., face locators), or (3) text processing algorithms, to name a few. Such integration will ultimately provide the balance between computational feasibility and the need for semantic retrieval.

REFERENCES

Chakravarthy, A. (1994). Representing information need with semantic relations. In *Proceedings of COLING-94,* 1994. Kyoto, Japan.

Chellappa, R.; Wilson, C. L.; & Sirohey, S. (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE, 83*(5), 704-740.

Chopra, R., & Srihari, R. (1995). Control structures for incorporating picture-specific context in image interpretation. In *Proceedings of IJCAI-95* (pp. 50-55). San Mateo, CA: Morgan Kaufmann.

Geman, D. (1994). The entropy strategy for shape recognition. In *IEEE-IMS Workshop on information theory and statistics* (October 27-29, 1994). Alexandria, VA:

ISX Corporation. (1991). *LOOM users guide, Version 1.4.* Available from: <http://www.isi.edu/isd/LOOH/documentation/loom.docs.html>.

Kass, M.; Witkin, A.; & Terzopoulos, D. (1987). Snakes: Active contour models. In *Proceedings of the First International Conference on Computer Vision* (pp. 259-268).

McKevitt, P. (Ed.). (1994). *Workshop on the integration of natural language and vision, AAAI-94* (Proceedings of the 12th national Conference on Artificial Intelligence). Seattle, WA: AAAI Press.

Niemann, H.; Sagerer, G. F.; Schroder, S.; & Kummert, F. (1990). Ernest: A semantic network system for pattern understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(9), 883-905.

Salton, G. (1988). *Automatic text processing.* Addison-Wesley.

Srihari, R. K., & Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of AAAI-94* (pp. 793-798). Seattle, WA:

Srihari, R. K.; Chopra, R.; Burhans, D.; Venkataraman, M.; & Govindaraju, V. (1994). Use of collateral text in image interpretation. In *Proceedings of the ARPA workshop on image understanding* (pp. 897-908). Monterey, CA:

Srihari, R. K. (1995a). Automatic indexing and content-based retrieval of captioned images. *IEEE Computer, 28*(9), 49-56.

Srihari, R. K. (Ed.). (1995b). *Computational models for integrating language and vision* (Proceedings of the AAAI-1195 Fall Symposium). Menlo Park, CA: AAAI Press.

Srihari, R. K. (1995c). Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review* (special issue on Integrating Language and Vision), *8*(5), 409-430.

Strat, T. M., & Fischler, M. A. (1991). Context-based vision: Recognizing objects using information from both 2-D and 3-D imagery. *IEEE Transactions on PAMI, 13*(10), 1050-1065.