Editor: Harrick Vin
University of Texas at Austin

Annotation and Education

Brian K. Smith, Erik Blankinship, and Tamara Lackner
MIT Media Lab

Because our research focuses on developing multimedia technologies for educational settings, we begin this article with an exercise. Throughout this article, you'll see a number of screen images from programs we're building. Try to understand what these pictures mean without reading the explanatory captions that appear below them. You'll probably find this difficult, as you're still unfamiliar with our work. The important thing to realize is that a well-written caption can provide a context for understanding. Without such an explanation, the visual information is almost meaningless. This may seem obvious, but let's continue with the exercise for a moment.

If you can't understand the images without reading the text, find a few of your colleagues and ask them to join you in this exercise. You'll likely have different opinions than your colleagues, and these differences might lead to arguments about the content. You may find yourself collaborating with others to create your own explanations of the imagery. Even if your collective hypotheses differ from our intentions, the process of collaboration and argumentation around the imagery may lead you to interesting insights about our projects.

Text captions (like the ones you've been ignoring if you've been playing along with the exercise) play an important role in education, focusing students on salient features of illustrations. The narratives that accompany video presentations are no different. The narrator of a documentary film presents viewers with a story that focuses attention on important information. In a sense, the visual information is far less important than the accompanying explanations. The captions and narratives give you a right answer, they tell you what you should learn from the figures.

But learning isn't always about the right answer. Often, it's about determining what you could learn from a text, photograph, or video clip. We often want learners to engage in a process not unlike the one you went through with your colleagues, a process of active observation and interpretation involving collaboration, argumentation, and critique. Rather than developing multimedia systems to efficiently deliver information to learners, we may want to design social interactions around multimedia tools, interactions that lead people to construct and defend their own hypotheses about events and processes.[1]

While we don't want to completely eliminate captions and narrations from photographs and video, we suggest that excluding them can lead to pedagogical benefits. Our research explores the types of learning that occur when people collaborate to develop explanations of multimedia content. We develop systems to help learners observe digital photographs and video, pose hypotheses about their meanings, and justify their assertions with evidence. In most K-12 classrooms, imagery is used as information, as a means of presenting facts about the world. We're trying to change this by developing tools to help learners use imagery as data for observation, interpretation, and argumentation.

Annotation as argumentation

Visual events are rich with opportunities for learners to pose their own questions and hypotheses about interesting events and processes. One of our goals is to get people to look beyond explanatory captions and narratives and to ask how and why questions about visual imagery. You can imagine students watching a nature film and later asking questions about the content of the narrative. When did they say that chimpanzees hunt? How fast does a cheetah run? Contrast these with questions that go beyond the content to seek additional explanations or to resolve discrepancies in knowledge. So why does the chimpanzee only hunt in the wet season? Why does a cheetah, unlike other cats, need to run so fast? These latter "wonderment" questions show a desire to extend knowledge,[2] yet they often go ignored when people simply turn to a narrative for answers.

1070-986X/00/$10.00 © 2000 IEEE

We've tried to build environments that encourage learners to pose and investigate wonderment questions. Each of the tools we describe has an annotation component, a way to explicitly connect questions or hypotheses to visual media. In most multimedia research, annotation is associated with search and retrieval of documents. How can an object be described such that users can find it? In our systems, similar descriptions are treated as part of the learning process. How can a process or event be explained through a set of indices? Through annotation, we try to help learners articulate more than content summaries—we try to help them make their thoughts explicit so that they can discover holes in their reasoning.

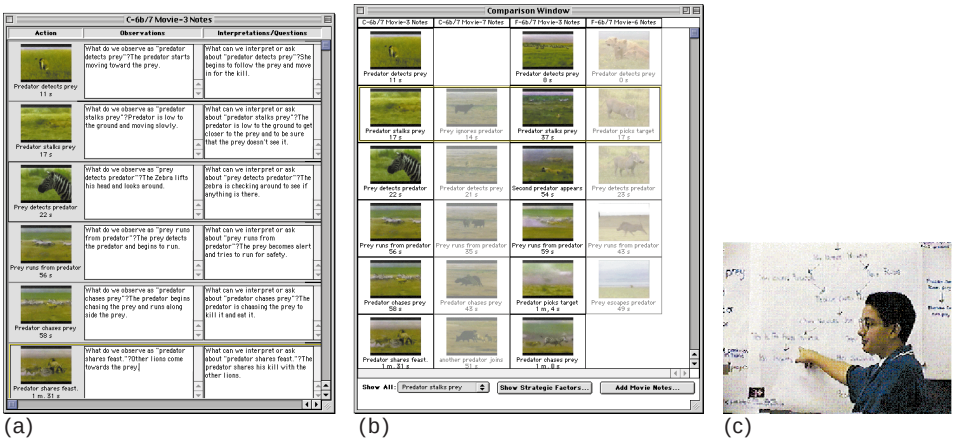

Figure 1. (a) Annotation, (b) comparison, and (c) modeling with Animal Landlord.

Annotation is about constructing ontologies to describe the world. When these ontologies are made explicit to learners, we suspect that interesting pedagogical outcomes will emerge. Instead of just watching nature films, for instance, we can have students describe the outcomes of filmed events and the intermediate states leading to the outcomes. Ordinarily, the narrator might explain these actions for the viewer. When students have to identify and annotate features of the video that seem to answer particular questions, they're performing tasks that resemble authentic scientific practice. That is, like scientists, they develop classification schemes to compare and contrast data. These classifications form the basis for developing models of behaviors and processes. More importantly, the annotation process has students developing their own explanations of visual materials.

If students are to rely less on explanatory captions and narratives and develop their own explanations and questions around images and video, we should make annotations visible. In two of our systems, learners construct ontologies to explain multimedia content. In the third system, "experts" annotate the content, but the annotations are made visible to students so that they can reflect on the underlying justifications (see the section, "The Parent Trap"). Getting students to "read between the lines" means adding additional context to the content, providing opportunities for them to question and explain for themselves.

In the systems described below, we'll show how annotation can lead to learning interactions rarely seen in classrooms and homes. By wrestling the annotation task away from computing professionals and placing it into the hands of learners, we intend to generate new ways for people to learn from multimedia content.

Animal Landlord

Animal Landlord is a video annotation system originally developed for high-school biology classrooms. Students explore issues in behavioral ecology using nature films as data. The initial curriculum focused on the hunting behaviors of the Serengeti lion. A second unit explores the relationship between conservation biology and animal behavior. In both cases, the narrations that typically accompany nature films have been removed, making the students responsible for annotating important events and assembling a story of how and why the animals in the films behave.

The initial task is to analyze behaviors in the videos and describe the intermediate actions that lead to outcomes. Groups of three to four students collaborate around an annotation tool to label important actions, their reasons for selecting the action, and any interpretations that can be drawn from the action (Figure 1a). For instance, a group of students might mark a video frame with the action "Predator picks target" because they see "the lion looking intensely at its prey." They would also make predictions or inferences about the reasons for this action (for example, "the lionesses probably chose the fat one because it can provide the most meat"). These actions form the video's plot structure. Students use the additional justifications to defend their plot structures in later class discussions.

The students create indices to actions in the video corpus. Once a classroom has indexed the entire corpus, the films can be compared to look

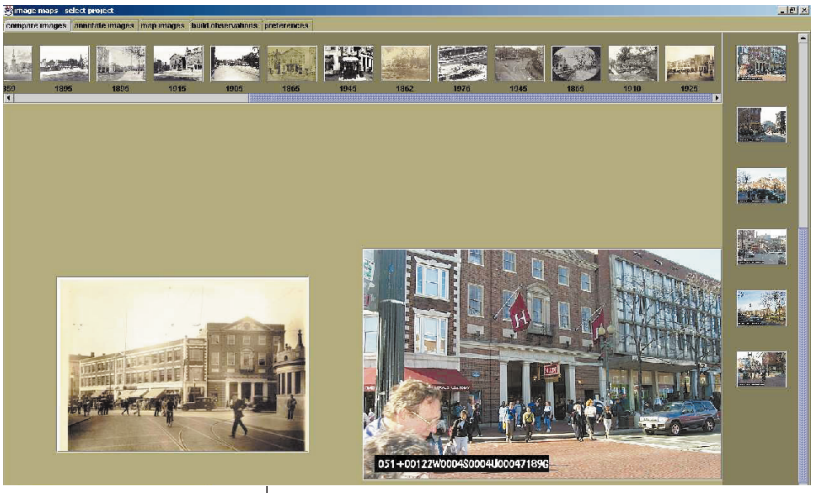

Figure 2. Retrieving historical images with Image Maps.

for similarities and differences between events. For instance, a lion is much slower than a zebra, therefore it needs to be careful when stalking one. On the other hand, it need not be as careful when stalking the large, slow Cape Buffalo. A graphical comparison tool lets students view all video frames indexed as "Predator stalks prey" to look for these sorts of variations (Figure 1b). Students also use this comparison tool to identify factors (such as type of prey, amount of ground cover, number of predators/prey) that may explain outcomes. By identifying these factors, students can begin to construct evolutionary accounts of behaviors.

The final task is to assemble annotation labels into decision trees that show all possible paths that predator and prey can take during a hunting episode (Figure 1c). These decision trees mark the end of a process that resembles the expert practice of behavioral ecologists: make observations of the world, decompose them into relevant actions, compare these actions across episodes to look for variations, and, finally, assemble them into a predictive model. When these trees are created and posted around the classroom, students argue their validity in terms of the annotations and comparisons they've made. For instance, many arguments arise around the node "Predator ignores prey." Why would a predator suddenly ignore its prey? This leads to discussions about energy consumption, the trade-offs between chasing and waiting for other, potentially more vulnerable prey to arrive, and so on. In other words, the students argue about the evolutionary consequences of decisions made by predator and prey.

When other predator-prey videos are shown in class, students return to their decision trees to see if their models still hold. If they find new information, they update their models much as scientists revise their theories in light of new evidence. Thus, the product of Animal Landlord persists and evolves beyond the initial intervention.

Image Maps

Our second project, Image Maps, applies annotation to the domain of history and urban planning using digitized photographs as the primary data source. As with Animal Landlord, students ask how and why questions about patterns that they find in imagery. In particular, they try to understand how and why their local communities have changed over time. The goal is for students to articulate features of their neighborhoods and argue about their influences on community change.

We want students to see the history of buildings in their communities. To make these histories explicit, we added a Global Positioning System (GPS) and a digital compass to a digital camera—this allows us to embed position and orientation metadata into the header of a JPEG image when it's captured. When students download photographs they've taken, the Image Maps software uses the metadata to search city maps stored in a geographic information system (GIS). By tracing a line along the orientation vector from the camera position, we can locate the first intersected building and return a set of historical images indexed to that site (Figure 2).

The thumbnail icons on the right of Figure 2 represent photos that students have taken. When one of these is clicked, it displays at right center and calls the search engine to find pictures of the same location. The retrieved images appear at the top of the screen, rank-ordered by the year the image was taken. At left center is an enlarged image from the historical set, a picture of Harvard Square taken during the 1920s.

Having historical photographs at your fingertips is interesting, but the classroom activity centers around making sense of the photographs. While asking questions about how and why their communities have evolved, students begin labeling the images with features that change over time. For instance, some students may focus on transportation, marking some images with "horse-drawn carriage" tags, others with "automobile" tags. Another group of students might be interested in land use, describing buildings with features like "commercial," "industrial," and so on. Figure 3 shows the annotation interface and an ontology being constructed by a group of students.

As students generate more annotations, they

use the labels to compare images and think about similarities and differences. More importantly, they can begin building models to explain how and why their local communities have changed over time. The models that they construct build on the architectural patterns described elsewhere.[3] A problem/theme is chosen ("crosswalks for people"), the context for the problem is described (pedestrian traffic is conflicting with transportation), and solutions are provided in the form of historical images. In the crosswalk case, students would construct a causal chain illustrating the progression from unmarked pavement to marked crosswalks. After constructing a number of chains, they can return to the field to see how well their generalizations hold up in unexplored parts of the city. That is, the exercise doesn't conclude with a single community outing; we expect students to iterate on their hypotheses. For instance, if they think that road layouts in Harvard Square were rearranged to minimize traffic flow, they need to return to the location to discover how traffic was rerouted. Taking additional photographs in the present leads to historical pictures that may help them discover the answer behind traffic routing issues.

As with Animal Landlord, students use image data to create models of behavior; in this case, the behaviors are changes in a community over time. Students will collaborate and argue around these data to develop hypotheses about change. For instance, a class can be divided into groups where each one studies a sector of the city. As a class, they can assemble a more complete model of community change than a single group could on its own. We also imagine that much discussion and debate will revolve around the causal chains that students produce. Teachers will be responsible for helping students adopt sensible investigation methodologies as they go into the world to collect their data and to moderate arguments around their hypotheses.

The Parent Trap

We often assume that television is a passive medium, one where viewers learn by absorbing audio and visual information. This isn't entirely true, since viewers must actively interpret what they see in light of what they know. But evidence exists that children gain more from educational television when they "co-view" programs with adults and other siblings[4] and can discuss content with others. In our newest system, the Parent Trap, we move from classrooms to homes to understand how to raise the level of parent-child discourse around educational television.

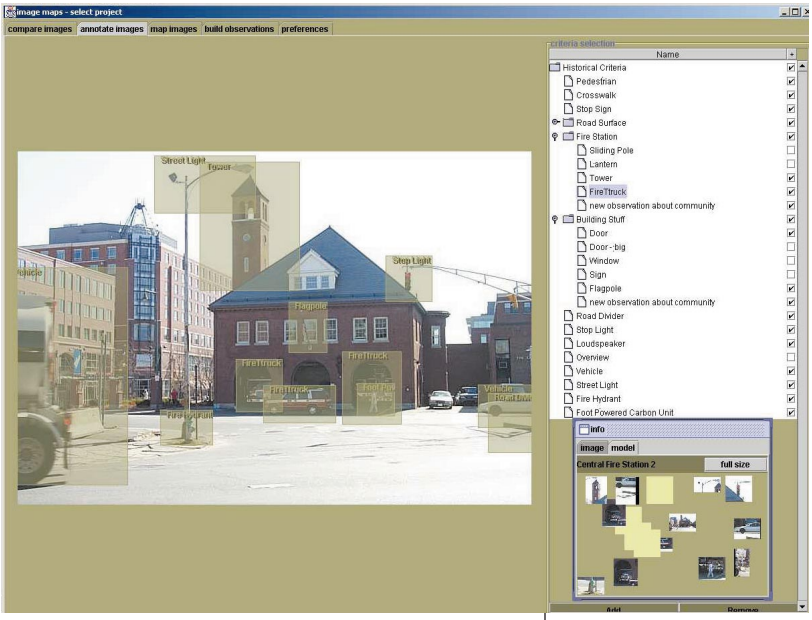

Figure 3. Annotating images with Image Maps.

Although parents and children frequently view television together, parents are more likely to offer evaluative comments ("Why are we watching this?") than to engage children in questioning, explaining, and critiquing content.[4] That is, instead of helping children to ask and investigate wonderment questions, parent-child conversations often summarize content without extending knowledge beyond the program. In classroom settings, various strategies have been identified that help children generate questions and predictions, identify factors related to these questions, and evaluate alternative hypotheses.[5] If parents and children had access to these strategies, they might be able to use them to extend the types of questions and explanations being offered during co-viewing.

Rather than giving parents and children an abstract strategy guide, Parent Trap models inquiry strategies by delivering program-specific, wonderment questions. By modeling inquiry questioning for parents and children, we hope to help them understand how they might begin asking similar questions before, during, and after television viewing. The best way to illustrate this is by example. Imagine that a child is at home watching today's episode of the popular children's program, *Bill Nye: The Science Guy*. This episode is about forests, and the child discovers how sunlight and rain make forests grow and how a tree's age can be determined by counting the rings on its trunk. When a section with singing trees

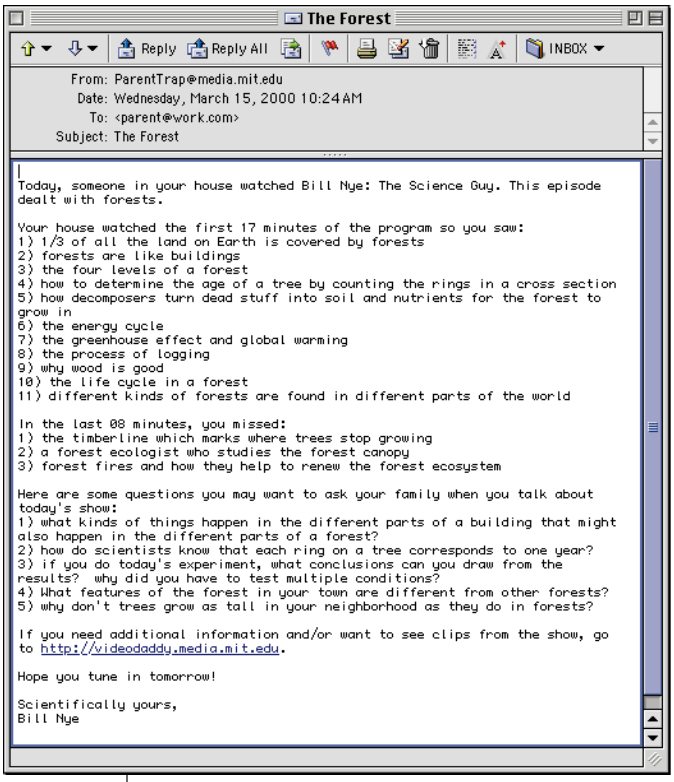

Figure 4. An e-mail message generated and sent by the Parent Trap application after a child has watched an episode of an annotated television program.

begins, the child changes the channel to view another program.

The "television" that the child has been watching is actually the Parent Trap's video player, a Quicktime streaming video object. When a program ends or is interrupted, the player records the completion time and sends it to a server. From this timestamp, we determine what pieces of the program have and haven't been viewed. This means we can create two lists:

1. the issues that the child should have seen while viewing, and

2. the issues that the child missed by changing to another program.

We also compile a list of questions that parents and children can discuss together, the wonderment questions. For instance, we know that the child in our example saw the segment dealing with the age of a tree, so we instantiate a wonderment question that asks, "How do you think scientists discovered that each ring on a tree equals one year of its life?" These three lists—what did you see, what did you miss, and what more could you ask about the program—are assembled and sent as e-mail messages to parent and child (Figure 4). Personalized Web pages are also created dynamically so that parents and children can browse the relevant video clips for each issue/question generated by the Parent Trap.

In Animal Landlord and Image Maps, students were responsible for annotating content; in Parent Trap, content designers add the annotation layer to their television programs. Behind each show is a set of time-coded, Extensible Markup Language (XML) annotations that generate the messages described above. Content designers use an annotation tool (Figure 5) to mark up their video clips, and this tool tries to prompt them to think about questions that would be useful to parents and children. For instance, we ask them to list the core educational issues for each scene in the video. In a typical children's program, there tends to be a core theme (such as forests) and a number of subthemes that children are expected to learn (such as the age of a tree and the effects of logging on forests). Making these explicit lets us generate the list of issues in the e-mail messages and Web pages.

In addition to the core issues, designers are asked to think beyond their content. This might mean questions about the following:

1. *Alternative viewpoints*. A segment of a program showing a logger discussing the value of trees to make furniture, paper, and so on might deserve additional annotation. A designer might include an alternative viewpoint question: What would an environmentalist say about cutting down trees?

2. *Experimental design*. Some shows have segments that ask viewers to try experiments at home. An experiment that has viewers place celery stalks into colored water to "simulate" water flowing through a tree's trunk might require additional questions. For instance, "Why do you need to have three conditions to understand what is happening in the tree experiment?" Such questions help parents and children reflect on the nature of science and that experimental methodologies frame the hypotheses developed by scientists.

3. *Justification*. Content is chosen for television programs, but viewers aren't privy to why various issues are thought to be important. We ask our designers to provide design rationales for each of their learning issues, helping parents and children see the importance of what

they're learning. More, we hope that these rationales will let learners reflect on and question the educational value of the content they're viewing.

While the target audience for the Parent Trap appears to be parents and children, we're also interested in helping content designers reflect on their practices. While we can't change the ways that television programs are created, we hope to make an impact by helping producers reflect on their content and to think deeply about ways to help parents and children converse and learn from educational television.

Future directions

Animal Landlord continues to be used in classrooms. Studies of its use in schools show promising results, as students seem to develop more sophisticated, causal arguments during and after the annotation, comparison, and modeling tasks. Image Maps is being prepared for deployment to students this year, and we will be studying its use to see whether it can produce similar results. The Parent Trap is currently a demonstration, but we're conducting studies in schools to understand how teachers work with television and how they prompt students to generate and investigate wonderment questions. These studies will be used to refine the prompts that television producers see as they annotate their content.

The common threads behind these applications are

- making annotations explicit to learners to help them generate questions, test hypotheses, and refine arguments, and
- creating software and social environments that encourage collaboration and argumentation around multimedia content.

Rather than simply providing information, we see great value in tools that facilitate the process of question asking and hypothesis testing. Most of our future work will involve testing these systems with learners to see if we can change the ways that they think and learn. MM

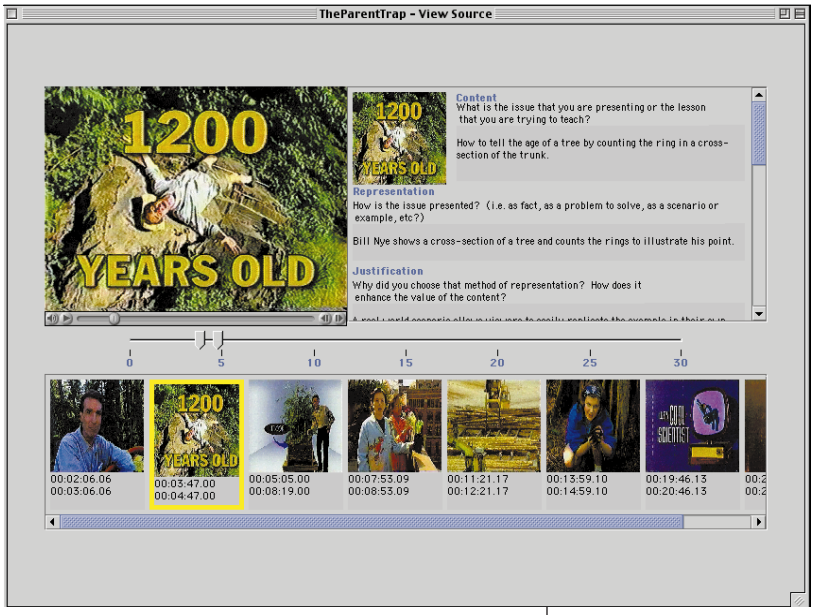

Figure 5. Annotating television programs with Parent Trap.

References

1. R.D. Pea and L.M. Gomez, "Distributed Multimedia Learning Environments: Why and How?," *Interactive Learning Environments*, Vol. 2, No. 2, 1992, pp. 73-109.
2. M. Scardamalia and C. Bereiter, "Text-Based and Knowledge-Based Questioning by Children," *Cognition and Instruction*, Vol. 9, No. 3, 1992, pp. 177-199.
3. C. Alexander, S. Ishikawa, and M. Silverstein, *A Pattern Language*: *Towns*, *Buildings*, *Construction*, Oxford Univ. Press, Oxford, 1977.
4. A. Dorr and B.E. Rabin, "Parents, Children, and Television," *Handbook of Parenting: Applied and Practical Parenting*, M.H. Bornstein, ed., Vol. 4, Lawrence Erlbaum Associates, Mahwah, N.J., 1995, pp. 323-351.
5. A. Collins and A. L. Stevens, "Goals and Strategies of Inquiry Teachers," *Advances in Instructional Psychology*, R. Glaser, ed., Vol. 2, Lawrence Erlbaum Associates, Hillsdale, N.J., 1982, pp. 65-119.

Readers may contact the authors at MIT Media Lab, 20 Ames Street, E15-001, Cambridge, MA 02139, e-mail {bsmith, erikb, tlackner}@media.mit.edu.

Contact Project Reports editor Harrick Vin at the Department of Computer Science, University of Texas at Austin, Taylor Hall 2-124, Austin, TX 78712-1188, e-mail vin@cs.utexas.edu.