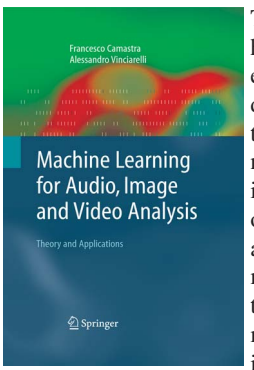
BOOK REVIEW

Machine Learning for Audio, Image and Video Analysis

Francesco Camastra and Alessandro Vinciarelli, 496 pp., ISBN 978-1-84800-006-3, Springer (2008), $109.00 hardcover.

Reviewed by Jie Yu, Kodak Research Labs, Rochester, New York



The past decade has witnessed an explosion of digital information and a phenomenal growth in the popularity of audio, image, and video multimedia. Due to the advancements in processing power and the proliferation of the Internet, people can easily capture, store, transmit, and share audio, image, and video content. However, efficient and effective indexing and retrieval of such data accumulated over time is still a formidable challenge. Researchers in industry and academia have spent a tremendous effort on developing sophisticated systems for processing and understanding this new information, at the core of which always lie machine-learning techniques. As a very broad research field, machine learning covers a broad set of areas ranging from uncertainty analysis to kernel methods. Although it closely interacts with other fields such as statistics, signal processing, and pattern recognition, it is almost impossible for the researchers in those areas to understand every detail of the state-of-the-art machine-learning techniques. While most of the books on machine learning cover classic techniques such as neural networks or support vector machines, few of them emphasize the recent advancements and their applications to audio, image, and video analysis. The book *Machine Learning for Audio, Image and Video Analysis* intends to fill this gap by bringing to its readers the latest developments in this fast-growing field.

The book consists of an introduction, three main parts (total of 13 chapters), and four appendices. The introduction explains how readers with various backgrounds can benefit from studying this book and provides the ABCs about acquisition and processing of audio and visual information for beginners. The appendices also prepare supplemental material for readers who lack the related statistical and signal-processing background. For experienced researchers, the recent advancements can be found in Part II, where classic machine-learning techniques are discussed along with their state-of-the-art developments. For practitioners, the authors analyze three typical applications to provide a sense of how machine-learning techniques can be applied to understand audio/video data.

Part I contains a brief description of how the human biological system perceives audio and video signals and how these signals are captured and digitized in a format that is amenable to computer processing. Chapters 2 and 3 further introduce the audio/video representation and coding standards without getting into too much detail. Besides color, texture, and shape, it is my opinion that local descriptors such as scale-invariant feature transform (SIFT) should be introduced at this point because they have demonstrated the strength of these "bag of words" representations in many applications.

Widely used machine-learning techniques are the focus of Chapters 4 to 11 in Part II. Chapters 4, 5, and 7 introduce the general objectives and approaches of machine learning and the means to evaluate its performance from a statistical point of view. Chapter 5 introduces the Bayesian decision theory, which leads to further investigation of Markovian models in Chapter 10. Kernel methods are discussed in detail in Chapter 9. It is worth mentioning that this book, unlike most other books in this field, not only introduces a few widely used techniques in audio and image analysis, but also discusses the latest advancements in the field. For example, most books would touch the surface of support vector machines (SVM) by introducing the original two-class SVM, yet Chapter 9 of this book goes one step further to discuss sequential minimal optimization, a powerful multiclass extension of SVM, which is more appealing in practice. Chapters 6 and 11 are concerned with clustering and dimension reduction via unsupervised learning. Specifically, Chapter 11 introduces several manifold learning techniques, such as locally linear embedding (LLE) and ISOMAP, which are particularly useful in handling nonlinear data in audio and video processing. Chapter 8 combines the discussion of classic neural networks and ensemble methods. My personal view is that ensemble methods, such as AdaBoost and random forest, deserve an independent chapter because of their outstanding reported performance on many applications and benchmark data sets. Similarly, the "topic model" methods, such as probabilistic latent semantic indexing (pLSI) and latent dirichlet analysis (LDA), should be added to this book to reflect the current research trend in analyzing text and visual data.

Part III showcases three applications of machine-learning techniques, namely speech and handwriting recognition, automatic face recognition, and video segmentation and keyframe extraction. In Chapter 13, the authors discuss the automatic face-recognition system. However, face image localization—one of the core problems in this application—doesn't seem to get enough attention. It would be desirable to add a section about the boosted cascade method proposed by Viola and Jones, which is one of the most successful machine-learning applications in image analysis. In addition, it is better to point out that the eigenface and its variants can suppress a lot of the luminance variation of the face images by removing the eigenvectors corresponding to the three largest eigenvalues.

There are several things that are unique in this book. In some chapters, the problem sections are included to challenge the readers to understand the discussed methods or apply them to solve some sample problems. Distinct from other books, it also points out several public software packages and benchmark data sets that encourage the reader to have a hands-on experience on how machine-learning techniques work to analyze audio and visual content. Its comprehensive coverage on recent development in this research area makes it easy for experienced researchers to further explore the latest techniques.

Compared with classic textbooks such as *Machine Learning* by Tom Mitchell (McGraw Hill, 1997) and *Pattern Classification* by R. Duda, P. Hart, and D. Stork (Wiley-Interscience, 2000), this book is more specialized and focuses on the machine-learning techniques that are applicable to audio and visual processing. Therefore, it is ideal as a textbook or supplemental material for senior graduate courses or advanced topic seminars.

Jie Yu joined Kodak Research Labs in 2007 as a research scientist. He received his PhD in computer science at the University of Texas at San Antonio and his BE in electrical engineering from Dong Hua University. His research interests include multimedia information retrieval, machine learning, computer vision, and pattern recognition. He has published over 20 journal articles, conference papers, and book chapters in these fields. He received several technical awards, which include the Student Paper Contest Winner Award of IEEE ICASSP 2006 and the Best Poster Paper Award of ACM CIVR 2008. He is a member of IEEE and ACM.