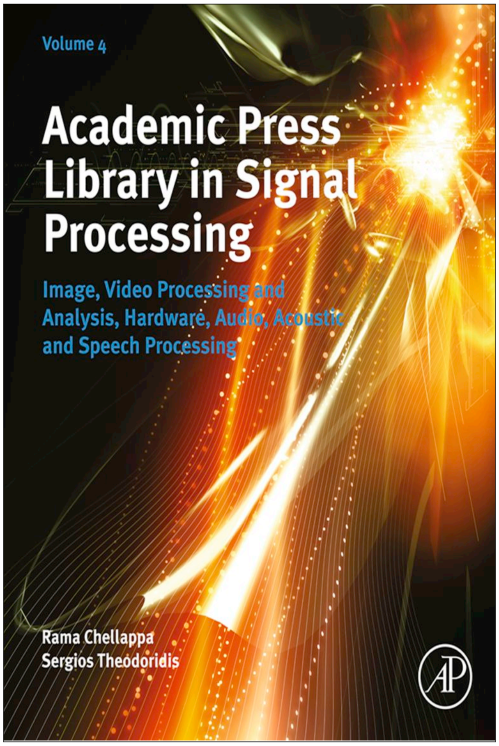

Volume 4
Academic Press Library in Signal Processing
Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing
Rama Chellappa
Sergios Theodoridis

From Amit K. Roy-Chowdhury, Video Processing—An Overview. In: Rama Chellappa and Sergios Theodoridis, editors, Academic Press Library in Signal Processing. Vol 4, Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing, Chennai: Academic Press, 2014, p. 343-347.
ISBN: 978-0-12-396501-1
© Copyright 2014 Elsevier Ltd
Academic Press.

CHAPTER

# Video Processing—An Overview 13

**Amit K. Roy-Chowdhury**

*Department of Electrical Engineering, Department of Computer Science (Cooperating Faculty), University of California, CA, USA*

Video processing can be defined as analysis of the content of the video to obtain an understanding of the scene that it describes. It is an essential component of a number of technologies, including video surveillance, robotics, and multimedia. From a basic science perspective, methods in video analysis are motivated by the need to develop machine algorithms that can mimic the capabilities of human (and other animal) visual systems. It is an area of research that has seen huge growth in the recent past. Researchers in video analysis have varied backgrounds, including signal/image processing, computer science, systems theory, statistics, and applied mathematics.

In this book, we will start by providing a broad overview of the tasks involved in the analysis of videos, including a summary of the research challenges in each, followed by a brief tour of the application areas and their significance, and finally an overview of the articles in this section. Note that we use the terms video processing and video analysis interchangeably.

## 4.13.1 Basic tasks in video analysis

The various tasks in video analysis can be categorized as low-level, mid-level, and high-level. Although there is some ambiguity on which tasks belong to each category, there are some broad trends in the literature. For example, some researchers consider edge detection and segmentation as low-level, 3D reconstruction as mid-level, while recognition is a high-level task. However, the categorization is probably less important than the actual tasks. Below, we provide a brief summary of the most important tasks. This section on Video Processing includes chapters that contain details of many of them.

At the level of a *single* image, two basic operations are fundamental to most video analysis applications. These include computing *gradients* which highlight the portions of the image where there is a substantial change in the image intensity, and analysis of the image *intensity* value, be it a gray-scale or color value. Computing gradients can lead to identifying the *edges* in the image, which can be combined together to identify structures like *lines* and *corners*. These basic structures can be the building blocks for identifying higher-level cues like shapes of objects.

Gradient estimation is also the foundation upon which different methods of image *segmentation* and object *detection* are built. Segmentation is the process by which coherent regions of the image are identified. These regions can be the basis of a higher-level analysis of the entire image. For example, an efficient segmentation algorithm of a natural scene may be able to divide the scene to consist of a

**Academic Press Library in Signal Processing.** http://dx.doi.org/10.1016/B978-0-12-396501-1.00013-3

green meadow, a blue sky, and a man-made structure like a house. The reason the algorithm is able to identify these regions separately is because each of them have characteristics that are common within that region and change significantly from one region to another. This involves analysis of the image intensity values to estimate the extent of similarities and differences.

*Detection* is also a basic low-level image analysis task as it can help identify the interesting objects in the scene, e.g., people, which can then lead to an understanding of the scene or an analysis of the actions of the objects. There are a number of detectors that have been built for different kinds of objects, the most common being person and vehicle detectors. Again, they combine analysis of the image intensities and their variations, often building statistical models that can serve as a signature for that object.

One of the factors that affects the performance of image analysis algorithms is the *quality* of the image. There could be a number of sources for this, including sensor noise, environmental conditions like lighting, and occluding objects that may temporarily mask the ones of interest. It is one of the most active areas of research and includes various trends—statistical modeling of image quality, machine learning based approaches to compensate for the variation in quality, and physics-based approaches that model the environmental factors to account for their effects.

Given a *sequence* of images (i.e., a video), which are usually highly correlated, an additional task is to compute the *motion* of the objects over the video. Again, a number of methods have been proposed for this purpose. *Optical flow* is a method for estimating the motion of each individual pixel. Combined with segmentation, it can provide a sense of how each part of the scene is changing over time. *Tracking* involves computing the location of each object over time, given the detections of the objects in each frame. Bayesian tracking approaches like the Kalman filter or the particle filter, combined with suitable data association strategies, have been adapted to the video analysis tasks. Other approaches have looked at how the distributions of certain object characteristics, like image intensity values, change over time to compute a track of these objects. Although motion analysis, including both flow computation and tracking, has been the mainstay of video understanding research for some time, robustness to environmental variations, as well as scene occlusions and clutter, remains dominant challenges for existing methods.

One of the special cases that needs particular attention is a multi-camera environment. In addition to the above, such an application domain introduces its own challenges. Images of the same object captured by different cameras need to be *registered* so that similar features are combined together for further analysis. Reidentification of a target over a network of non-overlapping cameras after it is not visible in any camera for a significant period of time is another important challenge. Moreover, multi-camera environments bring up interesting research problems in distributed image processing, whereby an understanding of the scene needs to be obtained by each camera acting as independent agents in coordination with other cameras in a local neighborhood, rather than sending all the data to a central processor.

Building upon the above-mentioned tasks, video analysis applications can obtain a higher-level understanding of the scene. Since an image is a 2D representation of the 3D world, a natural question to ask is whether it is possible to recreate the 3D scene given the images. This is an *inverse estimation problem* and is ill-posed unless proper constraints are imposed on the structure of the scene and the image acquisition process. The imaging process, which involves developing a mathematical model of the camera that maps the 3D world and 2D image, is a necessary first step in the process. Given the model and a collection of corresponding points in the image and the world, it is possible to then *calibrate* the camera.

Various cues can be used for 3D reconstruction of a scene from a set of images. One possible grouping is based on the use of multiple images. When there are two cameras used to image the same scene, *stereo reconstruction* approaches can be used. From a single camera, the motion between the frames in a sequence of images can provide an estimate of the 3D depth of the scene (*structure from motion*). Additionally, *shading* and *defocus* have also been used for estimating the depth of a scene. For a complex scene, like a city street, the basic depth estimation strategies can be combined together with natural constraints to provide a more robust estimate of the overall scene structure.

The overall goal of video analysis is to obtain an understanding of the scene. The tasks described earlier provide the building blocks for this. Scene understanding requires *recognition of objects* and *events*. Object recognition can be achieved at the level of a single image, while recognition of activities and events usually requires multiple images. In fact, various time scales can be used for recognizing activities over different temporal horizons. Moreover, the complexity of the scene can define the kinds of activities that need to be recognized, e.g., from a single-person activity to interactions between a group of people.

An active area of research in recognition is the use of *contextual information*. For example, the human eye recognizes a table not only from its looks, but also from the setting that it is in. An interesting question in video processing is how to model this surrounding information for more effective recognition of objects and activities. Machine learning based approaches that model the interrelationships between various detected objects and activities have become popular in this regard.

An overarching challenge in video analysis is to account for the errors in the lower-level modules, e.g., detection and segmentation, in higher-level modeling tasks. Integrated approaches that combine these various levels are increasingly becoming the trend in this regard. In these approaches, higher-level analysis, like recognition, can provide cues for better segmentation, detection, and 3D reconstruction. This requires designing suitable objective functions that can combine the results of the individual modules to satisfy the overall goal of the system. Alternatively, some researchers have taken the approach that precise segmentation, detection, and tracking which may not be necessary; rather, it may suffice to learn time-varying patterns of image features and use these learned patterns for recognition.

Over the past few decades, great strides have been made in video analysis leading to the development of many advanced technologies. Even as basic researchers work on some of the most challenging issues, application domains often provide enough constraints to lead to satisfactory solutions in that setting. Moreover, the growth in the Internet, social media, and video capture devices, coupled with the needs of security applications, has led to huge video repositories. Searching through them is opening up new research problems, which will probably call for novel solution strategies. Rather than fully autonomous applications, the convergence of man and machine may be the future trend in many practical video analysis solutions in the near future.

## 4.13.2 Applications in video analysis

Below is a list of some of the most common applications of video analysis. Many of them are covered in the chapters in this section.

*Video surveillance:* This is a traditional application domain that includes all the research tasks outlined above, with applications ranging from national security to environmental monitoring. The extraction of biometrics for person identification is a specialized sub-area of this application.

*Social media and the Internet:* The preponderance of videos, and the manner in which these are uploaded and used, on the Internet provide novel challenges to index and search such "big data" repositories.
*Mobile communications:* Advances in video analysis can lead to more efficient use of bandwidth and power in handheld devices, like smartphones, with users sharing videos among themselves and with cloud computing servers.
*Virtual reality:* Virtual environments can be made more realistic if information can be gleaned from videos of natural scenes and incorporated into the rendering process. This requires automated analysis of the content of the video.
*Vision-based robotics:* Robots equipped with cameras, working independently or alongside humans, can help in navigating through complex environments, like a disaster zone. Such applications require advances in various aspects of video processing—tracking, recognition, and distributed processing.
*Computational photography:* Recent advances in areas like compressed sensing have opened the possibility of designing efficient cameras that would reduce the amount of data that is collected, without affecting the fidelity of the analysis on the data.
*Biomedical applications:* Video processing can help medical practitioners in their diagnoses, as well as researchers working in various biological fields with automated analysis of larger volumes of data that is being collected, e.g., time-lapse microscopy images.

## 4.13.3 Overview of chapters

The chapters in this section span the broad spectrum of topics that have been identified above. While it is not a comprehensive review of all possible approaches, the section encapsulates some of the major trends in video processing. It provides a sampling of ideas that include the effects of the human visual system on video search, mathematical techniques from signal processing, and systems theory for video modeling, application domains like biometrics, rendering, and surveillance, core video analysis tasks like video tracking, and large multi-camera systems analysis and evaluation.

Chapter 14 titled *Foveated Image and Video Processing and Search* explains the role of foveation in the human visual system and how an understanding of this phenomenon can be useful for designing efficient detection and search systems. It is the only chapter that relates the human visual system with engineering applications in video analysis. The authors, Floren and Bovik, describe methods for modeling the effects of fixation, which can lead to certain interesting areas of the video to be represented at a higher resolution. Applications domains including teleconferencing, teleoperation, wide band imaging, and search are described, among others.

Rodriguez and Vijaya Kumar provide an overview of biometric recognition in Chapter 15 titled *Segmentation-Free Biometric Recognition Using Correlation Filters*. As noted earlier, low-level errors like registration and segmentation affect higher-level recognition tasks and reduce their performance. The authors in this chapter provide an overview of correlation filters that can achieve segmentation-free recognition. Applications are shown on recognizing people using their eye regions, recognizing faces, localizing pedestrians, and recognizing their activities.

Analysis of videos relies on understanding the dynamics in the scene. It is, therefore, natural to leverage upon a large body of work in the systems sciences on dynamical modeling of video sequences.

Chapter 16 on *Dynamical Systems in Video Analysis*, authored by Doretto, Ravichandran, Vidal, and Soatto, provides an overview of linear dynamical systems and their application in video analysis. The role of dynamic textures is studied, including learning the model parameters from natural videos, and synthesizing novel videos based on these models. Low-level video processing tasks like registration and segmentation are addressed.

One of the important application domains of video analysis is image-based rendering. As defined in Chapter 17, titled *Image-Based Rendering* by Chang and Chen, it is "a process of synthesizing images at novel viewpoints based on a set of existing images." The authors take a signal processing approach to what is traditionally a computer graphics problem, by posing it as a sampling and reconstruction problem of the plenoptic function which is used to represent the light field. The chapter starts with a historical perspective on the problem, describes the main challenges, and outlines a solution framework and application domains.

Video surveillance is the most direct application of video processing methods. Given that many tens of millions of surveillance cameras are being sold all over the world (over 30 million just in the US over the last decade), it is necessary to design effective methods to search through such datasets. Effective exploration and exploitation of such "big data" video repositories is the theme of Chapter 18 titled *Activity Retrieval in Large Surveillance Videos* by Castanon, Jodoin, Saligrama, and Caron. The authors define the challenges and then present an approach along with a thorough experimental evaluation.

Tracking is probably the most basic task for analysis of videos. It is often the theme on which the maximum number of papers is presented at major conferences in video analysis. Chapter 19, titled *Multi-Target Tracking in Video* by Poiesi and Cavallaro, provides a summary of the major research challenges that keep this topic at the forefront of work in video analysis, and presents a description of the various steps that are needed to develop efficient tracking algorithms. Both the batch and sequential approaches are described. The recent trends on exploiting contextual information for effective tracking are addressed at the end of the chapter.

A challenge that comes up again and again in current video analysis research is the preponderance of data that is being collected by large-scale sensor deployments. While other chapters in this section have addressed this problem from the perspective of efficient processing, Chapter 20, titled *Compressive Sensing for Video Applications* by Veeraraghavan, Sankaranarayanan, and Baraniuk, argues for the development of a "scalable theory of sensing." Building upon the recently developed theory of compressive sensing, they demonstrate the interplay of efficient signal models, imaging architectures, and signal recovery algorithms for designing effective visual sensing systems.

The last chapter in this section relates to an area that is of high interest currently. As large networks of cameras are deployed, it is important to develop methods that would scale up to these levels. While research in camera networks picks up, it is also important to understand how the developed algorithms would be evaluated since it is difficult to access a real-life large-scale camera network. In Chapter 21, *Virtual Vision for Camera Networks Research*, the authors Qureshi and Terzopoulos propose the development of a simulation paradigm that would enable efficient and rapid development of intelligent surveillance systems involving large numbers of camera spread over wide areas.