Introduction to the Special Issue on Video Object Processing for Surveillance Applications

Aishy Amer\textsuperscript{a} and Carlo Regazzoni\textsuperscript{b}

\textsuperscript{a}Department of Electrical and Computer Engineering, Concordia University, Montréal, Québec, Canada. E-mail: amer@ece.concordia.ca

\textsuperscript{b}Department of Biophysical and Electronic Engineering, University of Genoa, Italy. E-mail: carlo@dibe.unige.it

Introduction

The automated security market is growing at a constant and high rate that is expected to sustain for decades [DataMonitor (2004)]. For example, the annual growth rate of U.S. “Homeland Security” spending from 2003 to 2010 is estimated to be about 40\% [Inbar (2004)].

Video-based surveillance (or video surveillance) is one of the fastest growing sectors in the security market. This is due to the high amount of useful information that can be extracted from a video sequence. In particular, the automatic real-time processing of video objects, i.e., the extraction of video objects and related high-level content, is hereby of paramount importance. High-level video content, e.g., object activities and events, are generally related to the movement of video objects. This is related to the human visual system (HVS) which is strongly attracted to moving objects creating luminance change [Nothdurft (1993); Franconeri and Simons (2003); Abrams and Christ (2003)].

The main components of an automatic video surveillance system are shown in Fig. 1. A video surveillance system consists of video cameras connected to a video processing unit (either a general purpose PC or a dedicated computer equipment) to extract useful information that could be identified with some alert situation. This processing unit could be connected throughout a network to a control and visualization center that decides on the appropriate measures to take to manage extracted information (e.g., to handle alerts). As high-level object features occurs, alerts are sent to the control and visualization center. Video objects and events maybe stored for future references and retrieval.
Video processing for surveillance applications aims at describing the data in successive images of a video in terms of what is in the real scene, where it is located, when it occurred, and what are its features. It is the basic step towards an automated full understanding of the semantic contents in the input video. A generic block diagram of a video object processing module for video surveillance is shown in Fig. 2. As can be seen, various video processing steps are required to extract video objects and related high-level features: preprocessing (e.g., noise estimation or reduction), video analysis (i.e., extraction of video objects using object segmentation, motion estimation, and object tracking), object classification and recognition, video interpretation (i.e., extraction of high-level context-independent object information), and video understanding (i.e., extraction of context-dependent semantical object information).

Fig. 2. Video object processing for video surveillance.
The differentiation between video interpretation (context independence) and video understanding (context dependence) aims at facilitating the processing by dividing each processing level into simple but effective tasks. High-level video content can be divided into a context-independent component (or a fixed meaning) and a context-dependent component. For example, the event \textit{removal} has a fixed context-independent meaning that is common to all applications: “the object is removed from the scene” where the content is detected independently of where and what is being removed. But \textit{removal} may have a variable meaning in different contexts such as when “the object is removed at a significant time or location”. Context-independent content can be used in a wide range of surveillance application whereas context-dependent features are related to specific applications and are thus not generally applicable.

**Fundamental issues and challenges**

Effective video processing remains a difficult task despite the many advancements in the field. This difficulty originates from several issues that can complicate the design and evaluation of processing algorithms for video surveillance.

**Interpretation** Object-oriented video processing aims at extracting video objects and their spatio-temporal features. To extract object, technically the following are given: 1) a video is a finite set of images and each image consists of an array of pixels; 2) the aim of analysis is to give each pixel a label based on some properties; and 3) an object consists of a connected group of pixels that share the same label. The technical definition of an object may not be, however, one that video interpretation needs. For instance, does interpretation consider a vehicle with a driver one object or two objects? a person moving the body parts one or more objects? These questions indicate that there is no single object-oriented video processing method that is valid for all applications. Object processing is subjective and can vary between observers and the evaluation of one observer can vary in time. This subjectivity cannot always be formulated by a precise mathematical definition of an analysis concept that also humans cannot define uniquely. For some application, the use of heuristics is an unavoidable part of solution approaches [Pavlidis (1977); Correia and Pereira (1998)].

**Generality** Much research has been concerned with the development of object processing methods to be applied in general applications. Specific applications require, however, specific parameters to be fixed and even the designers of general systems can have difficulty adapting the system parameters to a specific application. Therefore, many research efforts are dedicated to develop object processing methods that focus on a well-defined range of specific applications.

**Automation** For a video processing algorithm to be useful in a surveillance
system it should allow automatic selection of algorithm’s parameters (or thresholds) to control its performance under different conditions.

**Efficiency** For a video processing algorithm to be useful in a surveillance system it should be real-time, i.e., outputs informations such as events as they occur in the real scene. Although progress in micro-electronics has enabled increasingly more sophisticated video processing techniques to be exacted in real-time, computation complexity aspect of new techniques remains crucial for a wide applicability of these new techniques. In general, the wide use of an analysis tool strongly depends on its computational efficiency [Garrido et al. (1998)].

**Robustness** A video processing algorithm is required to be robust, i.e., deliver good performance for a give application. Achieving robust algorithms is a challenge especially (a) under illumination variation due to weather conditions or indoor lighting changes, for example, in the presence of open doors or windows, (b) in case of shadow and reflections, (c) multiple objects with partial or complete occlusion. Robustness remains a major issue to avoid system failure.

**Trade-off: accuracy versus efficiency** A video processing task is complicated by various image and object changes, such as noise, artifacts, clutter, illumination changes, and object occlusion. This complicates further two conflicting requirements: accuracy and efficiency. In a surveillance application, for instance, the emphasis is on the robustness of the video processing task (e.g., no system failure) with respect to varying image and object conditions rather than on accurate and precise object extraction. In object-based retrieval, on the other hand, obtaining pixel-accurate objects is not necessary but the representation of objects must have some meaning. Beside accuracy and robustness, algorithm complexity has an impact on the design of video processing modules. Off-line applications, such as object-based coding, tolerate analysis algorithms that need processing power and time. Other applications such as surveillance require real-time analysis.

**Feature selection and filtering** A key difficulty in selecting features to solve a video processing task, such as segmentation or matching, is to find useful features that stay robust throughout an image sequence. Main causes for inaccuracy are sensitivity to artifacts, object occlusion, object deformation, articulated and non-rigid objects, and view change. The choice of these features and their number varies across applications and within various tasks of a processing algorithm. In some tasks, a small number of features is sufficient, in other tasks a large number of features may be needed. A general rule is to select features that do not significantly change over a video and that can be combined to compensate for each other’s weakness. For example, the most significant features can be noisy and thus difficult to analyze, therefore requiring a filtering procedure to exclude these from being used in a processing task.

**Feature integration** Since features can be noisy, incomplete, and variant, the issue is to find ways to effectively combine these features for robust
processing. The widely used methods for feature integration are linear. The HVS performs many vision tasks, however, in a non-linear way. In high-level video processing, HVS-oriented integration is needed.

**Multiple camera processing** Another challenge arise in extracting video objects and related meaning across multiple cameras.

**Performance evaluation** This is a crucial issue for the acceptance of video surveillance systems in real world applications. Due to the difficulties of evaluating video processing in general, this is an open problem that has gained many research activities. For example, some progress has been made towards defining a standard set of test data [PETS (2001-2005)].

**In This Special Issue**

This special issue brings together expertise of outstanding researchers in different video object processing topics for surveillance applications. We received 16 submissions for this issue and after a rigorous reviewing process, we selected six papers to illustrate relevant solutions to fundamental challenges in video surveillance research. Two common features of the selected solutions are their suitability for real-time video surveillance and the handling of video objects.

In this special issue, the first paper proposes a solution to object segmentation over a long period of time; the two following papers deal with object tracking where the first paper introduces semantic to support tracking and the second paper uses non-prior training active feature model for tracking. The fourth paper is on the detection of cyclic human activities based on inter-frame similarity matrix. The last two papers handle the detection of event in video sequences where the first paper applies to car parking and the second in case of context-independent processing.

Kim *et al.* present in their paper “Real-Time Foreground-Background Segmentation using Codebook Model” an algorithm for foreground-background segmentation where background values at each pixel are quantized into codebooks which represent a compressed form of background model for a long image sequence. This allows to capture structural background variation due to periodic-like motion over a long period of time under limited memory. For performance evaluation, they have applied perturbation detection rate analysis to four background subtraction algorithms and two videos of different types of scenes.

Jones *et al.*, in their paper “Learning the Semantic Landscape: Embedding scene knowledge in object tracking”, introduce a method to automatically parametrize the semantic landscape for object tracking. The accuracy of object tracking methodologies can be significantly improved by utilizing knowledge
about the monitored scene. They demonstrate how, over a sufficient length of
time, observations from the monitored scene itself can be used for such param-
eterization. The accuracy of object processing can be significantly improved
by utilizing knowledge about the monitored scene.

Kim et al. proposes a method to “Optical Flow-Based Real-Time Object
Tracking Using Non-Prior Training Active Feature Model”. They present a
feature-based object tracking algorithm using optical flow under the non-prior
training active feature model framework. The proposed algorithm aims at
tracking both rigid and deformable objects, and at robustness against object’s
sudden motion. For that they use both a feature point and the corresponding
motion direction at the same time for tracking and separate feature points
inside an object from background. They also handle heavy object occlusion.
The training set used for model fitting can be updated at each frame to make
more robust object’s features under occluded situation.

Branzan-Albu et al. in “Detection of Cyclic Human Activities Based on the
Morphological Analysis of the Inter-Frame Similarity Matrix” describe a method
for the temporal segmentation of periodic human activities from continuous
real-world indoor video sequences acquired with a static camera. The pro-
posed approach is based on the concept of inter-frame similarity matrix. The
pattern associated with a periodic activity in the similarity matrix is decom-
posable into elementary units. They propose a morphology-based approach
for the detection and analysis of activity patterns. The approach used for ex-
perimental evaluation is based on a statistical estimation of the ground truth
segmentation and on a confidence ratio for temporal segmentations.

Diamantopoulos et al. introduce a system for “Event Detection for Intelligent
Car Park Video Surveillance”. This system is applied for detecting tailgating,
an example of complex interactions and activities within a vehicle parking
scenario. They use an adaptive background learning algorithm and intelligence
to overcome the problems of object masking, separation and occlusion.

Amer et al. present a system for “Rule-Based Real-Time Detection of Context-
Independent Events in Video Shots” where they reason with video objects’
low-level features and mid-level features, to infer events related to moving ob-
jects. This is done by continually monitoring changes and behavior of features
of video objects. When certain conditions are met, events are detected. Their
goal is to detect generic events, i.e., events that are independent of the context
of where or how they occur. They classify events into four types: primitive,
action, interaction, and composite.
Acknowledgments

Special thanks are due to the Editor-In-Chief, Nasser Kehtarnavaz, of the Journal for real-time imaging for offering us the opportunity to edit this special issue and for handling the paper of Amer et al. in this special issue. We express our sincere thanks to the reviewer that without their help this issue would not have been possible. Our thanks goes also to the authors of the articles as well as to Sarah XXXX, who provided coordination and logistical support.

Aishy Amer
Carlo Regazzoni

References