MPEG95/127, Lausanne.
M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 295.
Statement for Panel on "Computer Vision and Image/Video Compression", ICPR94, Jerusalem

Content Access for Image/Video Coding: "The Fourth Criterion"

Rosalind W. Picard

MIT Media Laboratory; 20 Ames Street; Cambridge, MA 02139
picard@media.mit.edu

"Minimal rate-distortion" is perhaps the shortest description of the quest of the image coding community. More precisely, there have been **three criteria to minimize**:

1. Bit rate
2. Distortion (ideally, perceptual)
3. Cost (computation, complexity of implementation.)

For over a decade people have tried to improve bit rates by pursuing "Nth generation" coding, or model-based coding. Model-based coding is where computer vision has played its primary role: finding edges, segmenting regions, modeling texture, objects, and optical flow. Research results for the last decade have indicated that model-based methods decrease the bit rate, but increase the distortion and computational complexity. By the time the visual distortion is reduced to look as good as a competing universal {DCT, VQ, QMF} configuration, the bit rate is only comparable to these waveform methods.

The model-based methods have not won with the above three criteria. Maybe they will win after another decade or two of research. But, I would like to propose that even with present rates, there is still a SIGNIFICANT potential win with the use of model-based methods.

The potential win comes with a new fourth criterion. The fourth criterion is due to the fact that the world has changed significantly since image compression researchers posed the compression problem. It used to be that all anyone wanted to do with image and video was efficiently STORE and/or SEND it:

Old scenario

picture → code → (STORE, SEND) → decode → *picture*

So optimizing with respect to the above three criteria was the right goal. No access was made to the pictures until after they were decompressed; it only mattered that they took up the least space possible.

But now, not only do scientists want to "search for volcanos on Venus," but art students want to find all "Kandinskys that have eyeballs" and kids want to fast-forward the video "to the dinosaurs." The world has changed. There are huge collections of digital images and soon there will be frequent attempts to access their content. I propose that now we need:

New scenario

picture → code → (STORE, SEND) → decode → *picture*
⇕ ⇕
CONTENT
ACCESS

Users will want the system to answer queries, find a particular image or video event, or change the content (colorize, replace actors, change viewpoint, enhance lighting, for example). In short, people want to query, retrieve, and manipulate the content.

Not only do people want to access content, but they should do so midstream when they can, where I show it above. Although fine detailed manipulation may not be suitable for midstream, the most frequent tasks, especially on large databases, should be performed midstream. Why midstream, vs. after coding and decoding, at *picture*? **Midstream access is better** because it:

* Has fewer bits to look through
* Saves time decompressing unwanted data
* Saves space storing decompressed data
* Has the same information as *picture* at the end

But accessing content midstream calls for a different representation of the image by the coder. The representation needs to make query, retrieval, and modification fast and easy. These tasks call for a model that parses the content into meaningful chunks, a model which gives the user "control knobs" that are perceptual and/or semantic. The knobs, or model parameters, allow the user to interact with the content in a natural and efficient way. A good model allows image query, retrieval, and modification to proceed on the compressed representation.

The demands of content access can be summarized by the introduction of a new FOURTH CRITERION:

4. Content access work.

Optimal compression in the new scenario requires minimizing "content access work." For query, retrieval, or modification, content access work measures the work needed by pattern recognition, computer vision, or image processing algorithms to search the content of an image (perhaps including its "index"), identify the region(s) of interest, and analyze or extract the requested information. (Extraction work may be doubled to include replacement work if modification is the aim.) Content access work can be measured both on the compressed data and on the decompressed data. The two can be compared, alongside the other three criteria, to determine which representation gives the best performance for the complete new scenario.

When we consider the new goal of accessing content, as opposed to merely making it take up less space, then an algorithm that wins in the old scenario may no longer be optimal. In other words, optimality will now be achieved by JOINTLY minimizing the first three criteria with the fourth – "content access work." Coding that optimizes all four criteria, "content coding," is where model-based compression appears likely to win. If rate-distortion levels are the same for model-based and waveform coders, then the winning coder is the one that minimizes content access work and total cost.

There is a fundamental challenge to both the compression and vision communities in this joint optimization: optimal representation and optimal discrimination can conflict. A famous example occurs with the use of the Karhunen-Loeve transform (KLT) of image coding, and the principal components analysis (eigenvectors) of pattern recognition and computer vision. Although projecting the signal onto the eigenvectors corresponding to the largest eigenvalues is optimal for efficient representation, it is not necessarily optimal for discrimination; it can actually be the worst thing to do.

The proposed new fourth criterion requires significant contributions from both communities to determine which bits are most important perceptually and semantically. Just like there are keywords and keyframes, we can expect to find (or have users identify) "key bits." The two communities can combine their expertise to determine how to encode the key bits most efficiently, while making their content easiest to access. Progress in identifying these bits should also yield more meaningful error criteria than the woeful MSE-based SNR.

Computer vision is becoming more "semantic," linking high-level contextual information with low-level bits. Progress is steady toward recognizing "what's in this picture?" Scene description is the ultimate "semantics-preserving" compression. Progress in the latter is currently boosted by the availability of databases of common theme; knowing the context is "soccer" allows one to use prior knowledge to both recognize and code texture, objects, and action more efficiently. Compression researchers know the importance of a good source model; computer vision researchers have not usually thought of motion or texture modeling as source modeling. The best source model as measured against the proposed four criteria will be the one which exploits context to give efficient compression and access to content.

Clearly, image compression needs computer vision; is the reverse also true? (Other than to save disk space.) The answer is "no" if one refers only to the "get an extra dB" standards-driven subset of the compression community. In contrast, if the focus in compression would shift to the fundamentals of what I'll call "semantic information theory," the goal Shannon missed of putting "meaning" into "information," then the two communities would find an exciting new common ground. Both know that the smallest picture description requires both good content understanding and good modeling. What kind of picture description will be optimal with respect to the four criteria above? Is it possible for "content coding" to be lossless? Here lie issues to challenge both communities!