IMAGE AND VIDEO ANALYSIS: TRENDS AND CHALLENGES POSITION STATEMENT

Touradj Ebrahimi

Signal Processing Institute
Swiss Federal Institute of Technology – EPFL
CH-1015 Lausanne, Switzerland
Touradj.Ebrahimi@epfl.ch

ABSTRACT

This text provides my views on trends and challenges in image and video analysis. I end the position statement by providing examples of applications where I believe in short and medium terms such tools will be used.

1. INTRODUCTION

Analysis of visual signals has been among pre-occupations and major challenges in signal processing since the early days of this discipline. We, as humans, but also animals and insects, have largely been relying on our visual sensory organs to interact with our environment, because visual information is one of the richest sources of information in our natural surroundings. Understanding the visual information processing system in humans and other animals by itself would be an interesting objective as it contributes to a better understanding of nature. But, making machines which can see and can understand their surroundings as other living entities do, has its own merits, and would open the door to an unlimited number of applications.

Despite an important effort in design and implementation of visual information processing algorithms during at least four decades, we have not been able to approach performance of even simple insects. Why? One could justify this lack of performance in machine vision when compared to living entities, on the required complexity in today's visual processors. It is indeed a fact that nature has concluded that a large portion of human brain, as those of other animals, should devote to visual information processing tasks. But, is the lack of complex enough processors the only reason behind this misachievement? After all, human visual system is a complex system also because it copes with all sorts of complex tasks in complex environments. In addition, today's computer processors approach the same computing power as in an insect's brain, and still, machine vision algorithms remain behind even simple tasks when compared to a vulgar fly.

Scientists active in fields such as artificial intelligence, neuroscience and machine learning believe that beside the required additional complexity, the actual architecture for vision tasks in living animals is largely different from those in machines. They say that a combination of complex enough artificial neural networks similar to insects, animals and humans brains, and a good training (learning) protocol is the key to the success of future visual processors. This is an attractive perspective and the author believes that such an approach bears a definite potential. Copying the nature where it has succeeded is an alternative not to ignore. There is a need for further effort and more attention from the machine vision and signal processing community on bio-inspired signal processing. But, at the same time, the author believes that many achievements, and work in progress within the current more conventional contexts need to be further investigated, and it is likely that a combination of what nature has concluded and what engineering science proposes could provide a better answer. After all, airplanes do not fly like birds in all aspects and mimicking nature has its limits as well.

2. APPLICATIONS

There are many applications where image and video analysis would bring an important impact to their performance. Some of the earliest goals such as image segmentation for medical imaging, pattern recognition (face recognition, gesture recognition …) and robot vision are as valid as ever. More recently, and thanks to digital revolution and progress in networking and communications, other already identified applications have increased in importance. Examples are search and retrieval of image and video information in distributed

environments, visual surveillance, and visual biometrics. The idea here is that machines that understand content can help humans in multiple ways, and provide new content based representation techniques facilitating many tasks for which you need a human operator at the end of the chain. Standards such as MPEG-4 and MPEG-7 provide syntax and tools which not only provide efficient content based representation methods, but also open the door to a seamless exchange of content and information in addition to interoperability. Such standards deliberately do not tell you how to extract content, but only tell you how to represent any content component once extracted. In many situations (in particular in case of MPEG-7), their success will largely depend on the existence of efficient and practical image and video analysis tools required in systems making use of them.

3. CHALLENGES

I have already mentioned some of the challenges facing visual information processing. In this section I would like to mention some of them in a more structured way.

3.1. Visual information acquisition

What is visual information? A trivial answer to this question is that an image or a video, if digitized is represented by a number of frames per unit of time, with each frame in turn represented by a number of components (three colors or more), each again represented by a set of pixels at a given precision (say 8 bits, or more), scanning the frame component on a raster, line by line. This is often referred to as first general representation, and was introduced taking into account practical issues such as camera and scan technologies, as well as simplicity of their representation. First generation image and video can be represented as one or more matrices whose elements correspond to a frame's component pixel. When compared to the first, second generation representation approach represents image and video as set of what is called attributes. A largely popular second general representation is that of object-based representation where to each object has been assigned some color, texture or motion attributes. Some researchers such as the author of this paper further make a difference between region-based and object-based representations in which regions are set of pixels with some meaningful homogeneity criteria well understood mathematically, whereas objects bear some more semantic notion such as a person or a car. The majority of image and video segmentation techniques try to take a first general image or video as an input and provide as output a second generation representation of them. Other image and video analysis tools extract other (often incomplete sets as far as the representation is concerned) and provide what one generally calls a content-based representation in form of edges, feature points, and others.

The challenge in first generation representation has mainly concentrated in coming up with more efficient acquisition devices such as cameras with either better quality, more natural representation of colors, efficient automatic gain control, etc. Image and video analysis algorithms have played and continue to play an important role to overcome these challenges. As mentioned earlier, more efficient segmentation and tracking algorithms, as well as feature extraction and tracking contribute to better performance of second generation acquisition devices. To this, one has to add other image and analysis algorithms which would make use of extended optical and other information to come up with richer representations of visual information. As an example, one can mention 3D acquisition devices (such as laser scanners and range cameras) which would represent visual information in 3D or 4D, or those making use of multi-view cameras. In fact, nature teaches us that humans, animals and insects often do not limit their visual processing to one single acquisition source but make use of stereo, multi-view and arrays of visual sensors (in insects for example). Multi-view acquisition and processing, and analysis is a yet young and not enough explored domain.

3.2. Human and animal visual system

Despite some progress, our knowledge about mechanisms of human and animal visual systems is very scarce. As mentioned earlier in this paper, there is a general belief that the nature has chosen a very different path and approach to process visual information. A deeper understanding of this path, and design and implementation of machine vision algorithms based on such approaches are among the challenges facing us.

3.3. Image and video analysis quality metrics

It is interesting to note the little importance devoted to this fundamental issue. How can we improve on something if we cannot measure its performance? Some efforts with limited success have taken place in image and video quality assessment. Challenges in image and video analysis quality metrics are far more difficult to overcome. As a starter, although arguable in many cases, there is at least already some widely used methods available to assess quality of an image and a video. For lack of any better alternatives, one can at least measure the distance between an image whose quality is to be measured and a reference image which is meant to be the original. In most image and video analysis problems, such as segmentation,

there is no definite reference to use. So, even trivial and limited solutions providing partial and incomplete answers are not at hand. The situation is no better for subjective tests. Image and video subjective quality evaluation protocols have found their ways in some standards for applications such as broadcasting, despite their limitations. The situation is far worse in image and analysis subjective metrics where again no systematic and well organized efforts has been made, with the exception of some limited studies with no serious impact or even results whatsoever.

3.4. Illumination invariant/aware image and video analysis

One of the most important shortcomings of current image and video analysis techniques is the inability of such techniques to cope with changes in illumination. Humans have this ability to cope with illumination by either filtering or by taking advantage of its effects when analyzing a scene (e.g. shape from shading). Visual processing which takes into account effects of illumination can on the one hand eliminate noise due to it (shadow removal for instance), and extract a more accurate information about the content such as its constituting objects, or on the other hand get a better understanding of the scene structure (extraction of source of illumination) for a more efficient further processing and image understanding. In general, one can say that in most natural and uncontrolled environments, changes of illumination are a major source of failure for image and video analysis.

4. TRENDS

In this section, I would like to discuss some short and medium term trends in image and video analysis where I believe there are opportunities and real identified needs.

4.1. Image and video search and retrieval

The advent of the Internet has been a real boost to increased interest and activities in the general field of information search and retrieval, including various filtering of specific information. Most efforts have been naturally concentrating on text based content. The main reason is that until recently, the majority of content in web pages are under this form. The situation is no more the same. Increasingly, computers and databases including websites contain sensory information of which visual information in form of graphics, image and video do not cease to increase. At least in terms of volume such data will soon represent a significant portion of accessible information. Search and retrieval methods for images already exist in popular search engines but they rely on text search approaches, often looking for words in the file name of an image or in the text surrounding it. This by no means provides a reliable and efficient way to search for a visual content. In addition, the approach to search for visual information could be very different. Search by query (by giving an example, or a sketch when looking for an image) is often a more natural way to look for a visual information than a solely text based request. There is a real and growing problem for search and retrieval of visual information to which efficient and reliable solutions should be found. Image and video analysis tools are very likely to be one of the most important components of such systems.

4.2. Visual surveillance and monitoring

Insecurity in community, national and international levels has been one of the growing problems in society during the recent years. Visual surveillance and monitoring is felt to be and often mentioned as one among other solutions in a complex and sensitive problem. Surveillance does not come only with advantages. It can also intrude the privacy of individuals. Efficient image and video analysis techniques not only could contribute in reducing the cost of monitoring by providing help to human operators, they can also help in creating surveillance systems that protect the privacy of individuals. Such a protection of privacy can be achieved by either conceiving fully automatic systems where no humans will be present in the system, or can be used at pre-processing in surveillance cameras which will only send to operator information which will not compromise the privacy of the individuals under surveillance. Examples of applications which benefit from image and video analysis include traffic surveillance, intruder detection, unusual behavior detection, etc.

4.3. Visual identification and verification

Related to the previous trend but not only limited to it, visual identification and verification is one of the essential technologies of the future. Already today, some of the major airports in the world provide access to staff and passengers by visual person verification techniques based on iris or hand recognition techniques. Other trials are in progress to identify individuals from video surveillance cameras, with far more limited success. More efficient and reliable techniques based on more advanced image and video analysis algorithms are yet to be found.

4.4. Human machine interface

The law of Moore has played an important role in the last four decades and seems to do so for at least another three decades. Thanks to this prediction, computers process increasingly faster, or can process more information, they store increasing more data, and they communicate increasingly more information among each other. In the chain of communication, the bottleneck has become the interface between humans and machines. The speed of communication between human and machines does not follow the same pace as those mentioned above. We interact with computers and machines pretty much with the same devices and roughly with the same speed as has been the case in the last half century. One way to increase the rate of information exchange between humans and machines is to make use of other parallel channels. Image and video analysis here again can play an important role in providing additional channels. Computers that can see and understand gesture, intent, and emotions of their users and their environment can contribute to an increase in rate of communication between man and machine, but also to provide alternative communication channels and more natural means to achieve interaction.

4.5. Mixed reality

Mixed reality is referred to a continuum in the space of representations covering on one extreme the real world objects captured by sensors such as cameras, and on the other extreme virtual objects entirely generated by computers. A mixed reality scene is a mixture of real and synthetic objects coming from various real scenes or generated by computer. In mixed reality often the aims is to not to be able to make the difference between what is real and what is synthetic as all should look photo realistic. In order to achieve this goal, objects from different sources should behave in a physically coherent way, not only they should move coherently with respect to each other, but also their conditions of illumination must be coherent. Here again, image and video analysis tools are required to extract various objects and parameters wherever needed. Mixed reality systems are being investigated by major mobile communication companies as one possible future of personal communication. The author believes that an increasing number of mixed reality applications will be used in near future and that image and video analysis will play an essential role in them.

5. CONCLUDING REMARKS

In this position statement, I provided some of my views on challenges and trends in image and video analysis, and gave some examples of potential applications where in near and mid term, they will be used. It is important to mention that many of the issues discussed here are not limited to visual information, and that other sensory information such as audio face the same challenges. In addition, it should be noted that the solution to many of the challenges and trends won't be found within a mono-modal framework but rather through multi-modality in the sense that it is by processing multi-modal information, and by combining them that a good solution can be provided.

ACKNOWLEDGEMENT

This work has benefited in part from activities of the EC funded network of excellence VISNET and various grants from the Swiss Federal Office for Education and Science (OFES).