Information Fusion 24 (2015) 16–30

Contents lists available at ScienceDirect

Information Fusion

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

Saliency-directed prioritization of visual data in wireless surveillance networks

Irfan Mehmood [a], Muhammad Sajjad [a], Waleed Ejaz [b], Sung Wook Baik [a,*]

[a] College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea
[b] Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada

ARTICLE INFO

Article history:
Available online 11 August 2014

Keywords:
Image prioritization
Wireless visual sensor networks
Monitoring applications
Salient activity detection

ABSTRACT

In wireless visual sensor networks (WVSNs), streaming all imaging data is impractical due to resource constraints. Moreover, the sheer volume of surveillance videos inhibits the ability of analysts to extract actionable intelligence. In this work, an energy-efficient image prioritization framework is presented to cope with the fragility of traditional WVSNs. The proposed framework selects semantically relevant information before it is transmitted to a sink node. This is based on salient motion detection, which works on the principle of human cognitive processes. Each camera node estimates the background by a bootstrapping procedure, thus increasing the efficiency of salient motion detection. Based on the salient motion, each sensor node is classified as being high or low priority. This classification is dynamic, such that camera nodes toggle between high-priority and low-priority status depending on the coverage of the region of interest. High-priority camera nodes are allowed to access reliable radio channels to ensure the timely and reliable transmission of data. We compare the performance of this framework with other state-of-the-art methods for both single and multi-camera monitoring. The results demonstrate the usefulness of the proposed method in terms of salient event coverage and reduced computational and transmission costs, as well as in helping analysts find semantically relevant visual information.

1. Introduction

Wireless visual sensor networks (WVSNs) have recently emerged as a new type of sensor-based intelligent system. The performance and complexity challenges of WVSNs transcend those of existing wireless sensor networks. In contrast to scalar sensors, WVSNs capture visual data, which offers rich information. Moreover, owing to their high versatility, small size, and dense spatial coverage, WVSNs can be flexibly deployed in various applications, such as remote patient monitoring, distributed multimedia-based surveillance, and security systems [1–4]. In such setups, visual data about the event area has significant potential to influence decision making, because the WVSNs allow the real-time observation of events occurring within a busy environment. Despite the existence of numerous ideas for surveillance applications, the actual implementation of WVSNs remains a challenge. This is because surveillance systems consist of numerous sensors, and thus require a large amount of bandwidth to transmit their raw video streams. Further, these video streams require extensive processing to detect events and anomalies. Recently, there have been attempts to advance current WVSN technologies by implementing intelligent methods that can automatically extract relevant data from the source node, thus reducing the consumption of network bandwidth.

WVSN-based monitoring systems have two main requirements: robustness and efficient resource utilization. In real surveillance environments, the robustness of the system is often compromised by the failure of camera sensors in unpredictable conditions such as natural catastrophic events, technical malfunctions, and human intrusion. A multi-camera WVSN ensures the robustness of the system by providing visual access to the region of interest, even if a number of cameras have failed, and has the ability to collaboratively share monitoring tasks and minimize the energy consumption of each camera. However, a multi-camera wireless network introduces the problem of full or partial coverage overlaps in the camera's field-of-view (FoV), resulting in the production and transmission of highly redundant imaging data. The processing and transmission of such massive video data to the sink node (SN) causes unnecessary resource consumption in the network.

A wireless surveillance system is subject to a unique set of resource constraints, such as finite on-board battery power and limited network communication bandwidth [1,5]. Wireless transmission is close to its energy efficiency limit [6]. Energy consumption and the probability of packet collision increase when all of the multi-view data is transmitted. Moreover, the huge volume of

* Corresponding author. Tel.: +82 02 3408 3797; fax: +82 02 3408 4339.
E-mail addresses: irfanmehmood@sju.ac.kr (I. Mehmood), sajjad@sju.ac.kr (M. Sajjad), waleed.ejaz@ieee.org (W. Ejaz), sbaik@sejong.ac.kr (S.W. Baik).

http://dx.doi.org/10.1016/j.inffus.2014.07.002
1566-2535/

visual data collected by WVSNs poses a critical hurdle for human analysts, as it becomes challenging to discern actionable intelligence. Thus, it is imperative to leverage visual analytics at the point of data collection to enable the analysis of video content. This requires the development of intelligent and power-efficient methods for extracting semantically relevant data at the source node to reduce network bandwidth consumption. These methods should determine whether the sensed data is informative, and thus requires transmission, or if it is better to wait for useful images before sending any data. The decision must be based solely on the coverage of the phenomenon being monitored.

In this paper, we describe an energy-efficient visual data prioritization framework for multi-camera wireless networks. Our framework is based on image prioritization and the reliable transmission of data to the SN. For the image prioritization, the camera sensors categorize visual content as either salient or non-salient. Saliency is a neurobiological notion that symbolizes the human ability to concentrate mental powers on certain areas by close observation while simultaneously disregarding irrelevant information [7]. For efficient monitoring, the proposed image prioritization calculates salient motion using a low-cost dynamic visual saliency model. This considers parameters such as motion contrast, the salient object's coverage ratio, and the user's preference for a specific camera. Moreover, the transmission of high-priority content to the SN can be made error-free by selecting a reliable channel for transmission. To ensure the selection of a reliable channel, we introduce cooperative spectrum sensing, whereby the sink acts as a fusion center. This framework introduces a ranking strategy that formulates how cooperative sensing selects a reliable channel from among the possible candidates for the efficient use of available bandwidth. This framework significantly prolongs the network lifetime and accelerates the decision-making process by providing concise information at the base station (BS). In the remainder of this paper, the terms "high priority," "important," and "salient", are used interchangeably when referring to information.

The main contributions of the proposed framework are as follows:

- A fully decentralized visual sensor-based intelligent system is proposed. This computes and disseminates aggregated multi-camera data with minimum processing and communication requirements.
- An energy-efficient image prioritization algorithm using a low-cost dynamic visual saliency model is presented. This extracts semantically important frames from the videos, thereby reducing the bandwidth requirements.
- The dynamic visual saliency method is based on the fact that human visual attention is attracted by salient motion, which can be estimated using temporal gradients. To reduce the computation costs of modeling dynamic visual saliency, novel integral-image-based temporal gradients are employed.
- Fully reliable data transmission is ensured by selecting the best available channel using cognitive radio networks.

The remainder of this paper is organized as follows. Section 2 summarizes related work, and Section 3 introduces the proposed framework. Section 4 describes the evaluation scheme and experimental setup, Section 5 concludes the paper.

2. Related work

The development of automated video surveillance using WVSNs has drawn considerable attention in recent years [8–12]. Aware of the limitations of WVSNs and the dynamic nature of the surveillance environment, recent work has focused on certain aspects of visual monitoring, including collaborative data processing, visual data coverage, and visual data coding [13–15]. The main goal of the surveillance system is to use as few camera streams as possible, to avoid possible redundancy in the multimedia data, while still providing the necessary event coverage. In this regard, Dieber et al. [16] focused on the WVSN configuration problem. They presented an efficient approximation method based on an evolutionary algorithm to select an optimal configuration for the cameras. However, in dealing with multi-camera video streams, resource optimization schemes are generally not set up to optimize the computation and data distribution power for such an enormous amount of multi-view visual data.

To overcome the problem of high power consumption while dealing with multi-camera wireless sensors, the optimal area coverage for efficient monitoring and resource utilization has been examined [17]. The authors of [18,19] proposed a solution for optimal FoV coverage by selecting the best subset of camera sensors to capture an event area. Similar work was presented by Johnson and Bar-Noy for active cameras [20]. They utilized the cameras' pan and scan properties to support full view coverage with the minimum number of cameras. Newell and Akkaya [21] proposed a heterogeneous sensor placement strategy that provides full coverage of the monitored areas using scalar and visual sensors. Other studies on area coverage deal with the orientation of the camera sensors and occlusion issues for better coverage of the event area [22,23]. However, these methods must consider critical issues, including the efficient identification of the target location and coordination of the distributed sensors over the network, with limited resources. In addition, no mechanism has been suggested for reliable data transmission, which is crucial in WVSN-based monitoring [24].

Ma et al. [25] mentioned that a majority of the literature has focused on conventional data transmission methods, without considering the resource-constrained dynamic nature of WVSNs. For efficient transmission, Li et al. [26,27] allocated more resources to parts of the video sequence that have a greater impact on the video quality, while assigning fewer resources to unimportant sections. In surveillance videos, humans usually focus on the motion of the objects rather than the background [28]. In general, the motion of objects is computed using the video frame difference, background subtraction, and optical flow [29]. The frame difference scheme is simple and fast to implement, but is not robust to various environmental challenges such as lighting changes, weather conditions (rain, storms), and dynamic backgrounds. Background subtraction methods perform efficiently in the case of a static background, but fail to adapt to various background transformations. Similarly, optical flow detects motion in surveillance videos by considering the moving object speed, but is not robust to illumination changes. To cope with these challenges, various motion detection techniques have been presented in the literature. Zahurul et al. [30] presented an improved and secure motion detection method for home and office environments. This incorporates both temporal change and optical flow detection methods. The major drawback of this scheme is that the selection of an optimal threshold hinders the implementation of the automatic surveillance system. An SIFT flow-based approach for modeling the background and detecting moving objects was presented by Dou and Li [31]. They modeled each pixel as a group of adaptive SIFT flow descriptors, which are computed over a rectangular region around the pixel, and dynamically updated the background model. Huang and Cheng [32] presented a pyramidal background matching structure for motion detection in surveillance systems. This method utilizes spectral, spatial, and temporal features to generate a pyramid structure of the background model. Based on the background model, the motion of objects is detected by subtracting the