Challenges and Opportunities for PR&CV Research in Year 2000 and Beyond

Dr. Dragutin Petkovic

IBM Almaden Research Center
San Jose CA USA

1 Introduction

In this short position paper we attempt to point out to Computer Vision, Pattern Recognition and Artificial Intelligence (AI) communities some important and challenging opportunities for future research, development, and applications related to visual information. The emphasis is on both "challenging and important", since we believe both need to be addressed in any successful R&D effort. This paper is intended for researchers, directors of R&D programs and young researchers planning their research programs. Paper is based on authors experience ranging from biomedical to industrial computer vision, multimedia information retrieval and user interfaces, both as a researcher and as a product manager. In addition, this paper is based on authors extensive communication with IBM colleagues, with academic colleagues, especially at MIT Media Lab (Prof. Roz Picard and Sandy Pentland), and with "customers" i.e. users of such technologies in manufacturing and media industry. At the end, we suggest some readings and Web sites simply as examples of interesting work known to the author, and certainly not as an exclusive list.

We all know that majority of human communication is based on visual information. However, there were two basic impediments so far to make machines an effective partner in this information processing: a) our basic knowledge how this process is done; and b) lack of adequate technology for processing, storage, capture and transmission. Last few years the technology situation improved rapidly. We have increasingly powerful processors, recently getting specialized instruction sets for signal/image processing (i.e. Intel Pentium with MMX). Such processors progressed so rapidly that making special purpose and expensive vision chips is not cost effective any more. On the storage side, we envision laptops having 10 GBytes disk space soon, with 128 Mbyte RAM memory. Tape/optical/disk libraries can manage Terabytes of data economically. Recent DVD standard for video CDRoms promises full featured movies (2 hours of digital video) on one CDRom. Networks are also improving dramatically. Another big change on the hardware and technology front was of course Internet. It finally made it clear to all commercial companies that they can reach all sorts of users, thus significantly speeding up the conversion of their main business data (including images and video) into digital form. Promise of Internet is one of the biggest changing factors where our community will have to play - it will drive both software, hardware and algorithms and the way computers are used. In addition, desktop machines increasingly are becoming equipped with

sensors like cameras, scanners and video cameras. Digital cameras are beginning to proliferate making it feasible to create a lot if images and video in digital (even MPEG compressed) form.

The above changes will enable radically new environment where visual data will be easily and economically available in all areas of human endeavor, and therefore will also be sought after and used by wide variety of the users. This is the key driving factor which our community will benefit from - without it our work would be largely relegated to narrow areas. Unless political situation in the word changes, the key driving factor for R&D will be commercial, not government funding. Therefore, our community has to look much more closely into these commercial applications in the future.

While the progress on technology was immense, our progress in basic understanding of image analysis and intelligence is still very moderate. While we have to continue to push for basic understanding of these processes, we also have to seize the opportunity new technology offers and look into what our community can contribute both in the near future (2-3 years) and long term (3-10 years). After all, Web years are equivalent to 3 months nowadays!

So far, majority of our focus was the "ultimate goal": full analysis and recognition from images, nothing less than completely replacing humans. This "noble" goal remains elusive and actually might be ill posed, maybe unsolvable in our times, but also not always necessary for quite effective progress in real applications. Discovering how to do full semantic analysis of images and video may take years, but why not try to go step by step? Instead of trying to recognize people in video, can we at least reliably find out are there people in the video and what basic actions they are doing? Instead of describing set of objects in the images automatically, can we combine some simpler pattern recognition with any other available information with the images like URL and text data pointing to it? Instead of trying to solve 100% or AI problem, why not let machines do what they are best (count, measure, analyze some narrow domain) and combine them with what people do best (meta knowledge, top control)? We are still far from famous Hal computer from "2001 - Space Odyssey "(4).

Here are some significant opportunities that should be attacked by our community. They are both challenging (require significant scientific and technological advances) and important (their solutions would bring real economic benefits). This list is of course not inclusive, but in authors opinion offers a good starting point. It includes content based retrieval; future video compression standards; and vision enhanced user interfaces.

2 Content Based Retrieval

This area already started and is attracting a number of researchers, conferences and companies (1-3). The idea is to automatically index images, video and audio using computer vision, pattern recognition and artificial intelligence, and thus augment (in some cases replace) tedious, expensive and inconsistent human indexers. In many cases, it is the only way to index large number of data elements (i.e. Web crawlers). While progress has been made, much more needs to be done. Here are some ideas:

- Automatically extract some simple semantics such as: is this black&white or color image, are there people or faces in the image, how many, indoor/outdoor, buildings or horizons etc.

- Segment videos into background and objects, compute some basic patterns (color, motion, texture, shape) from these objects, also compute some basic semantic attributes
(people vs. buildings etc.).

- Use more of machine learning to help index the images - let user annotate a number of them,
then let the system finish the job by for example trying several possible models.

- Use audio information for basic video segmentation (noise, pause, explosions, talking), spot some apriory defined words and index them, enable query by audio content ("get me tunes like this...").

- Make intelligent Web crawlers that can combine clues from ALL available sources (URLs, accompanied text, image/video/audio content baser retrieval) for automated indexing

- Enable very fast indexing (e.g. K-NN in very high multidimensional feature space) and integrate it with traditional data management models like relational databases and Web crawlers.

3 Video Compression

Currently two new MPEG standards are in the works: MPEG4 (5) and MPEG7. MPEG4 refers to very low bit rate compression involving both real and synthetic images/video and will require basic video segmentation - the idea is to segment objects from background and encode them separately. Emerging MPEG7 attempts to standardize content based retrieval of video. Both of these areas require significant image and video analysis of the type our community is doing. Challenge will be to develop a set of basic algorithms that are reliable and robust, and can also be

computed economically and work on all sorts of data. Similarly, for very low bit rate video conferencing extraction of participants' parameters like facial expressions can significantly increase compression ratio. Techniques for computer vision are in heart of such techniques.

4 Vision Enhanced User Interfaces

So far, user interfaces basically did not "close the loop" between the machine and the user. Computer has no knowledge of the state of the user (his/her gaze, facial position and expression etc.). The basic user input is still the keyboard and mouse. Why not gaze, gesture, body language - this is after all how we naturally communicate. Being able to understand this and add one more channel to user input has a potential to significantly improve the whole human - computer interaction, maybe first in some niche areas (disabled users, medical), but then many others. Current desktops are or can easily be equipped with video cameras and microphones and speech recognition is becoming a reality. Why not combine ALL of these inputs into one user interface system controlled by some intelligent central process? Success in this area will require computer vision and AI technology, but also significant human - computer interaction expertise, something our community has to incorporate in its research methodology. Doing gesture analysis for its own sake, without understanding how people could use it, will not be enough.

Once we can "sense" more information about the user, we can go even further, not only improve the basic control of computer input. "Affective computing" is another concept being discussed recently and is very promising (7). The basic idea is to enable computers to sense and use users' emotions (anger, joy, frustration etc.). Why would one want to do this? In many applications we want computers to be impartial and exact (like controlling a power plant). But in many applications like learning, we would like to make computers "softer" and more "human-like", allowing them the adjust their operation based among other things, on user emotion. This is after all characteristic of all good teachers, performers, speakers and coaches. Sensing user's frustration computer could change the training pace, maybe play some soothing music? It has been demonstrated that some basic emotions could be derived from videos of user's face (6). Once we perfect this process we can attempt to make much more human-like computer systems that could significantly improve training, learning, information manipulation/creating etc.

In summary, our community has important and challenging opportunities in the future, but we have to seize them and work effectively to solve them. technology, especially Internet in all its forms, works strongly in our favor. In order to make progress, we have to apply more strict research methodologies. For example, there is no excuse any more not to test our algorithms on very large data sets, they are available and there is compute power to do these tests. Another change for our community comes from the need to address a number of commercial applications, rather than government ones. For anybody who thinks they are less challenging, we

encourage to try them and to make an average person be able to use such systems - the challenges are many. Finally, user perspective has to be given much more weigh, since majority of systems will be used by people who are often non-computer experts.

Given all this, we have every reason to be optimistic and look forward to exciting, productive and fun R&D careers!

5 References, Suggested Readings and Web Sites

1. H. Zhang, P. Agrain, D. Petkovic: "Introduction to Special Issue on Representation and Retrieval of Visual Media in Multimedia Systems", Kluwer Academic Publishers, Vol. 4, No.1, January 1997

2. Interesting Web sites related to content based retrieval: webseer.cs.uchicago.edu; www.media.mit.edu; www.ctr.columbia.edu/webseek/; www.informedia.cs.cmu.edu; wwwqbic.almaden.ibm.com, www.virage.com

3. T. P. Minka, R. W. Picard: "Interactive Learning Using a Society of Models", Proceedings CVPR, San Francisco, 1996, pp 447-452.

4. D. Stork ed.: "Hal's Legacy - - 2001's Computer as Dream and Reality", MIT Press, 1997

5. Web site related to MPEG4: www.cselt.stat.it/ufv/leonardo/mpeg/cfp/snhccfp.htm

6. I. Essa, S. Pentland: " A Vision System for Observing and Extracting Facial Action Parameters", Proceedings CVPR, Seattle, WA, 1994, pp. 76-83

7. R. W. Picard: "Affective Computing", MIT Media Laboratory, Perceptual Computing TR, N. 321, MIT, 1995