Computer Vision and Image Understanding 139 (2015) 40–58

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Sparsity-driven bandwidth-efficient decentralized tracking in visual sensor networks

Serhan Coşar*, Müjdat Çetin

Faculty of Engineering and Natural Sciences, Sabancı University, Orta Mahalle, Üniversite Caddesi No: 27 34956 Tuzla-İstanbul, Turkey

ARTICLE INFO

Article history:
Received 22 January 2014
Accepted 9 April 2015
Available online 11 June 2015

Keywords:
Camera networks
Visual sensor networks
Human tracking
Sparse representation
Designing overcomplete dictionaries

ABSTRACT

Recent developments in low-cost CMOS cameras have created the opportunity of bringing imaging capabilities to sensor networks and a new field called visual sensor networks (VSNs) has emerged. VSNs consist of image sensors, embedded processors, and wireless transceivers which are powered by batteries. Since energy and bandwidth resources are limited, setting up a tracking system in VSNs is a challenging problem. In this paper, we present a framework for human tracking in VSN environments. The traditional approach of sending compressed images to a central node has certain disadvantages such as decreasing the performance of further processing (i.e., tracking) because of low quality images. Instead, in our decentralized tracking framework, each camera node performs feature extraction and obtains likelihood functions. We propose a sparsity-driven method that can obtain bandwidth-efficient representation of likelihoods extracted by the camera nodes. Our approach involves the design of special overcomplete dictionaries that match the structure of the likelihoods and the transmission of likelihood information in the network through sparse representation in such dictionaries. We have applied our method for indoor and outdoor people tracking scenarios and have shown that it can provide major savings in communication bandwidth without significant degradation in tracking performance. We have compared the tracking results and communication loads with a block-based likelihood compression scheme, a decentralized tracking method and a distributed tracking method. Experimental results show that our sparse representation framework is an effective approach that can be used together with any probabilistic tracker in VSNs.

1. Introduction

Over the past decade, large-scale camera networks have enjoyed increased use in a wide range of applications, especially in security and surveillance. With the developments in wireless sensor networks and the availability of inexpensive image sensors, a new field has emerged: Visual sensor networks (VSNs), i.e., networks of wirelessly interconnected battery-operated devices that acquire video data.

Using a camera in a wireless network poses unique and challenging problems that do not exist either in traditional multi-camera video analysis systems or in sensor networks. In most of the multi-camera video analysis systems, a centralized approach, in which the raw data acquired by cameras are collected in a central unit and analyzed to perform the task of interest, is followed. However, performing complex tasks, such as tracking, recognition, etc., in a communication-constrained VSN environment is extremely challenging. For such constraints, with a data compression perspective, the common approach is to compress images in the process of transmission to the central unit. This strategy essentially focuses on low-level data compression without regard to the final inference goal. Such a strategy may not be appropriate for use under scenarios with severe bandwidth limitations and might cause significant degradation in tracking performance with large compression ratios. We provide a more in-depth review of existing work in Section 2.

In this paper, we propose a different strategy that is better matched to the final inference goal, which, in the context of this paper, is tracking. We propose a sparsity-driven tracking method that is suitable for energy and bandwidth constraints in VSNs. Our method is a decentralized tracking approach in which each camera node in the network performs feature extraction by itself and obtains image features (likelihood functions). In scenarios with overlapping cameras, tracking is performed by fusing the likelihoods obtained from each view. Instead of directly sending likelihood functions to the fusion node, we compute and transmit sparse representations of the likelihoods. By sending such sparse representations to the fusion node, multi-view tracking can be performed without overloading the network. We design special overcomplete dictionaries for the sparse

* Corresponding author.
E-mail addresses: serhancosar@sabanciuniv.edu (S. Coşar), mcetin@sabanciuniv.edu (M. Çetin).

http://dx.doi.org/10.1016/j.cviu.2015.04.010

representation of likelihood functions. The main contribution of this work is building a sparse representation framework and designing overcomplete dictionaries that are matched to the structure of likelihoods. In particular our dictionaries are designed in an adaptive manner by exploiting the specific known geometry of the measurement scenario and by focusing on the problem of human tracking. Each element in the dictionary for each camera corresponds to the likelihood that would result from a single human at a particular location in the scene. Hence actual likelihoods extracted from real observations from scenes containing multiple individuals can be very sparsely represented in our approach. By using these dictionaries, we can represent likelihoods with a very small number of coefficients, and thereby decrease the communication between camera and fusion nodes. To the best of our knowledge, such sparse representation based compression of likelihood functions computed in the context of tracking in a VSN has not been proposed in previous work.

We have used our method within the context of two multi-camera human tracking algorithms [1,2]. We have modified these methods in order to obtain decentralized tracking algorithms. Both by qualitative and quantitative results, we have shown that our method is better than using the block-based compression scheme in [3], the decentralized tracking method in [4], the distributed tracking methods in [5,6] and a traditional centralized approach that compresses raw images acquired by each camera. The sparse likelihood representation framework we present can be used within any probabilistic tracking method under VSN constraints without significantly degrading the tracking performance.

In Section 2, existing pieces of work on tracking in VSNs are reviewed. Section 3 presents our decentralized approach for multi-camera tracking in detail. In Section 4, our sparse representation framework and the details of our specially designed overcomplete dictionaries are described. Experimental results are presented in Section 5. Finally in Section 6, we provide a summary and conclusions.

2. Related work

There exists some previous work on tracking in VSNs. In several pieces of work, basic features or techniques are used to adapt centralized tracking methods to VSNs. For instance, visual hulls are used in [7,8] to detect the presence and number of humans. However, since a visual hull presents the largest volume in which a human can reside, the exact number of humans cannot be determined when humans are positioned close to one another. To minimize the amount of data to be communicated between cameras, in some methods simple features are used for communication. For instance, 2D trajectories are used in [9]. In [10], 3D trajectories together with color histograms are used. Hue histograms along with 2D position are used in [11].

Moreover, there are decentralized approaches in which cameras are grouped into clusters and tracking is performed by local cluster fusion nodes. This kind of approach has been applied to the multi-camera target tracking problem in various ways [4,12,13]. For a nonoverlapping camera setup, tracking is performed by maximizing the similarity between the observed features from each camera and minimizing the long-term variation in appearance using graph matching at the fusion node [12]. For an overlapping camera setup, a cluster-based Kalman filter in a network of wireless cameras has been proposed in [4,13]. In this work, local measurements of the target acquired by members of the cluster are sent to the fusion node. Then, the fusion node estimates the target position via an extended Kalman filter, relating the measurements acquired by the cameras to the actual position of the target by nonlinear transformations.

To further increase scalability and to reduce communication costs, distributed estimation operates without local fusion centers. The estimates generated in a camera are transmitted to its immediate neighbors only. The received estimates are used to refine the estimates at these immediate neighbors, and these refined estimates are then transmitted to the next group of neighbors [5,6,14]. This process ends after a predefined number of steps after all cameras viewing the target are visited or when the uncertainty has decreased below a desired value. In [5], the Kalman–Consensus algorithm [15] is adapted to take into account the directional nature of video sensors and the network topology. Each camera estimates the locations of the people in the scene based on its own sensed data which is then shared locally with the neighboring cameras in an iterative fashion, and a final estimate is arrived at in the network using the Kalman–Consensus algorithm. As an extension of this approach, in [6] authors presented the Information Weighted Consensus filter that weights the estimates coming from neighboring cameras by their information. Thus a camera node which has less information about a person's state is given less weight in the overall estimation process. A wireless embedded smart camera system for cooperative human tracking has been proposed in [14]. At each camera lightweight foreground detection and color histogram based tracking algorithms are implemented and run. Important portions of video and trajectories are determined by detecting events of interest that are pre-defined by users. Communication in the network is minimized by sending messages only when an event of interest occurs.

There are certain limitations of previous work which motivate further research. The methods in [7–11] that use simple features may be capable of decreasing the communication, but they are not capable of maintaining robustness of tracking performance in the case of reduced communication. For the sake of bandwidth efficiency, these methods choose to change the features from complex and robust to simpler, but not so effective ones. Distributed tracking methods [5,6,14] fit well to the needs of VSNs, but suffer from several disadvantages. In the literature of multi-camera tracking, there are many algorithms that can perform robust tracking. In order to use such algorithms in VSN environments, we need to implement existing centralized trackers in a distributed way. In order to do that, usually, one needs to modify pretty much each step from feature extraction to final inference, which is not a straight-forward task and which can affect the performance of the tracker. Performing local processing and collecting features to the fusion node, as in [4,12,13], may not satisfy the bandwidth requirements in a communication-constrained VSN environment. In particular, depending on the size of image features and the number of cameras in the network, even collecting features to the fusion node may become expensive for the network. In such cases, further approximations on features are necessary.

Over the last decade, an alternative sampling/sensing theory, known as "compressed sensing" has emerged. Compressed sensing enables the recovery of signals, images, and other data from what appear to be undersampled observations. Compressed sensing is a technique for acquiring and reconstructing a signal from small amount of measurements utilizing the prior knowledge that the signal has a sparse representation in a proper space. As a consequence, compressed sensing and sparse representation (SR) have become important signal recovery techniques because of their success for acquiring, representing, and compressing high-dimensional signals in various application areas [16–19]. In the past few years, variations and extensions of l_1 minimization have been applied to many vision tasks, including face recognition [20], denoising and inpainting [17], background modeling [21], and image classification [22]. Compressed sensing has also been combined with distributed estimation to perform distributed video coding in VSNs [23]. In almost all of these applications, using sparsity as a prior leads to state-of-the-art results [24].

Following the observations about SR and considering the problems of existing methods, we propose a decentralized approach that fits well to the needs of VSNs and exploits desirable features of a successful centralized tracking algorithm. By transmitting sparse representations of image features, our method can reduce