What's Wrong with Today's Video Coding?

V. Michael Bove, Jr.
MIT Media Laboratory

From *TV Technology,* February 1995, reprinted with permission

As more video professionals work with JPEG and MPEG, more will experience a phenomenon that's long been familiar to those of us on the research and development end of digital video: In your post-production system, you take a video sequence that's flawless compressed to a particular bit rate, and overlay some text or graphics on it. After recompression, the video looks appallingly bad unless the bit rate is raised considerably.

What's happening? The answer, simply put, is that transform-based compression algorithms like JPEG and MPEG (and wavelet-based ones as well) are optimized for camera imagery at the expense of graphical patterns. In the case of motion-compensated coders and scrolling titles (or stationary text over a camera pan), the inefficiency is worsened by the lack of any notion of background and foreground: the encoder is spending its bit budget constantly retransmitting information about occluded and revealed regions of the image.

Of course, this isn't the only thing wrong with the current generation of video coding methods. It's not even the worst thing. But it is an illustration of the costs of a mismatch between a coding algorithm's model of the world and the information it's trying to compress. What's less well known is that the model-mismatch problem among current algorithms isn't limited to sharp-edged graphics; while the video production and distribution community is cautiously embracing transform-based coders, the research community has begun to look toward video representations that are more physically and semantically related to scene structure. Instead of breaking images into square tiles that have nothing to do with the imagery, methods under investigation use motion and other cues to identify coherent regions. Instead of image-plane motion vectors of the three-pixels-down-and-two-to-the-left variety, improved algorithms model object motions in three-space. The fundamental change is that of recasting the design goal: rather than trying to encode arrays of pixels with some optimal tradeoff between efficiency and fidelity, researchers are now looking to uncover information about scene structure such that a receiver can synthesize images from the scene description. In many cases, the outcome of the encoding looks suspiciously like a computer graphics database and script. Some of these concepts will probably see their first widespread application in MPEG-4, which -- though it may concentrate on low-bit-rate, low-fidelity applications that aren't what the readers of this magazine would call ``television'' -- will be of interest as the first attempt to standardize a higher-level representation for video.

So that I'm not accused of merely wanting to rouse more discontent, I'd like to note what's good about transform coding. It is (or soon will be) ubiquitous and inexpensive. It involves a short and deterministic encoding and decoding delay. It's a relatively low-risk method, as it never falls totally to pieces given a reasonable bit rate (this can't be said for some of the more sophisticated model-based coders, which generally give a lot more compression but occasionally fail to interpret regions or motions correctly). Thus it's likely that for many applications transform coders will be with us for a long time.

The fact is, there probably doesn't exist a single digital video representation that's always best. Rather than looking for one, we'd be better off with a sufficiently powerful and flexible decoder such that a digital video sequence could be represented in what the originator (or even a ``smart'' encoder) deemed the best form for

that application. Parts of the same scene might be represented differently, as in our text-overlay example. That this is a reasonable scenario to expect is supported by the appearance of flexible video processing chips like Texas Instruments' MVP, and by some even more powerful hardware still in design stages in various laboratories.

If you think this is a lot of trouble just to get a little more compression, I agree with you. Indeed, where I work we regard compression as secondary to enhancing the degrees of freedom. A model-based representation can give us a bitstream or database that's much more searchable and manipulable than an MPEG bitstream, whether for easing post-production or viewer browsing. Another reason has to do with degrees of freedom in display. One of the tenets of digital video is that it will end up being displayed in all sorts of ways -- from a personal digital communicator to a postage-stamp-sized window on a PC screen to a ``normal'' TV to a wall-sized projection. Composing and editing video so that it is attractive and understandable under all those circumstances is nearly impossible. But suppose the originator could specify that on a small screen the titles are set in a larger, more legible font? Or that viewing on a large, wide screen entails fewer pans and cuts than on a 4:3 screen? What's been called ``directed pan-and-scan'' is only the beginning of the possibilities. Imagine that the action could take place on a 16:9 stage set instead of a 4:3 set depending on the display, or that extraneous details and clutter could be minimized for viewing at postage-stamp size. A simulated lens focal length might change from wide-angle to telephoto. From a technological vantage point it's easy to forget, but television ultimately shouldn't be about transmitting pixels. It ought to be about conveying information through visual means, and conveying it most effectively may not involve showing everyone precisely the same picture irrespective of the viewing situation.

While the scene-analysis methods needed for more intelligent representation of digital video are still active areas of research that may not escape from the lab for a few years, the advantages accompanying such a representation suggest that designers of post-production tools, video decoding hardware, and bitstream headers and descriptors ought to look beyond the capabilities and requirements of the current round of technology. The creative community also should consider this a fair advance warning.

V. Michael Bove, Jr. is an Associate Professor of Media Technology at the Massachusetts Institute of Technology, working in the Media Laboratory. He is current holder of the Sony Career Development Professorship.