Visual Event Detection

THE KLUWER INTERNATIONAL SERIES IN VIDEO COMPUTING

Series Editor

Mubarak Shah, Ph.D.
University of Central Florida
Orlando, USA

Video is a very powerful and rapidly changing medium. The increasing availability of low cost, low power, highly accurate video imagery has resulted in the rapid growth of applications using this data. Video provides multiple temporal constraints, which make it easier to analyze a complex, and coordinated series of events that cannot be understood by just looking at only a single image or a few frames. The effective use of video requires understanding of video processing, video analysis, video synthesis, video retrieval, video compression and other related computing techniques.

The Video Computing book series provides a forum for the dissemination of innovative research results for computer vision, image processing, database and computer graphics researchers, who are interested in different aspects of video.

VISUAL EVENT DETECTION

NIELS HAERING
DiamondBack Vision, Inc
11600 Sunrise Valley Drive
Reston, VA 20191
USA

NIELS DA VITORIA LOBO
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816
USA

Springer-Science+Business Media, B.V.

Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN 978-1-4419-4907-3 ISBN 978-1-4757-3399-0 (eBook)
DOI 10.1007/978-1-4757-3399-0

Copyright © 2001 **by** Springer Science+Business Media Dordrecht
Originally published by Kluwer Academic Publishers in 2001.
Softcover reprint of the hardcover 1st edition 2001

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Springer-Science+Business Media, B.V.

Printed on acid-free paper.

Contents

Series Foreword

Traditionally, scientific fields have defined boundaries, and scientists work on research problems within those boundaries. However, from time to time those boundaries get shifted or blurred to evolve new fields. For instance, the original goal of computer vision was to understand a single image of a scene, by identifying objects, their structure, and spatial arrangements. This has been referred to as *image understanding.* Recently, computer vision has gradually been making the transition away from understanding single images to analyzing image sequences, or *video understanding.* Video understanding deals with understanding of video sequences, e.g., recognition of gestures, activities, facial expressions, etc. The main *shift* in the classic paradigm has been from the recognition of static objects in the scene to motion-based recognition of actions and events. Video understanding has overlapping research problems with other fields, therefore *blurring* the fixed boundaries.

Computer graphics, image processing, and video databases have obvious overlap with computer vision. The main goal of computer graphics is to generate and animate realistic looking images, and videos. Researchers in computer graphics are increasingly employing techniques from computer vision to generate the synthetic imagery. A good example of this is image-based rendering and modeling techniques, in which geometry, appearance, and lighting is derived from real images using computer vision techniques. Here the *shift* is from *synthesis* to *analysis followed by synthesis.* Image processing has always overlapped with computer vision because they both inherently work directly with images. One view is to consider image processing as low-level computer vision, which *processes* images, and video for later analysis by high-level computer vision techniques. Databases have traditionally contained text, and numerical data. However, due to the current availability of video in digital form, more and more databases are containing video as content. Consequently, researchers in databases are increasingly applying computer vision techniques to analyze the video before indexing. This is essentially *analysis followed by indexing.*

Due to the emerging MPEG-4, and MPEG-7 standards, there is a further overlap in research for computer vision, computer graphics, image processing, and databases. In a typical model-based coding for MPEG-4, video is first *analyzed* to estimate local and global motion then the video is *synthesized* using the estimated parameters. Based on the difference between the real video and synthesized video, the model parameters are *updated* and finally *coded* for transmission. This is essentially *analysis followed by synthesis, followed by model update, and followed by coding.* Thus, in order to solve research problems in the context of the MPEG-4 codec, researchers from different video computing fields will need to collaborate. Similarly, MPEG-7 will bring together researchers from databases, and computer vision to specify a standard set of descriptors that can be used to describe various types of multimedia information. Computer vision researchers need to develop techniques to automatically compute those descriptors from video, so that database researchers can use them for indexing.

Due to the overlap of these different areas, it is meaningful to treat *video computing* as one entity, which covers the parts of computer vision, computer graphics, image processing, and databases that are related to video. This international series on *Video Computing* will provide a forum for the dissemination of innovative research results in video computing, and will bring together a community of researchers, who are interested in several different aspects of video.

Mubarak Shah
University of Central Florida

Orlando

Preface

In this book we argue in favor of a bottom-up approach to object recognition and event detection. The underlying principle of the book is that many diverse pieces of evidence are more useful for object recognition and event detection than the most elaborate algorithm working on an impoverished image representation based on, say, edge information. Our approach is motivated by the data processing theorem, which states that the real world possesses a certain amount of information, only part of which we can hope to measure and extract. Processing the extracted information is leaking further information about the world. David Marr's [74] principle of least commitment and Rodney Brooks' [15] subsumption architecture are instances in Computer Vision and Robotics where researchers have consistently applied the data processing theorem.

We present a framework for the detection of visual events from video that is based on three principles, derived from the data processing theorem:

- Extract **rich image descriptions** to provide an expressive internal description of the world.
- **Process the extracted information in a fat, flat hierarchy.**
- **Missing information will prevent crucial inferences**, while additional information disguises relevant information in the worst case.

The structure of this book reflects these principles. We discuss the extraction of diverse image descriptions and their fusion in a fat and flat hierarchy into object recognizing, shot summarizing, and event detecting components. Rich image descriptions based on many diverse sources of information, such as color, spatial texture, and spatio-temporal texture measures greatly simplify object recognition and event detection tasks.

We present object recognition and event detection results for a number of applications and contrast our framework and its components with alternative solutions.

Acknowledgments

Thanks, Wen-Lin for your help, trust, time, and patience; and thank you, Niels Lobo, Mubarak Shah, and Richard Qian for your guidance and support.