Digital Libraries: Meeting Place for High-Level and Low-Level Vision

Rosalind W. Picard

MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139 USA
picard@media.mit.edu
http://www.media.mit.edu/~picard/

Abstract. The average person with a networked computer can now understand why computers should have vision – to search the world's collections of digital video and images and "retrieve a picture of ___." Computer vision for intelligent browsing, querying, and retrieval of imagery is needed now, and yet traditional approaches to computer vision remain far from a general solution to the scene understanding problem. In this paper I discuss the need for a solution based on combining high-level and low-level vision, that works in concert with input from a human user. The solution is based on: 1) Learning from the user what is important visually, and 2) Learning associations between text descriptions and visual data. I describe some recent results in these areas, and overview key challenges for future research in computer vision for digital libraries.

1 Introduction

Collections of digital imagery are growing at a rapid pace. The contexts are broad, including areas such as entertainment (e.g. searching for an actress washing her hair), education (e.g. finding illustrations of pilgrims), science (e.g. analyzing satellite imagery), medicine (e.g. comparing onset rates of osteoporosis), marketing (e.g. insuring the competitor hasn't used leopards in their ads), and design (e.g. selecting an oriental rug). In all these applications and more, vision tools can facilitate access to content.

1.1 Vision, signal processing, and common sense

Computer vision is not a solo solution to the problems of retrieval and annotation in image and video libraries. There are numerous signal processing issues such as compression, and analysis of accompanying non-visual signals such as the soundtrack. There are also business issues (e.g. billing for downloading data), database issues (e.g. fast algorithms for indexing), interface issues (e.g. visualization and manipulation of multiple video streams), and numerous other research problems. Natural language queries and common sense systems also play a significant role; for example, Lenat's CYC common sense reasoning system can take a request such as "find someone wet," and return an image with a label such as "man finishing a marathon" [1]. The computer vision solutions will work best if wisely integrated with solutions from these other domains.

In [2] I overviewed the latest digital library research issues for the image processing community to address. The emphasis for that community is on finding models for simultaneous compression and content description, and improving measures of visual similarity for comparing images. However, the domains of the image processing and computer vision communities overlap, so it is wise for them to watch what each other does to insure complementary efforts and avoid wasteful duplication of errors, especially in spatiotemporal segmentation and modeling where research seems to overlap the most. In this paper I will focus on important research problems that I did not discuss in [2], problems which are traditionally closer to the computer vision community – namely, generating descriptions of image content, and vision systems that learn.

Although my focus in the rest of this paper is on computer vision for digital libraries, much of the following discussion also applies to other perceptual domains such as acoustic scene analysis. Furthermore, construction of vision tools (such as a visual thesaurus) should not proceed in isolation, but in concert with the other perceptual domains (e.g. audio thesaurus and text thesaurus), since humans, the ultimate judges of these systems, make associations which weave through all the perceptual senses.

1.2 The importance of bias in learning

Saying someone is "biased" is usually not a compliment; however, bias ultimately is what leads someone to a conclusion, right or wrong. Bias can not only be favorable, but it can also be essential for good performance, especially when the number of possibilities is large. The word "bias," as used in the rest of this paper, refers to that which can be controlled, and which guides a system to its answer.

Consider a system where all solutions are equally likely, and a solution is pursued by an optimization algorithm. If the algorithm falls into local minima, or can be varied to provide different equally optimal solutions, then the algorithm itself is a type of *procedural* bias. If the space of the optimization contains regions with different likelihoods, then these likelihoods are a form of *declarative* bias. In fact, the initial constraints established to set up the problem can be considered a form of declarative bias, ad infinitum. Moreover, the distinction between declarative and procedural is not firm; a procedure's text is also a "declaration." Bias can be anything – priors, weights, procedures, criteria, the space of possibilities permitted – which the designer gives the system for use in choosing its answers.

In theory, each 2-D image has an infinity of 3-D images which could have given rise to it. In visual recognition and retrieval where the space of possibilities is astronomical, a system will effectively never return from its search for a solution unless it has some bias.

In fact, there is recent neurological evidence supporting the role of emotions as a critical biasing mechanism in human decision-making [3]. The evidence indicates that humans with a particular kind of brain damage do not have these emotional biasing mechanisms, and consequently suffer a loss of decision-making ability. When asked, say, to find a good time to schedule an appointment, they disappear into an endless space of possibilities and cannot reach an answer on their own.

2 Combining low- and high-level vision with learning

Traditional low-level vision consists of the underlying representations used for the image and video – edges, color, texture, etc. The stored feature values comprise a low-level declarative bias, which is used upstream to make decisions.

Higher-level vision is usually concerned with interpretations of low-level features. Consider the selection of a recognition procedure that is well-suited to labeling one of the categories in a scene. The available procedures, and the mechanism(s) for their selection comprise what may be considered a high-level bias.

Successful biases must happen at both the low and high levels, and depend both on the data and the goals. With retrieval systems, where a user is in the loop, it is impossible to specify all the goals in advance. A key challenge for vision researchers is to develop a system with a good bias, that also knows how to intelligently change this bias as the goals and data change.

The strategy outlined in this paper addresses the problem of developing such a learning vision system, that learns its biases continuously. It provides a powerful solution for digital libraries and other domains where the following two Learning Criteria are met:

1. The system can pre-compute low-level biases and can analyze its own learning performance. These may be performed as *offline learning.*
2. The user of the system is present, and interacts with the system. Ideally, this allows *online learning.*

2.1 Teaching the system to learn

In traditional pattern recognition, the designer of the system runs extensive experiments comparing different features of the data and different recognition strategies, until finally the best combination of features and decision-making algorithms is found. Usually this process is conducted for a specific type of data, and a specific set of goals.

In digital libraries, there is not time for the researcher to follow this traditional paradigm for *every* new set of data, for *every* set of recognition goals. In other words, it is now necessary to automate more of the iterative learning process usually done by the researcher.

Minka and Picard have built a system, "FourEyes," which satisfies the two Learning Criteria above. Although I refer the reader to [4] for details of how it learns, how it compares to other learning systems and how its performance has been evaluated, I will highlight a few of its features below to illustrate the

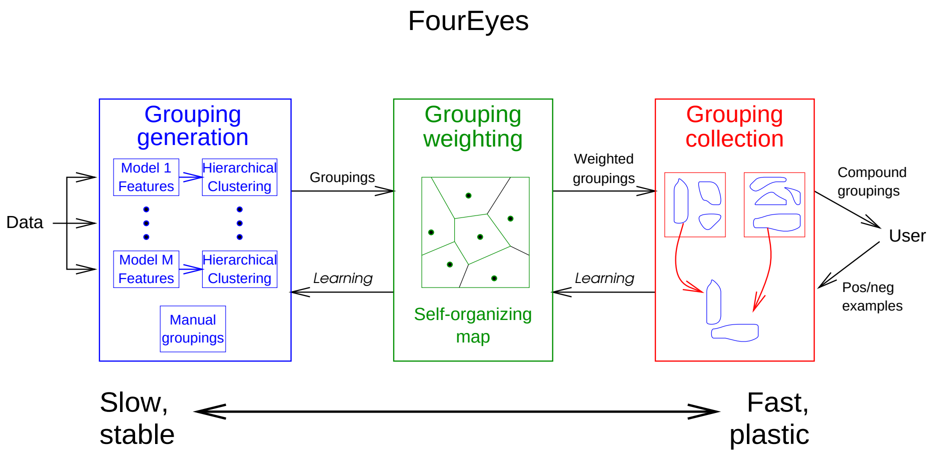

Fig. 1. The FourEyes learning system. The left-most box computes *groupings* for all the data in the database, learning offline. The right-most box interacts at run-time with the user, determining which groupings or combination of groupings best represent the data that interests the user, learning online. The middle box learns about learning, so to speak. It replaces a bias which would typically be procedural with one which is declarative, and therefore easier to update for changing the longer-term behavior of the system. For example, if the system sees the same problem repeatedly, then the middle box enables it to become faster at solving that problem, and problems similar to that problem. The middle box provides a way for a learning system to select and modify its own bias.

arguments in this paper. A diagram of the FourEyes learning system is given in Fig. 1.

FourEyes was originally constructed to assist the user in annotating, or attaching text descriptions, to databases of imagery. Because it is tedious for a user to label every region in every image, annotation is a good task for machine vision. However, because people segment and label regions differently depending on a variety of factors, it is not a task that has a single optimal solution. Instead, a system is needed that learns how to classify (annotate) regions the way that a particular user is annotating them.

2.2 Too much bias...too little bias...

Traditional segmentation and classification systems rely on a single model and set of decision-making rules. The results, such as for segmentation on satellite imagery, work well under carefully controlled conditions for the categories the system is trained to recognize, but can give wildly incorrect solutions when the controls are violated, such as when data outside those categories is shown to the system (an admittedly unfair test for the traditional purpose, but one which arises in the new diverse world of digital libraries). Such systems may be said to be over-biased.

Most of the latest segmentation and labeling systems try to improve flexibility by using one powerful model such as a Markov random field for the segmentation, and then allowing for model parameters to vary within regions, and for labels to be associated with particular parameter ranges. Posed probabilistically, these systems tend to be doubly-stochastic with astronomical computational demands for their simultaneous parameter estimation and segmentation. Because they allow too large a space of solutions, they rarely find optimal solutions, and rarely converge except for trivial problems. In short, they do not have enough bias.

In contrast with traditional approaches, the FourEyes system does not have a fixed bias. Instead, it is able to select and modify its bias, taking on some of the iterative learning that has usually been performed by the designer of a fixed-bias system. This new emphasis on a changing bias is an important research focus for digital libraries.

2.3 Associating text with visual features

Before proceeding, let's consider an example of retrieval in the context of video libraries of television shows. Consider the request, "Find comedians taking lie-detector tests." Like most requests, this one starts out with language, and will rely heavily on language-processing retrieval tools such as the caption-based retrieval tools of [5], [6]. However, the text had to get there somehow for these tools to be of use. Let's examine first some of the kinds of high-level text that tend to be available to help vision algorithms, and second, the use of vision-algorithms for helping generate text that can make the search easier.

2.4 Scripts and closed captioning

Figure 2 shows two frames from the "Seinfeld" TV series episode "The Beard." The frame at left shows the closed-captioning, where Cathy is saying "a polygraph." Figure 3 shows excerpts from the script corresponding to the two frames shown in Fig. 2.

The script can be used in several ways to simplify vision-based retrieval. First, the script tells which main actors will appear in a scene. Although they will not necessarily appear in every shot in that scene, this information can be used to bias the algorithm to try to verify first those actors listed in the script. Second, the script gives the location of the scene. Although this does not imply the backgrounds in each shot will be the same (cf. the two frames in Fig. 2) it is likely that certain objects in the scenes will reappear (cf. the green and white walls in the police station) and therefore become recognized as a feature associated with that set. Sometimes the script gives action information, although eye-catching actions are not always annotated in the script, but may be ad-libbed.

Closed-captions, intended for hearing-impaired viewers, can also be used to help vision algorithms associate actions with actors. The closed-captioning text is

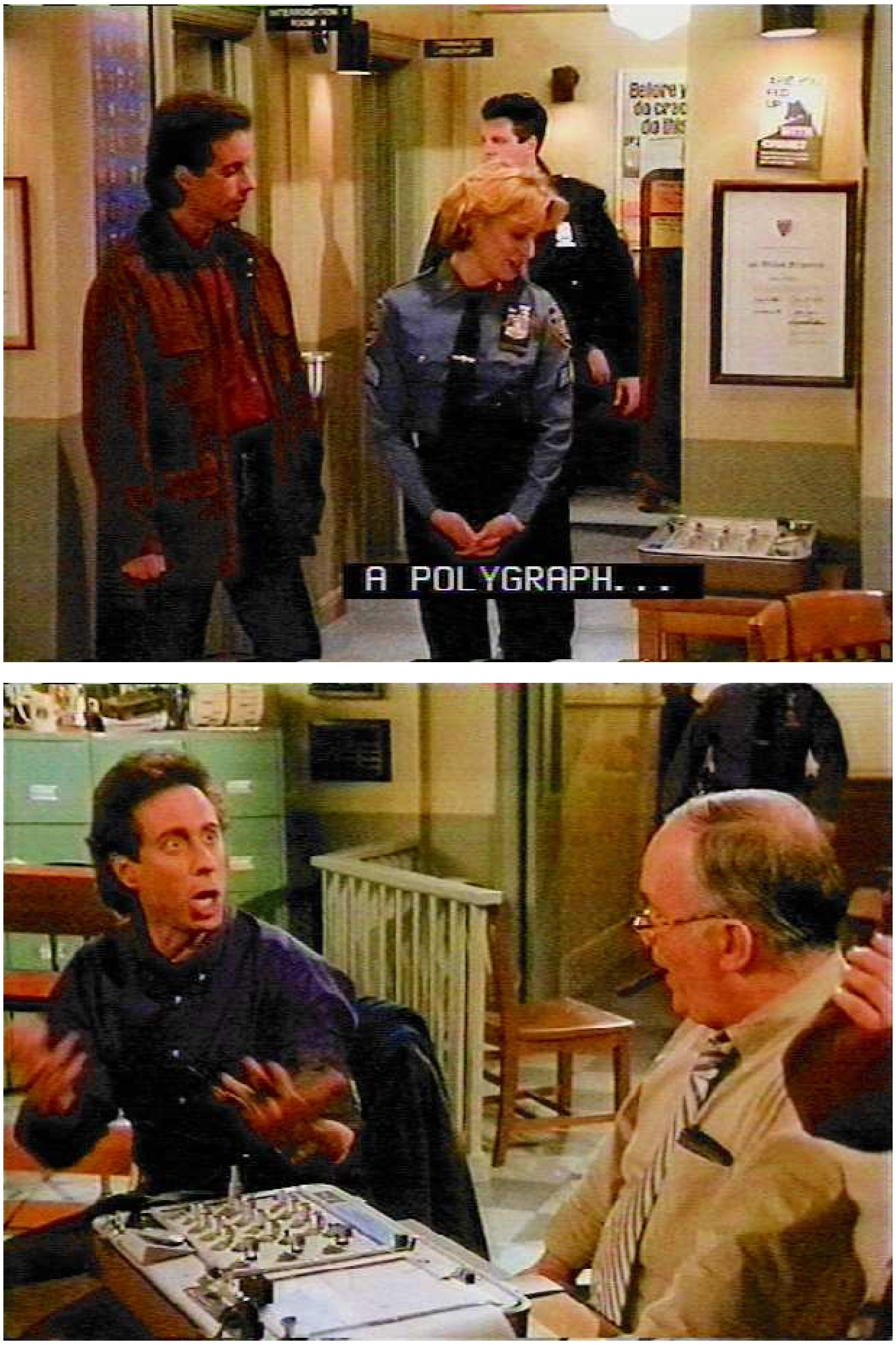

Fig. 2. Top: Shot where script mentions a polygraph. (See script excerpts in Fig. 3.) Note that the position of the closed-captioning text indicates which speaker (Cathy) is talking. Bottom: Shot where Seinfeld is hooked to a polygraph but the word "polygraph" and its synonyms such as "lie detector" do not appear anywhere in the scene containing this shot. Hence, script-based retrieval can get us near the answer, but vision is needed to find the answer. These images from "The Beard" episode of "Seinfeld" appear courtesy of Castle Rock Entertainment.

```
(Jerry, Cathy, Lou)
ACT ONE, SCENE N
INT. POLICE STATION HOUSE - DAY (2)
JERRY WALKING AROUND STATION HOUSE WITH CATHY.

JERRY
What's that?
CATHY
Polygraph.  What you civilians call a lie
detector test.
JERRY
Oh, Alright.  Let me ask you now when
someone is lying, is it true that their
pants are actually on fire?
CATHY
If I could tell you the famous faces that
have been up here.
...


(Jerry, Gus, Cathy)
ACT TWO, SCENE Y
INT. POLICE STATION - DAY (4)
JERRY IS WIRED UP TO A MACHINE THAT IS BEING
RUN BY GUS.  CATHY SITS NEXT TO HIM.

GUS
What's your name?
JERRY
Jerry Seinfeld.
GUS
What is your address?
JERRY
129 West 81st Street.
GUS
Did Kimberly steal Joe's baby?
...
```

Fig. 3. High-level information is available to work with low-level vision tools in most video productions today. The script says where and when the scene is set, and which main actors are present, but otherwise says little about scene content. These excerpts from "The Beard" episode of "Seinfeld" appear courtesy of Castle Rock Entertainment.

almost always a subset of the script, but its placement onto the image gives additional information. For example, the text appears closest to the person speaking if they are in the shot, and closest to their direction off the frame if they are not. The decoded closed captions can be located in the script to verify the name of the actor seen speaking. (Alternatively, a computer vision algorithm that recognizes the actors could assist a human in placing the captions for the hearing-impaired viewers.) The closed captions also display information that may not be in the script such as "knock knock" when somebody is at the door, or "ha ha" when it is not clear visually that one of the actors is laughing. These captions can also significantly aid in acoustic scene analysis.

2.5 Retrieval with low- and high-level learning

Now, let's return to the example request, "Find comedians taking lie-detector tests." Suppose that after going through a text thesaurus and other database tools, the retrieval system has determined that this episode of Seinfeld contains both the lie-detector synonym "polygraph" and the comedian Jerry Seinfeld. However, in the only scene where Jerry and the polygraph are both present according to the script (Fig. 2, left frame), he is not being given the test.

Several options are available at this point – one could watch the whole episode to see if he uses the polygraph, one could make a new query, or one could give up. With the aid of a smart-fast forward mechanism one could save time and watch just the scenes with Jerry in them. Alternatively, the user could combine some low-level vision into the query at this stage, by taking the frame at the left in Fig. 2, indicating to the learning system that the silvery box at lower right is of interest (query by example) and then asking the vision system to find all shots where that box appears with Jerry. The learning system determines which groupings best characterize the polygraph in the scene where the user labels it, and then uses these groupings, plus any prior knowledge about the data, to search for additional (unlabeled) appearances. Although general recognition of a polygraph is nearly impossible without, perhaps, a visual database of such devices (they come in many shapes, colors, and sizes), recognition of this particular polygraph can be accomplished, and used to retrieve frames such as the one at right in Fig. 2.

Moreover, after the retrieval has succeeded, the high-level polygraph label can be linked to the successful low-level visual features and stored, so that future queries for that situation will converge more quickly to the answer.

Although this is just one example, it illustrates several important points. First, there is no need to do "bottom-up" vision to label every segment in every image – this problem is notoriously ill-posed. (e.g. What *is* that stuff behind Cathy and Jerry? How should the railing be segmented or should it not be?) For decades, researchers have tried to generate edge-detection and segmentation outputs that are "right" even though there is no segmentation that people would agree on. Low-level vision should not proceed without high-level input.

FourEyes's gets around this problem by collecting groupings (leftmost stage) from which multiple segmentations can be constructed. Only after the user interacts with the imagery, identifying regions of interest, do selections get made from the groupings, and receive labels.

Another point illustrated by this example is that most of what is in these scenes is unimportant compared to the actors. When a visual item is important, it is likely to either be named in the script at some point, centered in the camera, or gazed at by an actor; a high-level production mechanism is used to draw attention to it.

All of these high-level production mechanisms simplify the visual retrieval problem; their detection is another important challenge to vision researchers. However, their detection still does not solve the retrieval problem of finding queried content. Although closed captions associate speech with speakers, there is still no mechanism for associating labels with objects, actions, and overall scene mood. The associations linking high-level text with low-level image features still need to be learned.

The digital library is an important meeting place for high-level and low-level vision. Having the human in the loop, and working simultaneously with scripts, closed captions and low-level visual features, provides a unique environment for learning vision, especially for learning links between low-level features and high-level descriptions.

3 Summary

Digital video and image libraries are not only a new application area for computer vision, but they provide a unique opportunity to conduct research linking high-level and low-level vision.

Because retrieval tools for digital libraries involve a human user, vision algorithms in this domain have the opportunity to learn from human vision and interaction. I have mentioned one prototype system, "FourEyes," that attempts to do continuous learning, and have walked through an example of how it is applied to real problems, problems demanding the linking of text descriptions with visual features.

Retrieval systems that can also collect annotations and learn their links to the visual features, are not only of immediate use in applications, but they can potentially gather a great amount of useful data to researchers. In particular, the large number of mappings which result between relatively low-level visual features and high-level descriptions will be interesting to study for stability, context-dependence, and user-dependence.

The most important issue I hope to have drawn attention to is that of the need for more research on learning – and not directed to the one-time training of neural network weights or development of a set of discrimination functions, but to a system that learns continuously, switching its bias depending on the data and goals at hand, and updating that bias for long-term performance improvements.

Continuous learning for vision is a difficult area of research with far-reaching implications. Digital libraries provide a fertile arena for learning with a user in the loop, learning which high-level descriptions and low-level visual features are best associated, and ultimately learning how to retrieve the most relevant and interesting data.

4 Acknowledgements

FourEyes is the thesis research of Tom Minka, and the script-based retrieval examples are drawn from the thesis research of Joshua Wachman. I would like to thank both Tom and Josh for numerous fun discussions which have influenced my ideas in this area. I am grateful to HP Labs, BT PLC, and NEC, for their sponsorship of this research.

References

1. D. B. Lenat, "Artificial intelligence," *Scientific American*, pp. 80–82, Sept. 1995.
2. R. W. Picard, "Light-years from Lena: Video and image libraries of the future," in *IEEE Second Int. Conf. on Image Proc.*, (Washington, DC), Oct. 1995. To appear; Also appears as MIT Media Lab Perceptual Computing TR #339.
3. A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain.* New York, NY: Gosset/Putnam Press, 1994.
4. T. P. Minka and R. W. Picard, "Interactive learning using a 'society of models'," *Submitted for Publication*, 1995. Also appears as MIT Media Lab Perceptual Computing TR#349.
5. A. S. Chakravarthy, "Toward semantic retrieval of pictures and video," in *RIAO'94, Intelligent Multimedia Information Retrieval Systems and Management*, (New York), pp. 676–686, Oct. 1994.
6. R. K. Srihari, "Combining text and image information in content-based retrieval," in *IEEE Second Int. Conf. on Image Proc.*, (Washington, DC), Oct. 1995. To appear.