Information Sciences 345 (2016) 199–216

Contents lists available at ScienceDirect

Information Sciences

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

Combining keypoint-based and segment-based features for counting people in crowded scenes

Mahdi Hashemzadeh*, Nacer Farajzadeh

Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran

ARTICLE INFO

Article history:
Received 16 July 2015
Revised 8 January 2016
Accepted 11 January 2016
Available online 2 February 2016

Keywords:
People counting
Crowd counting
Interest points
Keypoints
Occlusion
Video surveillance

ABSTRACT

The counting of the number of people within a scene is a practical machine vision task, and it has been considered as an important application for security purposes. Most of the people counting algorithms generally extract the foreground segments and map the number of people to some features such as foreground area, texture, or edge count. Keypoint-based approaches, on the other hand, have also been proposed, which involves the use of statistical features of keypoints, such as the number of moving keypoints to estimate the crowd size. In contrast to the foreground segment-based methods, keypoint-based approaches are not sensitive to background changes, illuminations, occlusions, and shadows. However, they have limited performance due to the lack of sufficient features. In this paper, in order to estimate the crowd count, the combination of keypoint-based and segment-based (foreground) features is proposed. However, the whole approach is based on the keypoints and not all the image pixels. The proposed method, firstly, extracts the salient keypoints in the scene. Then, foreground segments are obtained by a simple morphological operation on the moving keypoints and hence the system does not suffer from difficulties associated with foreground/background segmentation. Various features are extracted from each foreground segment together with the corresponding keypoints which are highly correlated with the size, density, and occlusion level of the crowd. Finally, a combination of the segment-based and keypoint-based features is used to estimate the number of people in crowds. The experiment demonstrates that the proposed method achieves lower counting error rates compared to the existing approaches.

1. Introduction

The reliable estimation of the number of people in crowded scenes is an important task for both operational and security purposes. It has been widely applied in fields such as the monitoring of public security, gathering of business intelligence, scheduling of public transportation, resource management, and optimization. The estimation of the number of people is also applied in other surveillance applications, such as detection of congestion, delay, unsafe situation, or other behavioral analysis of crowds [11,13]. Hence, many methods based on computer vision techniques have been proposed to perform this task automatically.

Recent approaches on this problem are generally categorized into two groups: (1) detection-based methods and (2) map-based methods (also referred as feature-regression-based methods). The first group tries to detect every individual in the

* Corresponding author. Tel.: +98 9144039592.
E-mail addresses: hashemzadeh@azaruniv.edu, meh_hashemzadeh@yahoo.com (M. Hashemzadeh), n.farajzadeh@azaruniv.edu (N. Farajzadeh).

http://dx.doi.org/10.1016/j.ins.2016.01.060
0020-0255/

scene, using some prior knowledge of the human body shape and appearance [8,15,26,30,32,36,42,46,49,54]. Some of the works in this group also discover independent motions in the scene by clustering motion trajectories of the interest-points which have been tracked over time [4,9,17,39]. Upon detection of individuals, the crowd size is calculated. Conversely, the map-based methods estimate the number of people without having to identify or segment each single person in the scene [5–7,10,22–24,27,33,34,37,40,44,45,53]. These methods, typically work by extracting some holistic features (that is, features from the entire scene) from the foreground pixels, such as foreground area, texture, edge count, etc. This process is followed by a classification method which estimates the crowd count.

With occlusions present in the crowded scenes, the first group of the methods, which are based on a model or appearance of human, cannot perform properly due to partial occlusions. To overcome this challenge, some recent approaches tried to utilize more efficient feature sets [14,20,50,51] or robust classifiers [19,25,52]. However, the hypothesis that a distinct visual separation between people can segment individuals in dense crowds will fail. For such circumstances, map-based methods are considered more robust than the detection-based methods. However, they still suffer from some issues as follows:

- Difficulties associated with the background/foreground segmentation process: this task requires a well-estimated background model, accurate binarization, and an effective shadow removal algorithm in order to provide a reliable foreground image for extracting the target features. These are difficult to achieve due to the changes in illumination, occlusion, and movement of camera, especially in outdoor environments.
- Erroneous image features: textural and edge based features which are usually utilized in these methods can be extremely inaccurate. Textural features are very sensitive to the scene background and illumination changes. Also, a complex background and rough textures of human clothes results in completely messy edges. Moreover, these features are computationally expensive to extract.
- Difficulties of data gathering: the extracting of holistic features from scenes can give rise to extensively different features. Therefore, a great quantity of training data must be provided to cover a wide variation in crowds. This task, however, is not practical in real world applications where numerous cameras are usually installed in different situations.

A recent method following the map-based approach was proposed by Albiol et al. [1], which utilized some keypoints (corner-points) as features to estimate the crowd size within a scene. According to this method, firstly, some corner-points were detected using Harris corner detector [16]. Then, moving points were separated from static ones (background points) by estimation of motion vectors of corner-points between adjacent frames. Finally, the number of people was estimated from the number of moving points based on a direct proportionality relation. Although the algorithm in this method may seem rather simple, it achieved the best performance among all the participants of PETS 2009 [38] contest on people counting task [12]. Beside the good performance, other advantages of this approach include: (1) there is no need to estimate the background, (2) no need to segment foreground areas or individuals, (3) no need to cope with shadow problems, and (4) no need to extract complicated image features. However, the attained accuracy of Albiol's method [1] is limited due to the lack of sufficient features, effects of perspective, instability of the Harris corners, and the presence of occlusions.

In this paper, an approach which utilizes both the keypoint-based and segment-based features is proposed to accurately estimate the number of people in crowded scenes. However, it is important to note that the proposed approach is only on the basis of some keypoints and does not take the entire image pixels into account. In this work, the foreground segments are also obtained by using the pixels of moving keypoints and hence the system is not sensitive to issues associated with the background/foreground segmentation. Since the importance of the features which can be computed from the foreground segments is not dispensable, the motivation is to estimate the crowd size by using a combination of the keypoint-based and segment-based features.

At the first step, the proposed method extracts the salient keypoints in the scene. Then, foreground segments are obtained by a simple morphological operation on the moving keypoints. Afterwards, several local features (local with respect to the groups of people moving together) are extracted from each foreground blob and the corresponding keypoints. Finally, a combination of the segment-based features and keypoint-based features is used to estimate the number of people in the groups. The groups are independently counted, so that the total number of people in the crowd is the sum of its parts. It is worth mentioning that as the scene is analyzed based on the local information, more training cases will be available in one training image [44]. Therefore, it is possible to capture various crowd distributions from a small amount of training data.

The performance of the proposed algorithm is evaluated on large benchmark pedestrian datasets, featuring a wide variety of environmental conditions and crowd configurations. The proposed approach is compared with Albiol's method [1] and the other state of the art approaches. The results confirm that the proposed method outperforms the competitors. The results of our extensive experiments demonstrate that the extracted features are very informative and highly correlated with the size, density, and complexity of the crowds. Moreover, it is shown that the proposed method is highly generalizable and can be applied to count the crowds not encountered during the training.

In brief, the contributions of this paper are as follows:

- A novel crowd counting system is proposed which utilizes both keypoint-based features and segment-based features together.
- The foreground segmentation scheme designed for this system works without having to estimate the background image and hence is not sensitive to the difficulties associated with common foreground/background segmentation techniques.