Show&Tell: A Semi-Automated Image Annotation System

Rohini K. Srihari and Zhongfei Zhang
State University of New York at Buffalo

A multimedia system for semi-automated image annotation, Show&Tell combines advances in speech recognition, natural language processing, and image understanding. Show&Tell differs from map annotation systems and has tremendous implications for situations where visual data must be coreferenced with text descriptions, such as medical image annotation and consumer photo annotation.

Show&Tell takes advantage of advances in speech technology and natural language/image understanding research to make the preparation of image-related information more efficient. Specifically, we aim to identify relevant objects and regions in the image, as well as to attach text descriptions to them. We use a combination of automated and semi-automated image understanding tools in object and region identification.

Image analysts can use Show&Tell in applications where text descriptions must be coreferenced with image areas, such as medical image annotation, National Aeronautics and Space Administration (NASA) space photo annotation, and even consumer photo annotation. Medical images suit our system well, since radiologists already employ speech to dictate their findings and robust image understanding technology is available for several areas, such as chest and lung radiographs.

In a joint effort with Kodak, we are adapting our system for consumer photo annotation. Since still cameras can be fitted with microphones, speech annotation of photos is now possible. Consumers will be able to easily create searchable digital photo libraries of their pictures and focus primarily on pictures of people in various contexts.

Multimedia input analysis

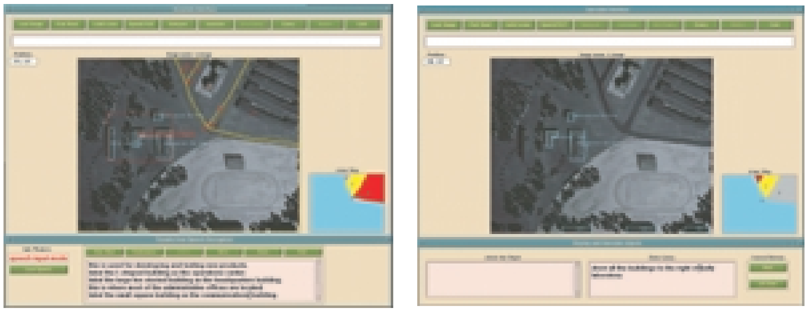
Multimedia systems involving speech and deictic input can be classified into two major categories: multimedia input analysis and multimedia presentation. Our work focuses on the former.

Our system differs from previous work in the area of adding text annotations to pictorial data in the following ways. Most systems assume that there already exists an underlying semantic representation of the pictorial data. We don't. We used Clark's terminology.[1] The region to which the user points is the *demonstratum*, the descriptive part of the accompanying text is the *descriptor*, and the region to which the user intends to refer is the *referent*. Much of the recent work in multimedia input analysis concerns disambiguating ambiguous deictic references, that is, determining which of the possible referents that map to the same demonstratum is intended by the user.[2] Accompanying linguistic input, in the form of speech, is used for this purpose. Such systems assume that the type of deixis being used, known as *demonstratio ad oculus*, is distinguished by the fact that the objects on display have already been introduced and that the user and the system share a common visual field. For example, in the case of maps, a graph represents the semantics. Thus, deictic references using a mouse or other pointing devices can be associated with the underlying geographical entity (or relationship).

Our work pertains to situations where the image hasn't been subjected to any previous semantic interpretation. Thus, when a user clicks on a building and supplies a name for it, the system doesn't initially know about the building. The mouse click could correspond to a single pixel, a region of pixels, or the entire image. In fact, one goal of our work is to examine the usefulness of text descriptions in image interpretation.

Here, we focus on the design issues of developing an efficient easy-to-use system. The major bottleneck in such situations occurs in image interpretation. At one extreme, a highly automated system would first perform image interpretation without user input. In this case, the user could subsequently attach text annotations to automatically detected objects and regions. This would make the annotation task similar to the map annotation systems previously mentioned and permit nonexperts to use it.

Recent approaches by the Defense Advanced Research Projects Agency (DARPA) image understanding community (Radius program[3]) have targeted developing robust image understanding



(a) **(b)**

Figure 1. An example of annotating and querying an image through speech interface. (a) In the annotation mode, the upper window displays the image interpretation result while the lower window displays transcribed speech. (b) In the query mode, the upper window displays the query result while the lower window allows a user to enter a query.

technology by providing detailed 3D site models[4] of a particular civilian or military facility. The community is developing techniques for automatic registration of a given 2D view to the 3D site model, thus enabling automatic object detection and change detection in a given area. However, this technology is still under investigation; it's not practical to assume that such detailed 3D information is always available. On the other hand, a manual approach would require substantial input from the user in terms of object identification.

Currently, intelligence experts are trained to use highly sophisticated image domain processing and understanding tools, such as 3D snakes,[5] in the task of object identification. Typically, it takes several hours to annotate an image using these techniques alone.

Our goal is to develop a golden mean between these two extremes, whereby a nonexpert could use the system to annotate images. This can be achieved by the intelligent use of speech, mouse input, and adaptation of image interpretation technology.

A new paradigm

Through speech annotation, Show&Tell exploits linguistic context for image interpretation. The interlingua between image understanding and language understanding is expressed in terms of object models. Linguistic context permits dynamic construction, or modification, of object models. In turn, these drive the image interpretation engine.

Testing the validity of this approach required us to develop capabilities in several areas, including image interpretation, speech understanding, natural language understanding, knowledge representation, and spatial reasoning. The model that exploits the correlation between these functionalities and leads to an integrated system represents a major achievement.

As an initial testbed in which to develop and evaluate the new paradigm, we're working with aerial images of various military and civilian sites. Currently, image analysts spend many hours labeling these images, even though they use state-of-the-art graphical tools. These images must be labeled to detect any changes over time.

Annotation and querying

Show&Tell provides two stages of functionalities. In the first stage—annotation—the system automatically interprets and indexes images. An expert user 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 transcribes the input and synchronizes the speech with the mouse input.

A natural language understanding component processes the resulting narrative and generates visual constraints on the image. An image interpretation system uses this information to detect and label areas, roads, and buildings in the image.

Finally, the system attaches additional collateral to objects it has identified. Thus, the system output is a natural language collateral description and an annotated image. The annotated image consists of a semantic representation of the text and locations of objects (regions identified in the image). Show&Tell represents information in a way that enables spatial reasoning, temporal reasoning, and other contextual reasoning capabilities.

In the second stage—querying—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. Users could further query each of these areas for corresponding textual information.

Figure 1 illustrates the annotation and querying functions provided in Show&Tell. Using the annotation tool, an image analyst can point to an object (which has already been segmented) and add any further descriptions. Such information would be available for querying at a later point, either with respect to a single image or across an image database. To date, we've focused on spatial and ontological queries, such as "Show all buildings to the right of Kelly Laboratory" or "Show all athletic facilities."

We make several assumptions in specifying the task and our proposed solution. First, from the perspective of an image analyst, this approach constitutes a healthy compromise between tedious manual annotation and completely automated (image-based) interpretation. Since the task involves coreferencing image areas with textual descriptions, our system uses the text for dual purposes—coreferencing and assisting image interpretation.

The second assumption concerns the use of preconstructed geometric site models. The image understanding community has used these effectively for registering new images of a known site and subsequently for change detection. Since these kinds of site models may not always be available in many applications, and since the construction of these models is typically expensive, we assume that no site models are available.

Finally, for reasoning purposes, we assume that an approximate shape representation for objects suffices. We represent objects such as buildings or areas by polygons with orientation. Polylines indicate roads and rivers. Our system also allows more exact representations for display purposes.

Architecture

Figure 2 illustrates the functional architecture of Show&Tell, which runs on a Sun Ultra. It consists of five principal modules:

1. Graphical user interface (GUI)
2. Preprocessing or landmark detection
3. Speech and language processing
4. Image interpretation
5. Knowledge base

The user interacts with the system through the GUI, both in the annotation and querying stages. The entire system contains several knowledge sources, both static and dynamic. Examples of static knowledge sources are the domain ontology (required by the speech and language module) and the object schemas or descriptions (required by the image interpretation module).

The key knowledge source is the visual model, which is incrementally constructed. It consolidates both visual and linguistic information. This dynamic, image-specific knowledge is represented in the knowledge base. Information about specific images, as well as a collection of images, are maintained in this knowledge base. Subsequent queries build on this information.

We wrote Show&Tell's natural language processing modules in common Lisp and the image interpretation code and GUI in C and Motif. We employed Bolt Beranek & Newman's Hark speech recognition software. Processing time is close to real time and varies with the amount of speech input and the number of objects to detect and label.

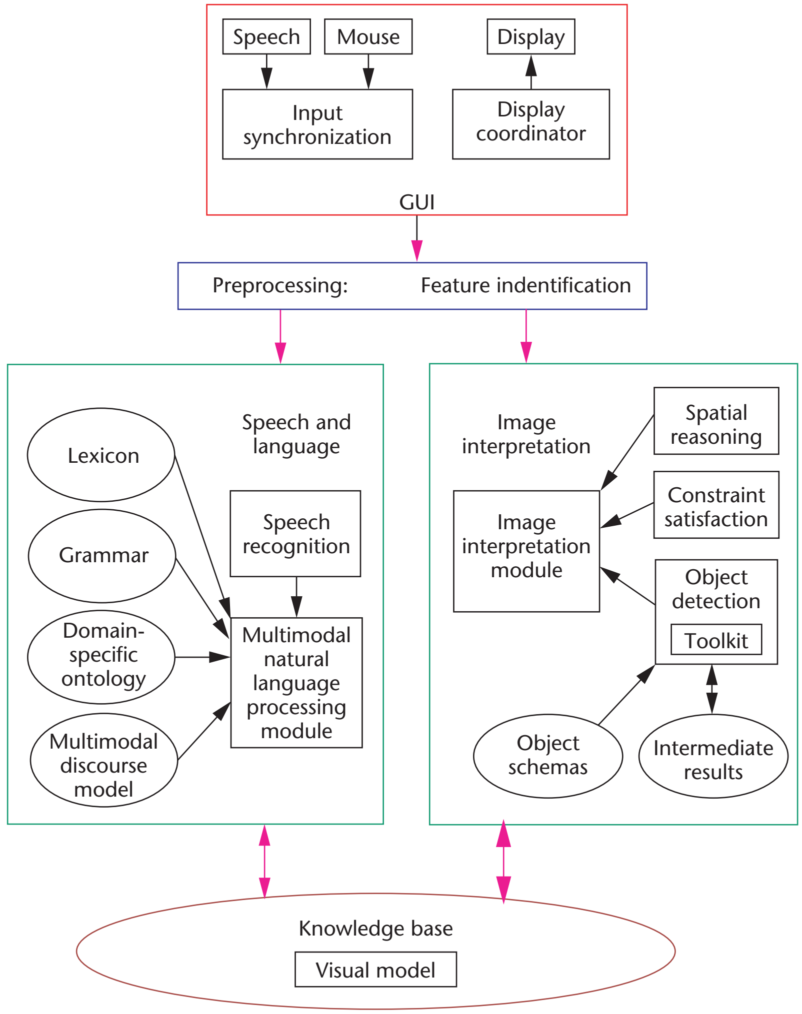


Figure 2. Functional architecture of the Show&Tell system. The round boxes indicate knowledge sources. The rectangles indicate processing modules. The five colored boxes represent the five basic functional areas. The pink arrows show the control flow.

Image interpretation

The main task of the Show&Tell system is image interpretation. The goal of this component is to "understand" salient portions of an image based on the qualitative annotation provided from the speech by image analysts and/or the mouse input. Understanding is reflected in a set of symbolic representations of the image events,

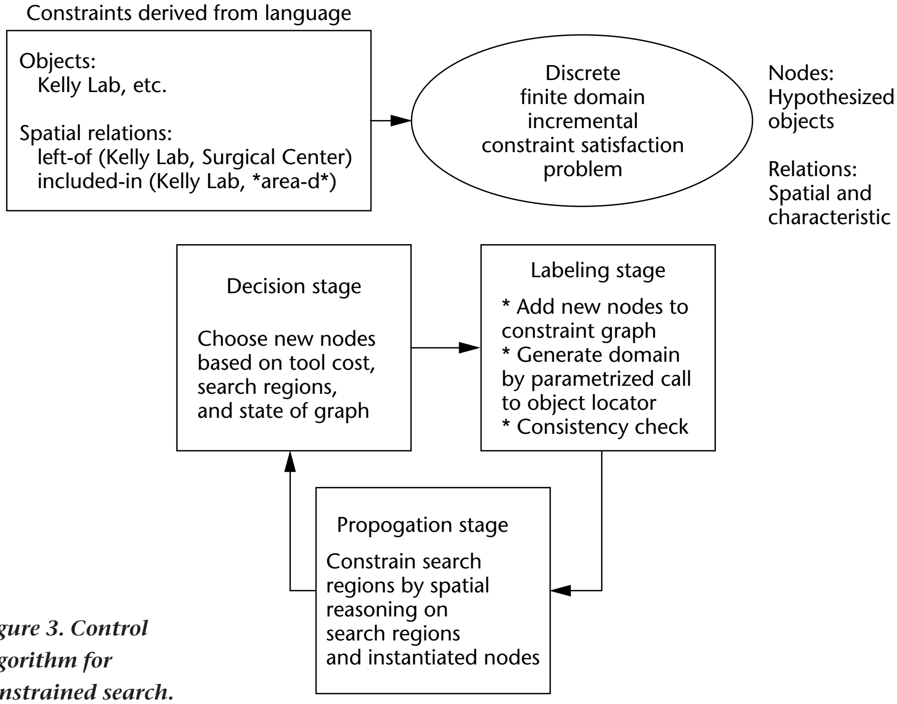


Figure 3. Control algorithm for constrained search.

including objects and their relationships. This information is stored in the knowledge base for future querying and retrieval.

The first step is to detect reliable landmarks in the image on which analysts can base their description. With roads forming natural partitions in an aerial scene, analysts provide an initial seed for the detection—a single mouse click at some point in the road network. Every connected network requires a separate seed. The algorithm builds on controlled continuity splines and entropy minimization[5] and is commonly referred to as a 2D snakes technique.

Once users identify landmarks, they begin to describe other salient features in the image using a speech interface. If speech provided by an image analyst contains complete information about all the objects expected to appear in a given image, Show&Tell initiates a search to detect and locate these objects. Some buildings may be easier to find than others. These in turn constrain the search for other buildings.

Recently, academics have shown a lot of interest in the Integration of Natural Language and Vision (INLV).[6] 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. Collateral-based vision exploits a reliable hypothesis of scene contents. We obtain this hypothesis from sources other than bottom-up vision to aid visual processing. Srihari[7,8] showed how collateral information from captions accompanying news photographs could be used to identify people in the photographs: Generate a hypothesis of the image contents, then invoke a routine to locate face candidates in the image, and use graph matching techniques to establish correspondence between face candidates and persons mentioned in the caption.

Constraint satisfaction problems

Search problems in computer vision have frequently been posed as constraint satisfaction problems (CSP). A static CSP[9] represented by *(V, D, C, R)* involves a set of m nodes, $V = \{v_1, v_2, \ldots, v_m\}$, each with an associated domain D_i of possible labels. The search space consists of the Cartesian product of the nodes' domains, $D = D_1 \times D_2 \times \ldots \times D_n$. A set of constraints C is specified over some subsets of these nodes, where each constraint C_p involves a subset $\{i_1, \ldots, i_q\}$ of V. C_p is labeled by a relation R_p of R, a subset of the Cartesian product $D_{i_1} \times \ldots \times D_{i_q}$ that specifies those labels of the nodes that are compatible with each other.

We choose to model the image interpretation problem as a CSP, but with some modifications. The initial 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.

Since domain generation is an expensive process in this application, the system incrementally creates a working constraint graph by adding nodes the decision module chooses. We assume that cost and reliability measures for our object locators and attribute verifiers are available. These, along with the current status of the constraint graph, determine which nodes should be expanded. We attempt partial labeling and use the results for spatial prediction.[10]

The control algorithm depicted in Figure 3 loops over three stages as follows.

- *Decision stage.* A complex cost and utility function decides which object class to detect next, given the current state of the constraint graph and labelings. The factors we take into account for this function are: tool cost, reliability, number of objects to be found, size of the search region, the number of constraints between this class and the objects of interest, and the level of interest in this particular class of objects. Some measures are class specific (tool cost), others are instance specific (number of objects), and some are dynamic (size of search region).

- *Labeling stage*. We locate candidates for objects of the chosen group using the available collateral information. These objects, candidates, and the constraint relations between the objects are inserted in a constraint graph representing the partial interpretation. We apply constraint satisfaction techniques to uniquely label the objects. Constraints between objects in the new group, and between objects in prior groups and the new ones, are satisfied, thereby advancing the partial interpretation of the image.

- *Propagation stage*. Any object that has been labeled uniquely in the previous module potentially constrains the search region for those objects involved in a spatial relation with it. The propagation module computes the search region using the labeled object and the type of spatial relation. So far, this spatial prediction has been implemented for binary spatial constraints.

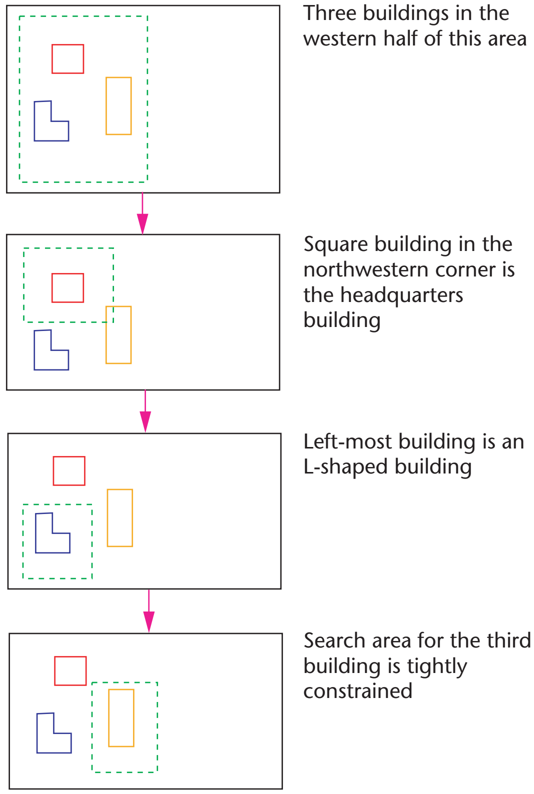


Figure 4. A hypothetical example to show how CSP would help identify and detect objects such as buildings efficiently.

Figure 4 shows how CSP works. If the annotation says three buildings lie in the western half of this area, CSP immediately focuses on the left half of the image. Then, if the annotation further explains that the square building in the northwestern corner is the headquarters building, the search area would be further focused on the top-left quarter. The program employs the building detection module to search in this quarter for a square building, significantly reducing the search. Since the CSP algorithm knows that there must be a square building in this quarter, the building detection module must return a square building even if the detection confidence is low due to image noise.

If the annotation further explains that the left-most building is an L-shaped building, CSP then instructs the building detector to focus on the area below the square building. The L-shaped building should be the first one from the left. Once the system detects the L-shaped building, the search area is tightly constrained.

We've implemented a spatial reasoning module that verifies the various spatial constraints between objects. Currently, we've implemented only simple 2D relationships such as "left-of," "above," "beside," "between," and "contains." We use polygonal representations for buildings and regions, and polylines for roads. This suffices for the initial prototype of our system, but would require significant enhancement to include other object types and shape representations. Spatial reasoning is a field of research on its own.

The final output of the image interpretation module is either a set of image coordinates representing an object in the image if it can be labeled uniquely, or the constrained search region for the object or region in case unique labeling wasn't possible. This information is recorded in the visual model, thus completing the integration of spatial and linguistic information.

Qualitative building detection

In addition to roads and areas, buildings are a commonly seen object type in an aerial image. In the image understanding community, researchers have proposed several successful algorithms. [11]

Unlike most of the existing building detection algorithms, which detect buildings based on site models, the algorithm proposed in this research is

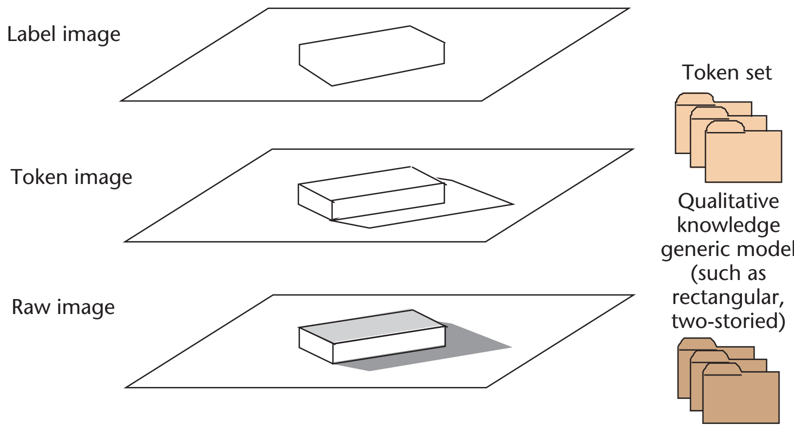


Figure 5. Illustration of building detection. Given a raw image, we apply a line-finding algorithm to the raw image to obtain a line token image. This is followed by line grouping to detect buildings.

completely independent of any site models and/or camera pose and calibration information. Instead, the building detection module is driven only by qualitative information provided by image analysts through speech.

Here, the goal of building detection is to approximately delineate building boundaries in an annotated monocular image. We call this approach qualitative building detection.[12] Since the Show&Tell system assumes that the quantitative 3D site model and/or camera information may not always be available, it's necessary to develop this qualitative approach to building detection to complement the mainstream approaches in the image understanding community.

Our building detection is based on knowledge-supervised perceptual grouping (KSPG). Figure 5 shows the general scenario of how KSPG works. This grouping has three layers. Given a raw image, we apply a line-finding algorithm to the raw image to obtain a line token image.

Because of the potential performance errors of line-finding algorithms, plus the fact that there is always noise in an image (shadows, for example), the tokens need to be grouped to form a semantically correct segmentation. Therefore, perceptual grouping must be applied to detect all the semantically correct lines, such as building boundaries, and ignore "noisy" lines, such as the cultural and textural lines inside the building boundaries, as well as lines caused by shadows.

The result of this detection is called the label image. If the grouping was conducted only at the token image without any context from the raw image, it would be difficult to make a decision whether a line token is part of a noise or part of a building boundary.

Knowledge must be used to disambiguate the ambiguities. We exploit two types of knowledge in KSPG. One is from the qualitative description by an image analyst through speech. This information contains qualitative descriptions of building shapes (rectangular, L-shaped), qualitative viewing directions (a nadir view, an oblique view), the number of stories (a two-storied building), and shadow information (yes/no shadow). Other qualitative information also may be used as part of the input, such as a qualitative description of the location (this building is located in the southwest corner), but it isn't necessarily essential for the building detection system.

The other type of knowledge used to supervise the perceptual grouping is obtained directly from the raw image per se. In perceptual grouping, whenever we're unsure whether we need to group two tokens together, the best way to make the decision is to return to the raw image to check the context of these tokens. If edge information exists (even though it may be very weak) between the two line tokens, these two tokens are likely to be parts of a physical line and thus should be grouped together. Otherwise, the two tokens represent two different contexts. We call this process of going back to the raw image in perceptual grouping reinvestigation.

KSPG occurs through dynamic matching, in which line tokens are gradually grouped together or discarded, based on repeated consultation of these two types of knowledge.[12] The detected buildings are represented in terms of a set of corner coordinates, which the system sends to the visual model for displaying and labeling in the image.

Speech and language processing

The speech annotation facility consists of two parts: speech processing, which results in transcribed ASCII text, and 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 6 illustrates a sample speech annotation used in processing the image in Figure 1. Several issues had to be addressed in the speech annotation facility design:

- 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" or "the." (Such words promote ambiguities, resulting in poor recognition performance.)

- Synchronizing speech input with mouse input; for example, "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 our design, all the information required for object and region labeling was provided solely by speech, with minimal mouse input. This information was presented using sentences describing visual characteristics of objects and regions, as well as spatial constraints.

We imposed these restrictions for several reasons:

- In many situations, speech input must be verbose, since it's used to provide general textual documentation of visual data. In other words, speech input is used for purposes other than simply labeling objects. This is the case with intelligence photos as well as radiologists' reports. We were interested in evaluating the efficacy of using language alone in image interpretation.

- In several situations mouse input isn't feasible since the user's hands may be otherwise occupied.

- Constantly using the mouse can interrupt the user's train of thought.

- Image interpretation is rendered more efficient by combining the information obtained from several utterances and performing a single search for all objects and regions, rather than finding objects one at a time. The system can select an interpretation strategy that best exploits the information.

Language processing output is a set of constraints on the image. These can be spatial, characteristic, or contextual. Constraints can be associated with single objects or a collection of objects, "the set of ten buildings," for example. These constraints fall into two categories of representation—aggregate concepts, such as "area" and "building," and relations between concepts indicating spatial and other relationships.

An important element in language processing is the construction of domain-specific ontologies. For example, it's important to know that a gym represents a building associated with athletic facilities. Construction of large-scale ontologies such as this remains an open problem. With the proliferation of machine-readable lexical resources, we can construct working ontologies that suffice for restricted domains. These constraints are critical to the subsequent image interpretation phase in which relevant objects are detected and labeled.

This image depicts Jarvis Park.
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.
Label the L-shaped building as the Operations Center.
Label the large two-storied building as the Headquarters Building.
Label the small square building as the Communications Building.

Figure 6. Sample speech annotation for the image in Figure 1.

Knowledge base

A common knowledge representation system facilitates communication between the natural language module and the image-understanding module. We currently use a description logic language Loom (http://www.isi.edu/isd/LOOM/documentation/LOOM-DOCS.html) to construct an object-oriented model of our world and employ multiple levels of representation 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.[13] Consider the entity "building," for example. We use a three-level representation as follows:

- a building concept where functional information is stored,

- building objects representing instances of buildings in an image, and

- polygons representing the shape (geometry) of a building object.

This representation scheme permits us to store visual and conceptual information in a distinct, yet shareable, manner.

Figure 7 (next page) illustrates a portion of this hierarchy, particularly the resulting visual model constructed from a speech annotation. The visual model includes information about physical objects appearing in the image and spa-

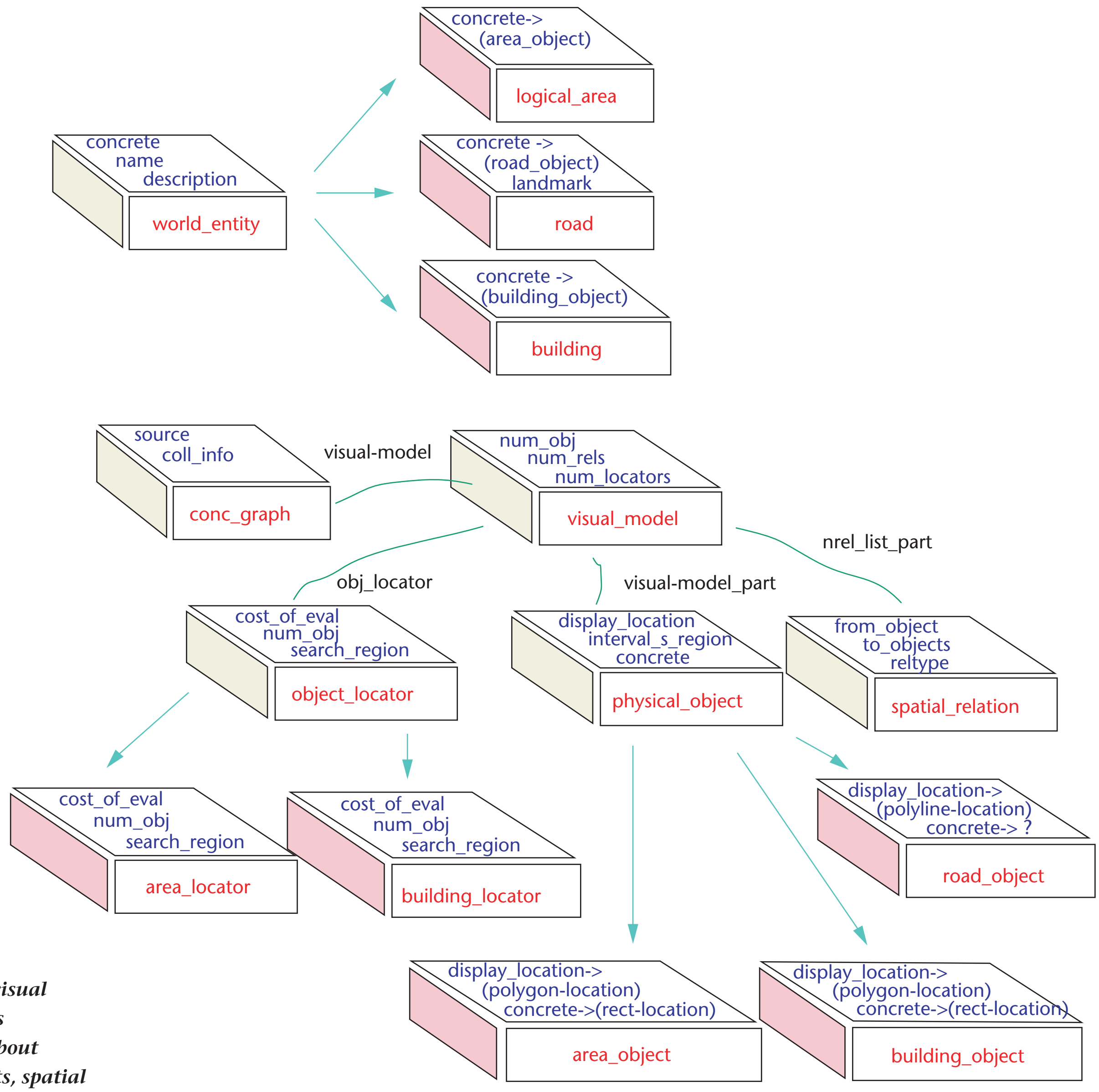


Figure 7. The visual model includes information about physical objects, spatial relations between them, and object locators. Physical objects have pointers to logical objects that store name and other text descriptions.

tial relationships between them, as well as information about object locators.

Physical objects have links to their logical object counterparts. The links store the names of objects and additional text descriptions. When the image interpretation module later detects and labels the required objects, their image coordinates are also recorded in this knowledge structure. Thus the visual model consolidates the linguistic and spatial information.

Evaluating Show&Tell

We evaluated Show&Tell on its system performance and as a new paradigm for image interpretation and annotation.

System performance

We tested our system on a set of images that DARPA provided. The images build on a synthetic model board depicting a typical military installation site. This permits a controlled testing environment where parameters such as roof topography can be controlled.

Evaluating the performance of such a complex system proves difficult because it may not always make sense simply to add the success rates of all the components of the system. We chose to report the evaluation of each module individually and the typical timing information for the whole system, which includes the time for query processing, reasoning, and retrieval.

Since the algorithms used in natural language processing, knowledge representation, and constraint satisfaction modules are all deterministic, there's no point in evaluating their success rates. We only need to evaluate the speech recognition and building detection modules.

Given a constrained vocabulary and syntax, speech recognition generally performed very well. In many cases, there were no errors in a set of 10 sentences. When there were errors, only one to two had to be manually corrected per 100 words spoken.

Building detector performance depends not only on the image quality of the buildings to be evaluated, but also on the accuracy of the qualitative information provided from the speech. To further simplify the evaluation process, we assume that for each of the building images being evaluated, the spoken description is unambiguous.

We evaluated five aerial images of 70 differently viewed building images for our building detection module. The detection rate was 76.5 percent, and the false positives were 2.9 percent. Poor results of line detection and limitations of the language capabilities in describing the building shapes were among the factors causing detection failures. The low false positives were due to reinvestigation techniques employed in our KSPG algorithm.

The time to respond to an annotation activity depends on many factors, including the quality of the speech, complexity of the speech semantics, and quality of the image. Poor quality speech in a noisy environment may necessitate manual correction of speech recognition output.

If the semantics required to understand the speech utterances is limited to making simple factual assertions, such as "This building is the headquarters," then speech understanding is virtually instantaneous. However, if the utterance involves spatial reasoning and interpretation, "In the western half of Area Delta, there are four buildings," for example, then more processing time is required.

Finally, poor image quality results in longer search times for building detection. For the standalone version of the prototype system of Show&Tell implemented on a Solaris 2.5.1, the typical response time for each utterance ranged from 0.1 second to several seconds. Once the annotation is completed, the time to respond to a typical query is instantaneous.

The paradigm: Lessons learned

The first lesson we learned was the need to restrict mouse input to a minimum during the speech annotation stage. An approach that relies heavily on linguistic descriptions has several drawbacks. Since users are speaking sentences continuously (with no break in between to wait for system output), these utterances constitute a narrative.

Processing narratives can be difficult. People may refer to the same entity in several ways, for example, "Baldy Tower," "the tall building," or "the skyscraper." Anaphoric references, such as "it" and "them," are ubiquitous in narratives and require maintaining previous history. In our system, the discourse model maintains a short history of the previous object and aggregate mentioned. Thus, some limited anaphora is permitted.

Second, understanding spatial language in context, "the buildings on the river," for example, can get arbitrarily complex. Processing a spatial language has two parts—the semantics of the utterance itself (which of the several hundred meanings of "on" is implied in a sentence), and the validation of a spatial relation in an image. Our spatial reasoning module validates the spatial relation and depends on the representations chosen for objects and regions. To ensure robustness, we curtail idiomatic use of prepositions. In the current system, we use simple projective prepositions, such as left, right, above, below, beside, and between, for expressing spatial relations.

Natural language understanding is itself a complex area. Since we use language to simplify the task of vision, we imposed constraints on the syntax and semantics of utterances to simplify processing. Although the image analyst can't use unrestricted natural language, there is sufficient flexibility to render it a convenient interface.

The most serious drawback of our approach to annotation however, is the lack of immediate feedback to the user. Users experimenting with our system (including intelligence analysts) told us when they speak, they expect the system to react. Based on their feedback, we decided to try a more interactive approach involving substantially more mouse input. Oviatt[14] discussed design issues involving robust speech recognition for human-computer interaction tasks. We incorporated many of her suggestions into this new scenario.

Ideally, a user should be able to click on an area in the image and say "Two-Story Building … Kelly Laboratory." Show&Tell attempts to locate and label the building immediately, based on characteristics centered around a point that enables building detection. In situations where this isn't possible due to a lack of information, Show&Tell displays the search regions for objects as they're

Label region Jarvis Park.
<click to indicate entire image>
Label segments roger two roger four and roger five as Main Street.
Label segment roger three as King Street.
Label region Baird Research Institute.
<click to indicate extent of bounding box representing Baird Research Institute>
Label object long rectangular building Kelly Laboratory.
<click on building>
Label object L-shaped building Operations Center.
<click on building>
Label object large two storied building Headquarters Building.
<click on building>
Label object small square building Communications Building.
<click on building>

Figure 8. Revised speech annotation for the image in Figure 1. This permits more interactivity and requires less speech input.

mentioned. These search regions are continuously refined based on further input.

Figure 8 shows a revised speech transcript—based on the above findings—for the example presented in Figure 1. The system automatically infers that the four mentioned buildings are part of the Baird Research Institute. The techniques used to find buildings in this case are very different. Previously, the region representing the cluster of four buildings was located and shape characteristics and spatial constraints were used to label the individual buildings. Based on the revised speech and mouse input, the system doesn't use spatial constraints in labeling.

In this scenario, the initial speech input is a very restricted syntax used solely for locating and labeling objects and regions. A user may attach further descriptive information to an object after it has been identified by clicking on it and describing it. These sentences only need be transcribed—no attempt is made to understand them.

Another lesson we learned is the limitation of qualitative building detection attributable to language restrictions in describing building shapes. This is one of the major factors contributing to the failure of the building detector. Even though many buildings can be described simply by a verbal term such as rectangular, or L-shaped, many buildings don't have simple verbal descriptions. One example is a long rectangular building with two asymmetric, long protrusions at both sides. It's neither a rectangular building, nor a T-shaped building, nor a cross-shaped building. What is it? If we approximate these building images into one of the existing qualitative building models, our building detection would fail. In this case, we would have to revert to quantitative site models for building detection.

The proposed paradigm will perform better in situations where strong site models are available. Examples of such situations include medical domains, such as chest X rays. Well-defined models for the chest region already exist. Language and deictic input indicate regions or special characteristics. Site model information is certainly useful in enhancing the robustness and reliability of the building detector.

Finally, in terms of the architecture of the system, we suggest a more flexible, open architecture so that changes made in one module—the user interface, for example—would not affect another module. We found that the Open Agent Architecture[15] fits these requirements. It facilitates multimedia interfaces, as well as permitting easy access to intermediate results via a browser interface.

We're currently working on porting the technology to other domains, which requires modifications to

- *Object ontology*. A new ontology of objects specific to that domain must be developed. For example, in the case of chest X rays, the anatomy of the chest along with typical abnormalities must be modeled.

- *Image processing tools*. It's necessary to develop a suite of automatic and semi-automatic tools for object detection and classification. For example, in the case of consumer photos, we're employing face detection modules.

- *Landmark identification*. In each application domain, some visual features may be used as landmarks to assist in describing the image. In the case of medical images, features such as arteries, ribs, and veins may serve this purpose. In the case of consumer photos, faces are typically the most salient features.

- *Spatial reasoning*. The requirements for spatial reasoning change from one domain to another. For example, in certain applications, simple bounding boxes may suffice for shape representation. In other applications, more sophisticated polygons, or splines, may be necessary. Thus, the interpretation associated with spatial primitives such as "left of" may change from one domain to another.

Conclusion

In spite of its limitations in robust image interpretation, the current version of the Show&Tell prototype has received positive feedback from the intelligence analyst community for its demonstrated advantages and viability. Even so, the judicious combination of semi-automated and automated techniques allows for the design of a robust system. We expect Show&Tell to realize major significance in the area of digital libraries. **MM**

Acknowledgments

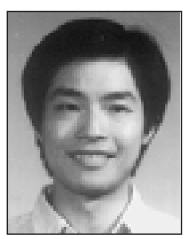
This work was supported in part by DARPA Contract 93-F148900-000.

References

1. H. Clark et al., "Common Ground and the Understanding of Demonstrative Reference," *J. Verbal Learning and Verbal Behaviour*, Vol. 22, No.2, 1983, pp. 245-258.
2. W. Wahlster, "User and Discourse Models for Multimodal Communication," *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, eds., Morgan Kaufmann, San Francisco, 1998, pp. 359-370.
3. O. Firschein and T. Strat, eds., *Radius: Image Understanding for Imagery Intelligence*, Morgan Kaufmann, San Francisco, 1997.
4. B. Bremner et al., "Integration of Image Understanding Exploitation Algorithms in the Radius Testbed," *Proc. DARPA Image Understanding Workshop*, Morgan Kaufmann, San Francisco, 1996, pp. 255-268.
5. M. Kass et al., "Snakes: Active Contour Models," *Proc. First Intl. Conf. on Computer Vision*, IEEE CS Press, Los Alamitos, Calif., 1987, pp. 259-268.
6. P. McKevitt, ed., *Integration of Natural Language and Vision Processing* (Volumes I-IV), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.
7. R. Srihari, "Use of Collateral Text in Understanding Photos," *Artificial Intelligence Review*, Vol. 8, No. 5, 1995, pp. 409-430.
8. R.K. Srihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images," *Computer*, Vol. 28, No. 9, Sept. 1995, pp. 49-56.
9. R. Mohr and T. Henderson, "Arc and Path Consistency Revisited," *Artificial Intelligence Review*, Vol. 28, No. 2, 1986, pp. 225-233.
10. R. Chopra and R. Srihari, "Control Structures for Incorporating Picture-Specific Context in Image Interpretation," *Proc. Intl. Joint Conf. Artificial Intelligence (IJCAI 95)*, Morgan Kaufmann, San Francisco, Aug. 1995, pp. 50-55.
11. C. Jaynes et al., "Task Driven Perceptual Organization for Extraction of Rooftop Polygons," *Proc. Image Understanding Workshop*, Morgan Kaufmann, 1994, pp. 359-365.
12. Z. Zhang and R. Srihari, "Qualitative Building Detection from Monocular Images Based on Linguistic Context," *Proc. Vision Interface,* Canadian Image Processing and Pattern Recognition Society, Vancouver, Canada, 1998, pp. 1-8.
13. H. Niemann et al., "ERNEST: A Semantic Network System for Pattern Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 9, 1990, pp.883-905.
14. S. Oviatt, "User-Centred Modeling for Spoken Language and Multimodal Interfaces," *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, eds., Morgan Kaufmann, San Francisco, 1998, pp. 620-629.
15. P. R. Cohen et al., "An Open Agent Architecture," *Proc. AAAI Spring Symposium Series on Software Agents,* AAAI Press, Menlo Park, Calif., 1994, pp. 1-8.

Rohini K. Srihari is an associate professor of computer science and engineering at the State University of New York at Buffalo. She has worked in both natural language processing as well as computer vision. Her current research interests include multimedia information retrieval and multimodal image annotation systems. She received a B. Math in computer science from the University of Waterloo, Ontario, Canada and an MS and PhD in computer science from the State University of New York at Buffalo.

Readers may contact Srihari at the Department of Computer Science and Engineering, State University of New York at Buffalo, UB Commons, 520 Lee Entrance-Suite 202, Amherst, NY 14228-2583, e-mail rohini@cedar.buffalo.edu.



Zhongfei Zhang is an assistant professor in the Computer Science Department at the State University of New York at Binghamton. He was a research scientist in the Center of Excellence for Document Analysis and Recognition (Cedar), at SUNY at Buffalo before he joined the faculty at SUNY at Binghamton. His research interests include multimedia information indexing and retrieval, image understanding, and pattern recognition. He has published extensively in national and international journals and conferences in these areas. He received a BS in electronics engineering and an MS in information sciences both from Zhejiang University, China, and a PhD in computer science from the University of Massachusetts.

Readers may contact Zhang at the Computer Science Dept., Watson School of Engineering and Applied Science, State Univ. of New York at Binghamton, Binghamton, NY 13902-6000, e-mail zhongfei@cs.binghamton.edu.