Image and Vision Computing 34 (2015) 11–26

Contents lists available at ScienceDirect

Image and Vision Computing

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

Visual re-identification across large, distributed camera networks ☆

Vildana Sulić Kenk [a,*], Rok Mandeljc [a], Stanislav Kovačič [a], Matej Kristan [a], Melita Hajdinjak [b], Janez Perš [a]

[a] Machine Vision Laboratory, Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
[b] Laboratory of Applied Mathematics, Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 12 July 2013
Received in revised form 18 September 2014
Accepted 1 November 2014
Available online 28 November 2014

Keywords:
Re-identification
Distributed sensors
Smart cameras
Visual-sensor networks
Surveillance

ABSTRACT

We propose a holistic approach to the problem of re-identification in an environment of distributed smart cameras. We model the re-identification process in a distributed camera network as a distributed multi-class classifier, composed of spatially distributed binary classifiers. We treat the problem of re-identification as an open-world problem, and address novelty detection and forgetting. As there are many tradeoffs in design and operation of such a system, we propose a set of evaluation measures to be used in addition to the recognition performance. The proposed concept is illustrated and evaluated on a new many-camera surveillance dataset and SAIVT-SoftBio dataset.

1. Introduction

The increasing demand for security leads to a growing need for surveillance in many environments [1]. This includes installations of vast closed circuit TV (CCTV) systems; at the time of writing, London Underground has more than 12,000, and a typical casino in Las Vegas has more than 2000 surveillance cameras. Since visual sensors generate large amount of data, scalability becomes important, which gives rise to solutions based on distributed architectures — distributed camera networks. Computer-vision-based methods in camera networks are useful for different tasks, such as object detection and tracking, recognition of problematic or unlawful behavior, and re-identification of objects of interest. In this paper, we focus on the problem of re-identification [2–6], which is the process of finding correspondences between images of an object, acquired at different moments in possibly different camera views.

1.1. Challenges in distributed re-identification

Visual sensor networks (VSNs) may provide relatively large amount of computing and storage resources, but these are typically, both spatially and topologically distant, and the computational capability of an individual node may be low to reduce per-node cost or preserve energy [7]. Consequently, random access to a distant resource may be prohibitively expensive in terms of required network bandwidth, especially in wireless multi-hop networks. While this may be trivially alleviated by replicating all data and processing across all nodes, this defeats the purpose of a distributed architecture and does not solve the polynomially-increasing communication burden. In a truly distributed system, both re-identification and learning are *expensive operations*; the input data appears randomly at multiple nodes, thus requiring constant exchange with all other nodes, which may or may not have relevant information about object's identity.

1.2. Our contribution

We present a holistic approach towards object re-identification in distributed camera networks, which specifically addresses the issues of distributed environments. Specifically, we claim the following contributions:

- Formalization of object re-identification problem in a distributed environment.
- Treatment of re-identification as an open-world problem, with novelty detection and forgetting.
- A set of performance measures that specifically address issues in open-world distributed surveillance.
- Reproducible experiments on a many-camera surveillance dataset ("Dana36", [8]), 8-camera SAIVT-SoftBio dataset [9] and publicly available experimental source code.[1] The code reproduces all results and graphs from this paper. Researchers are encouraged to use it for rapid evaluation of their descriptors or datasets, evaluation of parameter influence and learning and forgetting strategies.

☆ This paper has been recommended for acceptance by Rama Chellappa

* Corresponding author. Tel.: +386 1 4768 876.
E-mail address: vildana.sulic@fe.uni-lj.si (V.S. Kenk).

[1] The full source code can be downloaded from: http://vision.fe.uni-lj.si/research/reid/.

http://dx.doi.org/10.1016/j.imavis.2014.11.002
0262-8856/

The remainder of this paper is organized as follows. After the overview of related work in Section 2, we explain the concept of re-identification in large, distributed camera networks in Section 3. The core of the proposed re-identification mechanism and the experimental methods we used are presented in Section 4, followed by experiments and results in Section 5. Section 6 concludes the paper.

2. Related work

The task of identifying an object based on its previous appearance in some other parts of the camera network is called re-identification. In this respect, we can think about re-identification as form of large-scale tracking [10], which is comprised of several distinct challenges. Therefore, we address these separately.

2.1. Representation

The most frequently studied problem in re-identification is representation of object's appearance. We do not aim to improve the state-of-the-art in this respect, however, since object description is a necessary part of any re-identification system, we present the work done so far for the sake of completeness.

Several approaches model whole body appearance, and have recently been compared by Doretto et al. [10]. Overall appearance is commonly modeled by color or brightness histograms, as for example in [11–13]. Spatial information can be added by representing appearances in joint color spatial spaces [14]. One of the popular approaches is a mixture of color features and texture features [2,15,16]. Other representations include spatio-temporal appearance modeling, such as [17] or spatial and appearance context modeling, such as [18]. Authors in [14] train a multi-class classifier for recognizing people using low-level feature, i.e., color and height histogram. In some approaches, as for example in [19], primitive features such as color, height and body aspect ratio are used in combination with simple threshold-based classification. There is a group of approaches that strives to normalize object appearance across multiple cameras, to improve the performance of appearance descriptors [20,21].

Several approaches use training data to learn a holistic representation based on different low-level features, for example in [22] based on the bag-of-features representations, or in [23] based on Haar-like features and dominant color descriptors. Parts-based approaches are used as well. Part identification and correspondence can be carried out in several ways. One is to use interest point operators such as SURF [24] as in [25] or in [26] and SIFT [27], for example in [28].

Several authors identify body parts by other means. Bak et al. [29] propose an approach for person re-identification using spatial covariance regions [30] of human body parts, which are detected by using Histogram of Oriented Gradients (HOG, [31]). An approach proposed by Farenzena et al. [32] is based on a pondered extraction of local features that encoded different information: chromatic information, structural information through uniformly colored regions, and the nature of recurrent informative (in an entropy sense) patches. Recently, authors in [4] proposed a novel multiple-shot approach, which builds a specific human signature model based on Mean Riemannian Covariance (MRC) patches extracted from tracks of a particular individual. Authors in [33] evaluate different features, trying to find the most suitable ones for person re-identification. They conclude that despite recent advances, person re-identification using local features remains challenging, which might be due to existing descriptors describing mainly shape and texture.

There seems to be a consensus in scientific community that a person re-identification is a difficult problem and despite the best efforts from computer vision researchers, some claim that it remains largely unsolved [34]. Recently, topic models started to appear as a representation of choice in surveillance and re-identification tasks. Such models are usually based on the Latent Dirichlet Allocation (LDA, [35]), see for example [22]. When used for human appearance representation, LDA does not provide topics with obvious, humanly-understandable meaning. Therefore, Liu et al. [16] devised a semi-supervised method for topic generation that yields topics which can be easily interpreted.

2.2. Distributed surveillance systems

Further challenges arise from the need for *distributed representation*, which is especially important to guarantee efficient computation in large-scale networks.

As shown by recent work [26,22,23,28,36,37], the community is increasingly aware of constraints in distributed systems. The multi-stage approach proposed by Jungling et al. [28] provides local extraction of features on camera nodes, thus allowing the lower stages of re-identification to be performed by transmitting extracted features rather than images. Nevertheless, the approach builds its efficiency mainly on compact feature representation that is suitable for transmission and storage in distributed system, and does not provide a specific solution for efficient feature distribution in a large distributed camera system. In the system envisioned by Presti et al. [22], each node individually and autonomously processes the data acquired by its own camera. Communication among nodes enables knowledge sharing and is performed whenever an object leaves a camera's field of view. During the initialization phase, each node detects people and trains a LDA [35] model. These appearance models are propagated across the network and used both to describe incoming objects and to establish correspondences, but it is unclear how the underlying topic model is propagated. Authors claim that the knowledge of the camera network topology is not needed, but they only demonstrate results on data obtained from two cameras — a test case in which efficient feature distribution is obviously not an issue.

The issue of efficient feature propagation in large camera networks has been specifically addressed in our previous work [38]. We have shown that by using hierarchical encoding of features, it is possible to substantially decrease the amount of data transmitted across the network. However, such reduction is limited to *matching*, which is known in surveillance terminology as *matching to the gallery set* [15].

2.3. Novelty detection

An important concept in surveillance and person re-identification is the *novelty detection* [39]. Despite being a classic task in computer vision that had been previously addressed, e.g., [40,41], novelty detection in surveillance received only limited attention, and was to the best of our knowledge used mainly in tasks such as detection of anomalies [42–44], detection of new classes of objects [45] or detection of unusual pedestrian behavior [46].

2.4. Evaluation and datasets

A large amount of work on pedestrian detection, tracking and activity analysis has been done in the framework of the successive PETS workshops. However, to the best of our knowledge, there are only few datasets that are specifically designed for identification and re-identification of pedestrians: the VIPeR dataset [2], the GRID dataset [47], the Person Reidentification dataset [3], 3DPeS [48] dataset, SAIVT-SoftBio [9] dataset, CUHK02 [49] dataset, and our recent Dana36 dataset [8]. The first three provide only small number of images from one or two cameras, while 3DPeS contains video sequences for 200 people in a 8-camera multi-view setting, but provides bounding boxes for only about 1200 frames (a subset named 3DPeS ReId Snap). SAIVT-SoftBio consists of image sequences of 150 people, with average 400 frames per person observed with 8 cameras, but the observed persons pass a particular camera view only once. CUHK02 contains images of 1816 persons, but their identity is observed pairwise regarding the camera views, not on a global scale. The last one, Dana36 dataset, provides 23,683 images from 36 different camera views.