Computer Vision and Image Understanding 144 (2016) 258–275

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A scalable and flexible framework for smart video surveillance

Antonio C. Nazare Jr.*, William Robson Schwartz

Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 26 December 2014
Accepted 21 October 2015

Keywords:
Smart surveillance framework
Surveillance systems
Computer vision
Video analysis
Video surveillance

ABSTRACT

In the last years, the number of surveillance cameras placed in public locations has increase vastly and as consequence, a huge amount of visual data is generated every minute. In general, this data is analyzed manually, a challenging task which is labor intensive and prone to errors. Therefore, automatic approaches must be employed to enable the processing of the data, so that human operators only need to reason about selected portions. Computer vision problems focused on solving problems in the domain of visual surveillance have been developed aiming at finding accurate and efficient solutions. The main goal of such systems is to analyze the scene focusing on the detection and recognition of suspicious activities performed by humans in the scene, so that the security staff can pay closer attention to these preselected activities. However, these systems are rarely tackled in a scalable manner. Before developing a full surveillance system, several problems have to be solved, which are usually solved individually. However, in a real surveillance scenario, these problems have to be solved in sequence considering only videos as the input. With that in mind, this work proposes a framework for scalable video analysis called Smart Surveillance Framework (SSF) to allow researchers to implement their solutions to the surveillance problems as a sequence of processing modules that communicates through a shared memory.

1. Introduction

Due to the reduction in prices of cameras and the increase in network connectivity, the number of surveillance cameras placed in several locations increased significantly in the past few years. If on one hand, a distributed camera network provides visual information in real time covering large areas, on the other hand, the number of images acquired in a single day can be easily in the order of billions, preventing their manual processing and posing an intricate problem for monitoring such areas [1].

While the ubiquity of video surveillance provides safer environments, the monitoring of large amount of visual data is a challenging task when performed manually by human operators since most of the visual data do not present interesting events from the surveillance standpoint, turning it into a repetitive and monotonous task for humans [2,3]. Hence, automatic understanding and interpretation of activities performed by humans in videos are of great interest since such information can assist the decision making process of security agents [2].

The addition of automatic understanding and interpretation to surveillance systems does not entail the replacement of human operators as foreseen on several Sci-Fi movies, on the contrary, it aims at supplying information to the operator. For instance, instead of a security agent monitoring continually up to 50 screens with live security video feed (task which humans do not present high performance due to the lack of important events during most of the time [4]), an automated system might perform a filtering in the videos and indicate only those video segments more likely to contain interesting activities, such as suspicious activities that might lead to a crime.

In the last two decades, professionals of industry and researchers have dedicated their studies to improve surveillance systems. To understand the increase of works related to video surveillance and inspired by Huang [5] study, we searched the keywords *video* and *surveillance* in IEEE Xplore Digital Library[1] (within metadata only) and the IEEE Computer Society Digital Library[2] (by exact phrase). The findings are shown in Fig. 1 as a function of the publication year. The large number of publications in the past ten years indicates that research on surveillance video has been very active.

Smart visual surveillance systems deal with the real-time monitoring of objects within an environment. The main goal of these systems is to provide automatic interpretation of scenes and understand activities and interactions of the observed agents based on the visual information being acquired. Current research regarding these

* Corresponding author.
E-mail addresses: antonio.nazare@dcc.ufmg.br, acnazarejr@gmail.com (A.C. Nazare Jr.), william@dcc.ufmg.br (W.R. Schwartz).

[1] http://ieeexplore.ieee.org/
[2] http://www.computer.org/csdl

http://dx.doi.org/10.1016/j.cviu.2015.10.014
1077-3142/

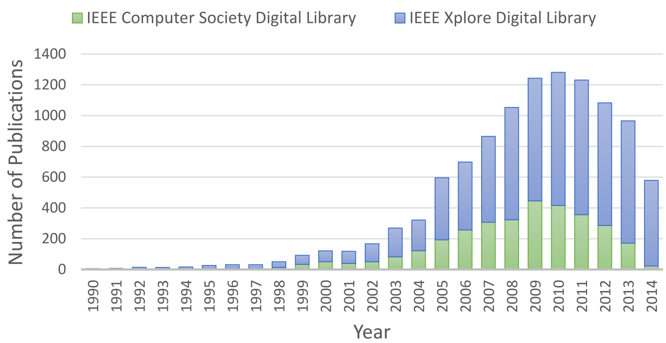

Fig. 1. Histogram of publications in IEEE Computer Society Library and IEEE Xplore Digital Library whose metadata contains the keywords *video* and *surveillance* (Adapted from [5] and updated on December 17, 2014).

automated visual surveillance systems tend to combine multiple disciplines, such as computer vision, signal processing, telecommunications, management and socio-ethical studies.

One of the great challenges of automatic surveillance systems is that to interpret what is happening in the scene, a sequence of problems need to be solved, which is highly prone to noise generated during the process. Among the problems are the background subtraction [6], pedestrian detection [7], face recognition [8], gesture recognition [9], pose estimation [10], person tracking [11] and re-identification [12], action recognition [13] and activity recognition [14]. Even thought each one of these problems present a vast literature, they are usually considered independently such as in currently available evaluation data sets, e.g., the evaluation of face recognition methods is performed using already detected, cropped and aligned faces [15], which cannot be accomplished in real surveillance scenarios where the only inputs are video feeds without annotations. Therefore, dealing with the problems individually does not allow one to identify what are the effects of the results obtained by solving one problem on the following steps in the processing sequence.

Although visual surveillance has been subject to a huge growth, there is still a lack of contributions from the field of system engineering to the area [16]. The small number of frameworks that are open and focus on visual surveillance usually require a steep learning curve. In addition, with the contemporary advances in video sensors and increasing availability of network cameras allowing the deployment of large-scale surveillance systems, distributed in a wide coverage area, the design of smart and scalable surveillance system remains a research problem: how to design scalable video surveillance systems considering aspects related to processing power, memory consumption and network bandwidth?

Motivated by the presented issues, this work proposes a framework for a scalable video analysis able to readily integrate different computer vision algorithms into a functional surveillance system. This framework, called *Smart Surveillance Framework (SSF)*, aims at bringing several improvements providing scalability and flexibility, allowing the users (researchers) to focus only on their application by treating the sequence of problems as a set of processing modules that communicates through data streams, stored in a shared memory. This framework was presented in a preliminary and reduced version focusing only on its main concepts in [17] and its feature extraction server was described in [18]. However, this work presents the SSF in details and focuses on the evaluation of its several components[3] In addition, this work also presents new features, such as the data stream and the image and feature managers.

More specifically, the *Smart Surveillance Framework* is a development environment in which the researcher can implement and evaluate his/her algorithms related to surveillance in an integrated manner, as illustrated in Fig. 2. It is based on execution modules that communicate to each other using data streams controlled by a shared memory. The framework provides the following features to aid the researcher: memory management to allow handling large amounts of data in regular computers; communication control among execution modules; predefined data structures specifically designed for surveillance environment; management of multiple data input, such as cameras or stored videos; feature extraction server to maximize the usage of the processing power available to compute local descriptors; query server to allow high level reasoning and scene understanding; and a configuration interface to help setting up sequences of execution. These features and the framework architecture will be discussed in details in Section 4.

The main contributions provided by the development of the SSF are the following: *(i)* A novel framework to allow the processing of large amounts of data provided by multiple surveillance network cameras; *(ii)* A platform to compare and exchange research results in which researchers can contribute with modules to solve specific problems; *(iii)* A framework to allow fast development of new video analysis techniques since one can focus only on his/her specific task; *(iv)* Creation of a high-level semantic representation of the scene using data extracted by low-level modules to allow the execution of video event analysis based on individual or group activities; *(v)* A testbed to allow further development on activity understanding since one can focus directly on using real data, instead of annotated data that may prevent the method from working on real environments;

[3] The Smart Surveillance Framework is available for download at http://www.ssig.dcc.ufmg.br/ssf/.

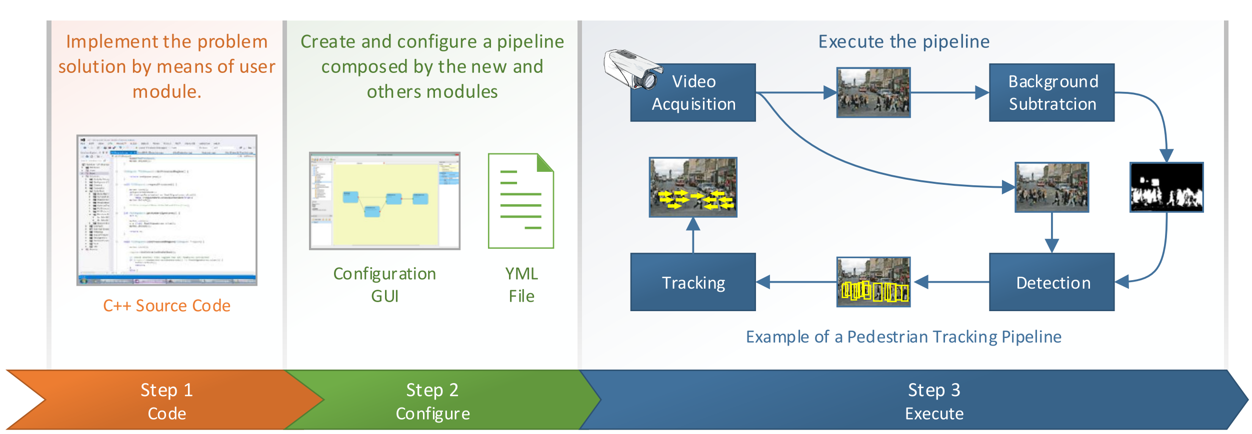

Fig. 2. Overview of the Smart Surveillance Framework, an development environment that allows the researcher to implement his/her surveillance algorithms in an integrated manner by setting up a sequence of modules that will be executed in a pipeline. The researcher can take advantage of the transparent communication control provided by a shared memory using surveillance-focused data structures (Section 4.2), a feature extraction server (Section 4.3) to reduce the computational cost, and a high level reasoning might be performed using information stored in the Complex Query Server (Section 4.4).