Pattern Recognition Letters 36 (2014) 125–134

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

People counting by learning their appearance in a multi-view camera environment

Lucia Maddalena [a,*], Alfredo Petrosino [b], Francesco Russo [b]

[a] National Research Council, Institute for High-Performance Computing and Networking, Via P. Castellino 111, 80131 Naples, Italy
[b] University of Naples Parthenope, Department of Applied Science, Centro Direzionale, Isola C4, 80143 Naples, Italy

ARTICLE INFO

Article history:
Received 24 January 2013
Available online 18 October 2013

Communicated by A. Fernandez-Caballero

Keywords:
Artificial neural network
Background subtraction
Multi-view
Self organization
Video surveillance

ABSTRACT

We present a people counting system that, based on the information gathered by multiple cameras, is able to tackle occlusions and lack of visibility that are typical in crowded and cluttered scenes. In our method, evidence of the foreground likelihood in each available view is obtained through a bio-inspired mechanism of self-organizing background subtraction, that is robust against well known foreground detection challenges and is able to detect both moving and stationary foreground objects. This information is gathered into a synergistic framework, that exploits the homography associated to each scene view and the scene ground plane, thus allowing to reconstruct people feet positions in a single "feet map" image. Finally, people counting is obtained by a *k*-NN classification, based on learning the count estimates from the feet maps, supported by a tracking mechanism that keeps track of people movements and of their identities along time, also enabling tolerance to occasional misdetections. Experimental results with detailed qualitative and quantitative analysis and comparisons with state-of-the-art methods are provided on publicly available benchmark datasets with different crowd densities and environmental conditions.

1. Introduction

Localization and counting of people in image sequences is a video surveillance task with useful applications. Indeed, people counting can be used for several aims, such as to survey passenger load in urban transportation (buses, ferries, railways, airports, etc.) in order to facilitate the service planning, or to obtain detailed collection of counting data of visitors and customers in public structures (museums, libraries, etc.) and in commercial areas (trade centers, supermarkets, etc.) in order to optimize the resource management.

Certainly, people counting in image sequences is a complex process. Indeed, objects in the scene interact, giving place to overlaps that lead to the temporary loss of some of them, the so-called "occlusions". If a person is visually isolated in the images, localization of its position and visual tracking are quite easy to obtain, because the information usually exploited to identify him (e.g., color distribution, shape, etc.) mainly remains unchanged when the person moves. If the density of objects in the scene increases, also occlusions intensify; consequently, a region of contiguous pixels of the image foreground (blob) could not belong to a single person, but to several persons. In such conditions of limited visibility and crowded scenes it is extremely difficult to correctly detect and track all the persons only based on the images coming from a single camera (single-view). Using several views of the same scene (multi-view) can allow to recover the information that could have been hidden in a specific view.

Several people counting approaches have been proposed in the past twenty years. They have been classified into *detection-based* methods, that determine the number of people, as well as their locations, by identifying individuals in the scene, and *map-based* methods, that exploit the relationship between the number of people and some features extracted from the images (Hou and Pang, 2011). More recently, they have been subdivided into *individual-centric* methods, based on the detection, tracking, and counting the number of tracks, and *crowd-centric* methods, based on the analysis of global low-level features extracted from crowd imagery to produce accurate counts (Chan and Vasconcelos, 2012).

Most of the literature concerning people counting relies on a single-view approach, due to the wide availability of single surveillance cameras and to the relative ease of implementation, since they do not require calibrated cameras nor specific knowledge of the scene geometry. Examples include Davies et al. (1995), Wren et al. (1996), Zhao and Nevatia (2003), Rabaud and Belongie (2006), Kilambi et al. (2008), Albiol et al. (2009), Chan et al. (2009), Sharma et al. (2009), Choudri et al. (2009), Conte et al. (2010), Patzold et al. (2010), Zeng and Ma (2010) and Subburaman

* Corresponding author. Tel.: +39 081 6139522; fax: +39 081 6139531.
E-mail address: lucia.maddalena@cnr.it (L. Maddalena).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.006

et al. (2012). Also neural networks can be exploited for people counting and crowd density estimation (Maddalena and Petrosino, 2012a), as, for instance, in Marana et al. (1998), Cho et al. (1999), Kong et al. (2006) and Hou and Pang (2011). Generally, single-view approaches present difficulties in the analysis of crowded scenes, due to highly possible severe occlusions, and some of them are not robust to illumination changes or have heavy computational load.

Several research directions have been taken in order to handle occlusions. For example, the adoption of cameras looking straight down from the ceiling greatly helps reducing the occlusions (Albiol et al., 2001; Kim et al., 2002; Velipasalar et al., 2006; Englebienne and Krose, 2010). However, the application is limited to indoor environments; moreover, either the acquired sequences still present occlusions in all but the central portion of the image, or the cameras have limited field of view (Harville, 2004). Also stereo cameras have been considered, in order to exploit depth information to project moving people to the ground plane, producing an occupancy map and reducing occlusions (Beymer, 2000; Harville, 2004; Qiuyu et al., 2010; Yahiaoui et al., 2010; van Oosterhout et al., 2011). The use of multiple cameras reveals as fundamental for localizing and counting people in crowded environments. Multi-view approaches aim at reducing hidden image regions due to occlusions, allowing at the same time to reconstruct the target 3D position based on the abundant information provided by different observation points. Examples include Kim and Davis (2006), Alahi et al. (2009), Krahnstoever et al. (2009), Stalder et al. (2009), Ge and Collins (2010) and Ma et al. (2012). However, multi-view approaches usually require calibrated and synchronized cameras, and have a complex structure, resulting in computationally demanding algorithms.

In this work we propose an individual-centric system for robustly counting the number of people under occlusion through multiple cameras with overlapping fields of view, characterized to be neural-brain-like inspired. Compared to the other occlusion handling methods, our multi-view approach relies on the learning of motion templates in time, can adapt to detect both moving and stationary people, and turns out to be robust to gradual lighting variations, moving backgrounds, and cast shadows. The proposed approach is based on the idea of performing an accurate moving object detection in each available view and suitably fusing such information in order to limit problems related to occlusions. To this end, the neural approach to moving object detection recently proposed in Maddalena and Petrosino (2013b) is adopted, where the background model is built by learning in a self-organizing manner image sequence variations, seen as trajectories of pixels in time. The neural model is here adapted to compute "foreground likelihood maps" from different views to be effectively merged together in the multi-view setting. Information fusion is based on the Homographic Occupancy Constraint (Khan and Shah, 2006), that exploits the homography associated to each scene view and the scene ground plane, in order to combine the visual information available by different view-points. This allows to reconstruct people feet positions on the scene ground plane in a single "feet map" image, through the homographies of the foreground likelihood maps. Subsequent tracking in the feet map images is adopted to support people counting, by keeping track of people movements and of their identities along time. Finally, people counting is obtained by supervised classification, based on learning the counts from the feet maps.

The paper is organized as follows. Section 2 describes the moving object detection approach, that allows to obtain a "foreground likelihood" information for each view, based on neural modeling on motion templates. People localization, achieved by reconstructing people feet positions on the scene ground plane, is described in Section 3, while tracking in the feet maps is described in Section 4, and people counting is described in Section 5. Experimental results and comparisons on different real datasets are reported in Section 6, while concluding remarks are provided in Section 7.

2. Neural modelling on motion templates

Foreground detection in each single scene view is the basic building block of our proposed people counting system and its accuracy is crucial for the entire process. Therefore, we adopt here the self-organizing background model for image sequences presented in Maddalena and Petrosino (2013b), whose high accuracy and robustness to well known moving object detection challenges has already been proven. Indeed, extensive experimental results on daytime, night-time, and thermal sequences made available in benchmark datasets (Maddalena and Petrosino, 2012b, 2013a) have shown the high accuracy achieved in handling gradual lighting variations, moving backgrounds, cast shadows, bootstrapping, moving and stationary objects, regardless of acquisition noise. The neural model relies on extensive studies concerning the self-organized learning behaviour of the brain, including Hebb's learning law (Hebb, 1949); Marr's theory of the cerebellar cortex (Marr, 1969); Willshaw, Buneman, and Longnet-Higgins's non-holographic associative memory (Willshaw et al., 1969); Gaze's studies on nerve connections (Gaze, 1970); von der Malsburg and Willshaw's self-organizing model of retina-cortex mapping (Willshaw and Von Der Malsburg, 1976); Amari's mathematical analysis of self-organization in the cortex (Amari, 1980); Kohonen's self-organizing map (Kohonen, 1982); and Cottrell and Fort's self-organizing model of retinotopy (Cottrell and Fort, 1986) (see Maddalena and Petrosino (2012a) for a survey). Here, we provide a concise description of the model and describe how it is adapted for the construction of the foreground likelihood maps. The interested reader is referred to Maddalena and Petrosino (2013b) for an extended description of the adopted approach, a detailed analysis of parameter values, extensive experimental results and comparisons with several state-of-the-art methods.

Given an image sequence $\{I_t\}$, for each pixel $\mathbf{x}$ in the image domain D, we build a neural map consisting of n weight vectors $m_t^i(\mathbf{x}), i = 1, \ldots, n$, which will be called a *model* for pixel $\mathbf{x}$. If every sequence frame has P rows and Q columns, the complete set of models $M_t(\mathbf{x}) = (m_t^1(\mathbf{x}), \ldots, m_t^n(\mathbf{x}))$ for all pixels $\mathbf{x}$ of the tth sequence frame I_t is organized as a 3D neural map $\mathcal{M}_t$ with P rows, Q columns, and n layers, where each layer L_t^i contains, for each pixel $\mathbf{x}$, the ith weight vector $m_t^i(\mathbf{x})$.

As in Maddalena and Petrosino (2013a), an initial background model E_0 is estimated on a subset of K initial sequence frames (in our tests $K = 30$ has been experimentally chosen) through temporal median (Gloyer et al., 1995), and all weight vectors of the neural map $\mathcal{M}_0$ related to a pixel $\mathbf{x}$ are initialized with the pixel brightness value $E_0(\mathbf{x})$. Subsequent learning of the neural map allows the background model to adapt to scene modifications, without introducing the contribution of pixels that do not belong to the background scene. The learning process consists of selectively updating the model by changing the neural weights, according to a visual attention mechanism of reinforcement. Specifically, temporarily subsequent samples are fed to the network. At time t, the value $I_t(\mathbf{x})$ of each incoming pixel $\mathbf{x}$ of the tth sequence frame I_t is compared to the current pixel model $M_t(\mathbf{x}) = (m_t^1(\mathbf{x}), \ldots, m_t^n(\mathbf{x}))$, to determine the weight vector $m_t^b(\mathbf{x})$ that best matches it:

$$d(m_t^b(\mathbf{x}), I_t(\mathbf{x})) = \min_{i=1,\ldots,n} d(m_t^i(\mathbf{x}), I_t(\mathbf{x})), \quad (1)$$

where the metric $d(\cdot,\cdot)$ is suitably chosen according to the specific color space being considered, e.g., the Euclidean distance of vectors in the HSV color hexcone, as suggested in Fisher (1999).