Computers and Electrical Engineering 000 (2017) 1–15

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Embedded deep vision in smart cameras for multi-view objects representation and retrieval☆

Jamil Ahmad[a], Irfan Mehmood[a], Seungmin Rho[b], Naveen Chilamkurti[c], Sung Wook Baik[a,*]

[a] College of Software and Convergence Technology, Sejong University, Seoul, Republic of Korea
[b] Department of Software, Sungkyul University, Anyang, Republic of Korea
[c] Computer Science and IT, La Trobe University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 9 November 2016
Revised 19 May 2017
Accepted 19 May 2017
Available online xxx

Keywords:
Embedded processing
Convolutional neural network
Transfer learning
Image retrieval

ABSTRACT

Active large scale surveillance of indoor and outdoor environments with multiple cameras is becoming an undeniable necessity in today's connected world. Enhanced computational and storage capabilities in smart cameras establish them as promising platforms for implementing intelligent and autonomous surveillance networks. However, poor resolution, limited number of samples per object, and pose variation in multi-view surveillance streams, make the task of efficient image representation highly challenging. To address these issues, we propose an efficient and powerful convolutional neural network (CNN) based framework for features extraction using embedded processing on smart cameras. Efficient, high performance, pre-trained CNNs are separately fine-tuned on persons and vehicles to obtain discriminative, low dimensional features from segmented surveillance objects. Furthermore, multi-view queries of surveillance objects are used to improve retrieval performance. Experiments reveal better efficiency and retrieval performance in different surveillance datasets.

1. Introduction

In wake of the rising security concerns, visual surveillance with multiple cameras have recently emerged as an interesting and challenging application, which generates huge volumes of video data rapidly. Timely and accurate access to semantic content in surveillance videos is a fundamental objective of surveillance systems. However, due to the huge volumes of multimedia data and the inherent complexity of visual contents, it still remains a challenging task. Typical surveillance systems sense the environment using multiple cameras and transmit video data to one or more servers where it is stored for future inspection and browsing. Locating and identifying particular objects of interest (e.g. person or vehicle) in surveillance data is a common but intricate task, with numerous applications. The massive volumes of multimedia data makes it very cumbersome for manual browsing and inspection. Hence, automated methods for analysis, processing, and searching surveillance streams are highly desirable.

Multiple cameras are usually deployed to obtain wide coverage of sensitive areas, where the field of view is partially overlapped, allowing multiple views of the same objects be captured. In such a setting, it becomes vital to collect and

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. A. K. Sangaiah.

* Corresponding author.
E-mail addresses: jamil.ahmad@icp.edu.pk, jamilahmad@sju.ac.kr (J. Ahmad), sbaik@sejong.ac.kr (S.W. Baik).

http://dx.doi.org/10.1016/j.compeleceng.2017.05.033
0045-7906/

analyze inputs from multiple visual sensors and generate a semantically meaningful interpretation [1]. Using the features-based representation of surveillance data, automated visual analysis and processing algorithms can be developed. Intelligent surveillance involves automated analysis and interpretation of object behaviors through detection, recognition, tracking, and scene understanding, as well as their indexing and retrieval [2,3]. Though these activities are often performed offline, their online processing can significantly improve usability of surveillance networks. The advancements in embedded computing capabilities in smart cameras have created new opportunities for intelligent surveillance [4].

Intelligent autonomous surveillance highly depends on effective representation and accurate interpretation of visual contents in real-time. Several attempts have been made in order to develop efficient algorithms and hardware-based implementations to perform these tasks. For instance, Botella et al. [4] introduced a bioinspired visual sensor based on the combination of optical flow and orthogonal moment invariants. The extracted features were used to generate an efficient mid-level visual abstraction, thereby allowing embedded processing on visual sensor nodes. The authors exhibited several tasks on the proposed visual sensor including segmentation, tracking, and motion estimation in real-time. Holte et al. [1] presented a thorough review on the problem of human pose estimation and activity recognition. In addition, they reported performance comparisons of several methods on popular challenging datasets. They also showed that pose estimation and action recognition using multiple views significantly outperforms single view based approaches. Implementing localized processing using smart cameras will lead to timely understanding of activities and events and would be able to generate autonomous responses in case of emergencies to prevent any undesirable delays. Major challenges in realizing this objective include 1) managing huge volume of data sensed by multi-camera surveillance networks, 2) limited processing capabilities of the embedded processors in smart cameras, 3) algorithmic complexity for high level interpretation of surveillance scenarios, and 4) difficulty in representing objects due to viewpoint induced pose and illumination variations. To cope with these issues, an efficient framework is required which allow robust representation of surveillance data in real-time, utilizing the embedded processing in multi-view smart camera networks.

The recent success of deep convolutional neural networks (CNNs) in image classification and object recognition has attracted research community to utilize these multi-layer end-to-end learning architectures for performing a wide variety of tasks. Typical CNNs consist of many convolutional layers, followed by some fully connected layers and a Softmax layer which produces a probability distribution over the training classes. The neuronal activations from the fully connected layers are commonly used as image representation in a variety of applications. Due to the representational capability of CNNs, they are widely used in object retrieval systems [5]. One significant challenge being faced by the use of CNNs in real-time embedded computer vision systems is the high computational cost of these models. Researchers are constantly trying to develop efficient models and the hardware community is aiming to develop efficient and powerful hardware to supports real-time inference using deep models. In this context, several techniques have been investigated to reduce the computations in CNNs for efficient inference including model simplification using smaller kernels, compression, pruning, and quantization. With the advancements in these techniques, we have seen several models small enough to fit the memory of an embedded system on a smart camera or FPGA, yet powerful enough to yield high performance. This achievement has further motivated the computer vision researchers to use these hierarchical architectures for real-time embedded vision in smart cameras.

The objective of this work is to develop an efficient framework for multi-view object retrieval using deep models for embedded computer vision. We have used an efficient CNN architecture inspired by SqueezeNet [6] to perform features extraction in surveillance networks. The pre-trained model is fine-tuned using surveillance datasets to function as a real-time feature extractor in the intelligent surveillance network. These features can be used to perform object recognition, scene understanding, behavior analysis, and event understanding in real-time for intelligent surveillance networks. We propose to utilize these features for indexing of surveillance objects for later retrieval and inspection. Our main contributions in this work are:

1. An efficient cooperative multi-view real-time feature extraction method has been proposed for intelligent surveillance networks.
2. We propose to use a small sized and discriminative deep CNN model for features extraction from surveillance data.
3. A re-ranking strategy is developed to accumulate and re-rank retrieved results from deep features of multiple view queries for improved retrieval performance.

The rest of the paper is organized as: Section 2 introduces the proposed framework and discusses its various components in detail. Section 3 presents results of experimental evaluations of the proposed method on various surveillance datasets. Section 4 concludes the paper with discussions on strengths and shortcomings of the proposed method along with some future directions.

2. Proposed method

Typically, image retrieval systems rely on visual features extracted from single view of an image [7,8]. The inherent absence of appearance information from single view images affect their overall representation. To overcome this deficiency, we utilize the multi-view images of the same objects available in surveillance scenarios. Visual features collected through multiple views of the same objects can help us retrieve images more effectively in surveillance CBIR systems. The overall framework for the proposed scheme is provided in Fig. 1. In the proposed multi-camera scenario of n nodes, n-1 camera nodes act as agent nodes, which capture the scene and remove backgrounds using bootstrapping procedure. The roughly

Please cite this article as: J. Ahmad et al., Embedded deep vision in smart cameras for multi-view objects representation and retrieval, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.05.033