Preface

Understanding of Complex Visual Content is essential for a wide range of important applications including automatic multimedia content indexing and retrieval, medicine, robotics, or surveillance. This is a difficult problem due to what we call the "semantic gap" or the distance between the raw representation of image or video contents (bit streams) and the concepts and relations between them that are meaningful and useful for human beings.

Many approaches rely on the joint use of content representation and supervised machine learning techniques though the recent approaches like deep learning now attempt to do both at once. Many alternative and complementary content representation and machine learning techniques now exist. While some single or elementary representation/classification combinations perform relatively well, none of them is currently able to capture and fully exploit the raw media stream for understanding its contents. Research is still trying to explore the best single combinations and their improvements in several complementary directions and the fusion of such elementary combinations is a way of further improving the overall system performance.

This book focuses on the fusion problem in a variety of domains and applications. It follows the workshop on Information Fusion in Computer Vision for Concept Recognition held jointly with the 12th European Conference on Computer Vision (ECCV2012). It contains extended versions of works presented in this workshop together with other works carried out by leading researchers in the domain. The different chapters cover many aspects of the problem and describe successful approaches evaluated in the context of international benchmarks that model realistic use cases at significant scales.

Visual and multi-modal scene understanding by humans is a result of high level interpretation of quantities of information we gather by different physiological channels. We are sensitive to colors, contrasts, motion, visual "roughness," "granularity," loudness of sounds and their nature, etc. We are fusing these different sources to recognize and understand the content. In computer vision and multimedia nowadays, we are imitating this fusion process at different levels. We are speaking of "early," "late," and "intermediate" fusion for scene understanding.

Under "early fusion," we usually understand building rich feature spaces for content description as well as the transformation of these spaces to get the highest efficiency in the content recognition task.

The "late fusion" term denominates the fusion of results of primary decisions, often using information from a single description subspace. The various combination operators, including cascade classification approaches, are applied for aggregating primary decisions in the overall recognition task.

In the "intermediate fusion" approaches we combine results obtained on description subspaces, often coming from different modalities.

Today, the research community in computer vision and multimedia can benchmark their methods on large datasets in the scope of evaluation campaigns, such as ImageCLEF, TRECVid, Pascal VOC, MediaEval, etc. These competitions show the interest and ever-growing performances of late fusion schemes.

The book is organized as follows. In Chaps. 1–3 we are interested in the late fusion approaches for concept recognition in images and videos. A specific accent is made on the study, in Chap. 2, of a very popular model of visual content, namely Bag-of-(Visual)-Words and various fusion aspects which are analyzed in this framework.

Chapter 4 presents an ever-growing trend in the interpretation of visual content by incorporating models of Human Visual System with content understanding methods. Here we are also speaking about fusion. To delimit the areas of potential attention in the so-called bottom-up image-driven manner, multiple cues have to be fused: motion, contrast, and geometry of scenes. The approach is also incorporated in the classical Bag-of-(Visual)-Words model.

In Chap. 5 fusion schemes are developed for a more focused task, such as example-based event recognition in video. Multi-modal features of different semantic levels, such as Bag-of-(Visual)-Words, motion features, audio features, but also results of semantic concepts detections are fused together to recognize events of interest. The interesting conclusion on a good performance of simple fusion operators, such as linear combination of intermediate classification results, is given.

Analyzing a very rich state-of-the art research in computer vision in the matter of scene understanding, one can roughly say that the most efficient approaches follow a threefold scheme: content description, classification, and fusion. All of them are important, nevertheless, the classification approach is the core. In Chap. 6, rotation-based ensemble classifiers for high-dimensional data are proposed, which encourage both individual accuracy and diversity within the ensemble simultaneously.

Chapters 7–9 are more application-focused and present the search of optimal strategies of fusion in such applications as video surveillance, violent content detection in movies, and biomedical information retrieval.

Information fusion is a model of human interpretation of complex visual content. Nevertheless, we are very far from saying today that the mechanisms of content understanding by humans are fully explored. We are only at the beginning in modeling the process. This is why we need to study on a large scale how humans interpret the content. The last Chap. 10 of the book is devoted to this key question.

We dedicate this book to researchers and students working in the domain of information fusion for complex visual contents understanding or working in other related domains where mentioned techniques can be applied.

We are deeply grateful to all those who have helped us to successfully organize the workshop and to produce this book. Special thanks to the reviewers who have contributed to the high quality of the work presented here. We are indebted to all authors for their contribution and hope that the readers of the book will appreciate their hard work and enjoy the reading. Finally, we thank our editor, Springer, who gave us the opportunity to bring this project to life.

Bogdan Ionescu
Jenny Benois-Pineau
Tomas Piatrik
Georges Quénot

http://www.springer.com/978-3-319-05695-1

Fusion in Computer Vision
Understanding Complex Visual Content
Ionescu, B.; Benois-Pineau, J.; Piatrik, T.; Quénot, G.
(Eds.)
2014, XIV, 272 p. 74 illus., 65 illus. in color., Hardcover
ISBN: 978-3-319-05695-1