M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 339
Also appearing: Int. Conf. on Image Proc., Washington DC, Oct. 1995.

Light-years from Lena: Video and Image Libraries of the Future

Rosalind W. Picard
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
picard@media.mit.edu, http://www.media.mit.edu/~picard/

Abstract

The average consumer with a personal computer will soon have access to the world's collections of digital video and images. However, the theory and tools that facilitate browsing, querying, retrieval, and manipulation of imagery are still in their infancy. For example, people would like to access content in movies, e.g. "fast forward to where they bicycle through the sky." This new application area reveals an abundance of unsolved scientific problems for image processing. In this paper I overview key technical challenges that the image processing community should embrace.

1 Introduction

For decades, image processing research has focused on problems arising with medical and scientific data, and with compression of "general" imagery such as the ubiquitous Lena images. There has been little change in the emphasis of image processing research. Although tremendous progress has been made, the focus has remained on pixel-level processing, with the goal to acquire, enhance, restore, or compress images better.

However, a great change is happening now. Whole libraries of film, video, photographs, paintings, drawings, and clip art, in addition to medical and scientific imagery, are becoming digital. Digital imagery is now in the hands of anyone with a personal computer. Multidimensional filtering is no longer the domain of researchers, but can be understood by children playing with over-the-counter software. Moreover, these "common customers" now pose some of the most challenging questions for image processing research. In particular, they want to access images by content, e.g. "Computer, find the scene where Fred Astaire dances up the walls." The new problem statements are simple enough for all people to understand; indeed, to expect image processing researchers to solve. But, like Fermat's last theorem, a simple problem statement does not imply an easy solution.

What research problems need to be solved to facilitate content-based access to digital libraries? In this paper I outline the key research problems, and review some of the related work beginning in this new area.

1.1 No image is an island

Facilitating content access is not just a problem of image processing. For example, suppose a user requests scenes of Jay Leno making funny jokes about Rush Limbaugh. This request not only involves processing the image frames and motion, e.g. scene detection, face and gesture recognition, but also audio processing (keyword spotting, laughter detection) and dozens of semantic and artificial intelligence issues (script and story knowledge, natural language, common sense understanding), interface issues, (query input and presentation of results) as well as networking and service issues (delivery, copyrights, billing.)

Although these domains can be treated separately, some should not be separated. For example, visual gestures often co-occur with Leno's jokes, and are repeated until the audience laughs. Combining visual and audio information will often make detection and identification more robust.

Despite the need for combining acoustic and visual signals, I will restrict the scope of this paper to image processing. In particular, I will focus on problems of representation and analysis of image content, and frame-to-frame motion.

2 Next generation problems

2.1 Different false alarm criteria

Image content recognition has been pursued by the military for years. However, the success criteria have changed with the new consumer domain, and hence the optimal solutions have yet to be determined. Detection algorithms for both military targets and consumer content generally attempt to maximize the probability of detection for a given false alarm rate. In consumer applications such as browsing an image database, false alarms might correspond to pictures the user didn't request, but which they might still enjoy seeing. In contrast, military false alarms can lead to shooting the wrong target. The natures of the signals and noise also differ greatly in these two domains, e.g., if a user wants scenes with one particular actor, then all other actors might be considered noise. Although problems in these two domains might be posed with the same Bayesian theory, the different cost criteria warrant a fresh investigation for the new consumer applications.

2.2 Find interest regions, not one segmentation

Segmentation has been a topic in computer vision for decades [1], and has recently become an area of image processing research. Traditionally, an image segmentation is defined as a *partition* of the image into homogeneous regions, where "region" depends on a specified definition of homogeneity. Frequently, edge detection (high-pass filtering followed by a nonlinearity) is posed as the dual problem of segmentation. It is easy to come up with new methods for edge detection and segmentation; consequently, there are more papers on these topics than pixels in the Lena image.

The main problem with traditional segmentation is that it assumes one optimal partition exists; this is rarely the case. Furthermore, obtaining pixel-precise edges and segments is unnecessary for most content-based retrieval. The emphasis on refining edge detection and traditional segmentation should be shifted to an important variation of the segmentation problem – finding regions of interest to a user.

Consider a picture of a crowd of people. One user may be interested in identifying individual faces; another might be interested in the crowd as a whole. Since regions of interest vary with an individual's goals, a hierarchy containing multiple segmentations is desirable. Ideally, the hierarchy is generated by multiple models which specialize at different goals (color, texture, people, etc.). However, the traditional approach to segmentation is very computational and is aggravated by the addition of multiple models.

Presently, at least two solutions exist, and these are due to the advantage of having a user interacting with the system. Users can give example regions of interest, and give the system (or agent) feedback as it selects other regions it thinks are similar. As the user interacts with the system, the system dynamically determines which model or combination of models best describes the regions of interest. This is the "society of models" approach of Picard and Minka [2]. This approach is similar in spirit to the learning algorithm of Delanoy and Sasiela [3]. In both cases, the system accepts positive and negative examples of regions from a user, then determines what is the best model or set of features for subsequent classification and identification. In the society of models, the system can also combine models dynamically, if no one model meets the user's criteria.

2.3 Edge detection in time

Instead of dealing with all frames of a video at once, it is often useful to separate them into shots, scenes, and segments. A "shot" is an unbroken sequence of frames from one camera, e.g., a zoom of a person talking. Before video footage is edited, two shots are always separated by a cut, which can be thought of as an "edge" in time. After editing, shots may be separated by cuts or by edited transitions such as a fade or wipe, which are usually functions of two or more frames (filtering in time). A "scene" is a sequence of shots that focus on the same point of interest, e.g., a person browsing through a store. A "segment" is a sequence of scenes that forms a story unit, e.g., a flashback.

Video parsing research has been directed so far toward the problem of detecting shots, e.g. Araman *et al.* [4], Tonomura *et al.* [5], and Zhang *et al.* [6]. Most of the methods have relied either on differencing (high-pass filtering) all the pixels or a subset of them in two frames, or on differences of gray-level or color statistics. In both cases a close analogy exists to early work in spatial edge detection. Matched-filter methods are also useful, and have the ability to both detect and identify edited transitions [7].

Parsing video into scenes and segments is a harder process that has not been carefully examined. The transition from shots to segments parallels the transition between low-level and high-level processing in computer vision. The former can be accomplished at the pixel level, while the latter ultimately requires high-level understanding in tandem with low-level features.

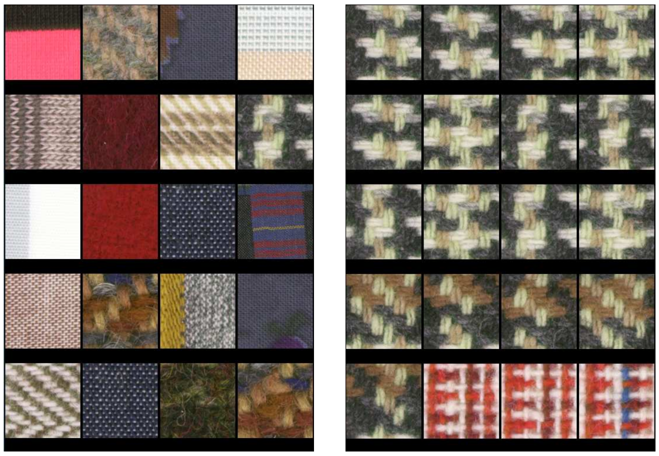

Figure 1: Browsing fabrics for similar texture patterns.

2.4 Spatio-temporal segmentation

Spatio-temporal segmentation is useful for tasks such as "extract the person walking" or "find shots with crashing ocean waves." Here, a combination of spatial and motion features is necessary to identify the regions of interest. The resulting regions are volumes in "xyt" – where x and y are the image axes, and t is the time axis. Very little work has been done in this area, although it appears to be an important step toward characterizing video content.

2.5 Content-based retrieval and organization

Research in image retrieval is relatively new, with the initial largest efforts by IBM [8], ISS [9], and MIT [10]. These systems have emphasized representations of shape, color, and texture. For example, in Fig. 1, a screenful of fabric samples is shown, part of a database used in the apparel industry. Each sample is represented by a set of precomputed color and texture features. A user can select features which are important, and then click on an image to retrieve more images that "look like" that image, perhaps subject to some desired constraints such as cost, availability, etc. An example of retrieval is shown at the right in Fig. 1, where the system has found images similar to the one at the upper left and displayed them in raster-scan order according to their texture similarity.

The features needed for retrieval vary with the domain. Rowe *et al.* [11] identified three types of indices for queries in video on demand systems: bibliographic (title, genre, etc.), structural (shot, scene, segment, etc.), and content (actors and objects in scenes, etc.) Romer [12] identified three types of requests in stock photo searches: subject and action content (three stooges, throwing a pie), picture syntax (horizontal structures, smooth open spaces), and subjective components (mood – frolicsome, powerful). In commercial photography for advertising the subjective features were most important. For editorial purposes, subject and action were most important. Identification of subject and action is a fertile field for image processing research.

Color, texture, and shape are useful when searching for an image "similar to" one you have, or when specifying an image you've seen before; more work needs to be done to make these features more robust to variations in natural scenes. These features also help identify smooth open areas for images where text will be overlayed. Much more work also needs to be done on detecting people in imagery, and identifying the category of a shot when such categories exist. For example, useful categories exist for photographs with people: face, face & partial body, face & full body, two people, picture within a picture, group, crowd, one in front of many, etc. [12].

In some cases the person does not know what they want to retrieve; they say, "I'll know it when I see it." However, they do not have time to look at even a thumbnail version of the millions of images available. To assist this search, how does one perform "image organization?" Alphanumeric data can be meaningfully ordered, but images cannot. The problem is not unlike that of building a universal codebook, where similar data must be clustered together. Hierarchical clustering methods from pattern recognition and vector quantization can be applied to images and their subregions to help find examples which are representative of the whole collection. These examples can be presented first to the user, reducing the time to browse. As the user provides feedback, the system can re-organize, presenting new representative images.

2.6 Wanted: good similarity criteria

Both cases above, where one is looking for an image "similar" to another, and where one doesn't know what they are looking for, are plagued by the same problem: a good criterion for similarity.

The problem is not new to image processing – in image coding, researchers have long looked for a measure of "perceptual similarity," recognizing that mean-squared error is not the criterion used by people to judge image similarity. In contrast, distance measures applied to suitable color and texture features can perform surprisingly well for image retrieval [13], [14]. Nonetheless similarity is complicated by many factors, and no one criterion of similarity, e.g. perceptual or semantic, will be appropriate for all applications [2]. More research is needed to determine suitable features and similarity measures for comparing visual information.

2.7 Motion analysis: camera, action

Motion analysis can be divided into scene motion (actions) and camera motion. When the scene is static and the camera motion does not induce much parallax, then the camera parameters can be solved and used to transform the coordinates of one frame so that it aligns with another; this problem has recently been solved using the exact projective coordinate transform [15]. When there are multiple motions in the scene, these can be approximately found using affine methods [16] or methods with also account for parallax [17].

Recognition of the extracted motion is an important new area of research which merits increased effort. Sometimes low-level features can be very successful for recognition of motion textures (water, leaves, etc.) and even human activities (see Polana and Nelson [18]).

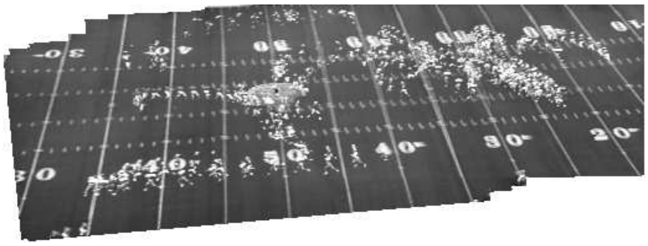

Figure 2: "Stroboscopic" image made from several frames of a noisy football video. See http://www-white.media.mit.edu/~steve/orbits/orbits.html for a closer look.

2.8 Browsing video

It is unclear at this stage what will be the best way to browse video. Tonomura *et al.* have outlined several possible solutions [5]. One of these options, the stroboscopic browser, attempts to make a still image out of a shot or scene. These "dynamic mosaics" [17] or "summary frames" [15] combine lots of frames seamlessly into a new image using either an affine model with parallax correction [17] or a direct projective method [15]. An example of the latter is shown in Fig. 2. Here, frames of video were automatically coordinate-transformed to align them on top of each other. If a football player moved between frames, he shows up in a "stroboscopic" trajectory according to the path he moved. This research facilitates subsequent analysis of what play was run, and helps in future automatic analysis and retrieval of successful plays.

2.9 Compression with content access

A typical movie can be compressed down to 2-3 gigabytes, but may occupy over 200 gigabytes when decompressed. If you are searching for a particular scene, then clearly it will save time and computation if the search can be done on the compressed data. In fact, most large searches will be more efficient if done on the encoded data. The problem is that coding has focused for years on maximizing rate while minimizing distortion and cost. Often this results in an encoded form which prohibits accessing the content. If the encoded form is also to be "searchable" then that adds a fourth criterion to the optimization [19].

However, jointly optimizing with respect to these four criteria can lead to much better systems. For example, designing a system that simultaneously compresses and classifies was found by Oehler and Gray [20] to give excellent results.

The more one knows about the content of an image, the better one should be able to compress it. Homogeneous regions, when they exist in images, can be compressed more efficiently [21]; this approach forms the basis of second-generation and model-based coding efforts. But rarely do large homogeneous regions exist to justify the extra gain in complexity for the goal of compression alone. However, if the goal is both compression and content-access, then the extra complexity is justified, for it saves enormous work during the searching stage.

Shot change detection has already been run successfully on compressed MPEG and JPEG data [4] [6]. Smith and

Chang have also run retrieval algorithms directly on compressed data [22].

3 Summary: Hard questions

In this short paper I have tried to overview key image processing research problems which must be solved to give people access to the content of digital video and image libraries. Perhaps the two most challenging questions are: "What is a good set of models to represent images, to facilitate both compression and content-access?" and "What is a good measure of visual similarity?" These, together with the suggested directions in this paper will hopefully assist the development of future research in this new and exciting area.

4 Acknowledgements

Thanks to S. Mann for making Fig. 2, to T. P. Minka for developing the system shown in Fig. 1, and to S. Mann and F. Liu for their comments on a draft of this paper. This research was sponsored in part by BT, PLC and NEC.

References

[1] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Comp. Vis., Graph., and Img. Proc.*, vol. 29, pp. 100–132, 1985.

[2] R. W. Picard and T. P. Minka, "Vision texture for annotation," *Journal of Multimedia Systems*, vol. 3, pp. 3–14, 1995.

[3] R. L. Delanoy and R. J. Sasiela, "Machine learning for a toolkit for image mining," Lincoln Laboratory 1017, MIT, Lexington, MA, March 1995.

[4] F. Arman, A. Hsu, and M.-Y. Chiu, "Feature management for large video databases," in *Storage and Retrieval for Image and Video Databases* (W. Niblack, ed.), (San Jose, CA), pp. 2–12, SPIE, Feb. 1993.

[5] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, "Structured video computing," *IEEE Multimedia*, vol. 1, pp. 34–43, Fall 1994.

[6] H.-J. Zhang, C. Y. Low, and S. W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tools and Applications*, vol. 1, pp. 80–111, March 1995.

[7] A. Hampapur, R. Jain, and T. T. Weymouth, "Production model based digital video segmentation," *Multimedia Tools and Applications*, vol. 1, pp. 9–46, March 1995.

[8] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: Querying images by content using color, texture, and shape," in *Storage and Retrieval for Image and Video Databases* (W. Niblack, ed.), (San Jose, CA), pp. 173–181, SPIE, Feb. 1993.

[9] H.-J. Zhang, S. W. Smoliar, J. H. Wu, C. Y. Low, and A. Kankanhalli, "A video database system for digital libraries." ISS, Nat. Univ. Singapore, June 1994.

[10] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases," *Int'l Journal of Computer Vision*, 1995. in press.

[11] L. A. Rowe, J. S. Boreczky, and C. A. Eads, "Indexes for user access to large video databases," in *Storage and Retrieval for Image and Video Databases II* (W. Niblack and R. C. Jain, eds.), (San Jose, CA), pp. 150–161, SPIE, Feb. 1994. Vol. 2185.

[12] D. Romer, "The Kodak picture exchange," April 1995. seminar at MIT.

[13] M. J. Swain and D. H. Ballard, "Indexing via color histograms," in *Image Understanding Workshop*, (Pittsburgh, PA), pp. 623–630, Sept. 1990.

[14] F. Liu and R. W. Picard, "Periodicity, directionality, and randomness: Wold features for perceptual pattern recognition," in *Proc. Int. Conf. Pat. Rec.*, vol. II, (Jerusalem, Israel), pp. 184–185, Oct. 1994.

[15] S. Mann and R. W. Picard, "Video orbits: characterizing the coordinate transformation between two images using the projective group," Tech. Rep. 278, MIT Media Lab, Perceptual Computing, Cambridge, MA, 1995.

[16] M. Irani, B. Rousso, and S. Peleg, "Detecting and tracking multiple moving objects using temporal integration," in *Computer Vision - ECCV*, (Santa Margherita Ligure, Italy), pp. 282–287, Springer-Verlag, May 1992.

[17] H. S. Sawhney, S. Ayer, and M. Gorkani, "Model-based 2d & 3d dominant motion estimation for mosaicing and video representation," in *ICCV*, (Cambridge, MA), June 1995.

[18] R. Polana and R. Nelson, "Low level recognition of human motion," in *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, (Austin, TX), 1994.

[19] R. W. Picard, "Content access for image/video coding: 'The Fourth Criterion'," Tech. Rep. 295, MIT Media Lab, Perceptual Computing, Cambridge, MA, 1994. MPEG Doc. 127, Lausanne, 1995.

[20] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE T. Patt. Analy. and Mach. Intell.*, vol. 17, pp. 461–473, May 1995.

[21] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second-generation image-coding techniques," *Proc. IEEE*, vol. 73, no. 4, pp. 549–574, 1985.

[22] J. R. Smith and S.-F. Chang, "Transform features for texture classification and discrimination in large image databases," in *Proceedings ICIP*, (Austin, TX), pp. 407–411, Nov. 1994. Vol. III.