Available online at www.sciencedirect.com

ScienceDirect

Computer Vision and Image Understanding 110 (2008) 43–59

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Estimating pedestrian counts in groups

Prahlad Kilambi, Evan Ribnick, Ajay J. Joshi, Osama Masoud, Nikolaos Papanikolopoulos *

Department of Computer Science and Engineering University of Minnesota, Twin Cities, USA

Received 7 August 2006; accepted 27 February 2007
Available online 12 March 2007

Abstract

The goal of this work is to provide a system which can aid in monitoring crowded urban environments, which often contain tight groups of people. In this paper, we consider the problem of counting the number of people in the scene and also tracking them reliably. We propose a novel method for detecting and estimating the count of people in groups, dense or otherwise, as well as tracking them. Using prior knowledge obtained from the scene and accurate camera calibration, the system learns the parameters required for estimation. This information can then be used to estimate the count of people in the scene, in real-time. Groups are tracked in the same manner as individuals, using Kalman filtering techniques. Favorable results are shown for groups of various sizes moving in an unconstrained fashion.

Keywords: Groups; Count estimation; Pedestrian tracking; Occlusions; Projection

1. Introduction

Using computer vision techniques to monitor humans in urban environments is a difficult problem. In many human and group activity monitoring applications, it is important to get a sense of scale in the scene and estimate the number of people present. However, most of the cues that aid in detection and tracking of individuals like shape, texture, and appearance break down in the case of crowds, especially when low-resolution surveillance cameras are used. As such, it can be helpful to use an approach that treats a group of people as a single entity instead of processing each person individually. In this paper, we address the important problem of estimating the number of people in a group and being able to track it reliably.

The ability to accurately estimate the number or density of people in a scene has useful applications in many areas. In traffic control, automatic pedestrian and crowd monitoring techniques can be used to increase safety and improve timing of traffic. For example, at traffic intersections, intelligent walk–signal systems could be designed based on the proposed method. The system could be automated, such that the signal decides when to change based on the number of pedestrians waiting to cross. This would be more efficient than most systems currently in use, in which people have to press a button in order to cross intersections.

Other important applications have to do with estimating the number of people walking through a crowded area. For example, knowing the size and density of a group outside a school or public event can help authorities identify unsafe situations and regulate traffic appropriately. This capability can also be useful in planning an environment. For example, a sidewalk or corridor could be designed based on typical pedestrian traffic patterns in the area. Additionally, retailers could use information about traffic in their stores to improve efficiency and design spaces appropriately.

Although there is a lot of work in the literature on tracking people in crowded scenes, there has been limited research on the specific problem of counting the number of people in a scene, and most of the work that has been done is restricted to counting individuals. Systems based

* Corresponding author.
E-mail addresses: prahlad@cs.umn.edu (P. Kilambi), ribnick@cs.umn.edu (E. Ribnick), ajay@cs.umn.edu (A.J. Joshi), masoud@cs.umn.edu (O. Masoud), npapas@cs.umn.edu (N. Papanikolopoulos).

1077-3142/$ - see front matter
doi:10.1016/j.cviu.2007.02.003

on detection and tracking of individuals suffer in areas with high crowd density and occlusion. As such, these systems are of limited use in urban environments, which often contain small or large groups of people which are dense, and in which many people are severely occluded.

We propose a method than can be used to accurately estimate the number of people in a scene without constraining ourselves to detection of individuals. The approach here is group-based, where the count of people in a dense moving group is estimated as a whole. In this paper, both, a heuristic-based and a shape-based method are presented for estimating group populations. Both methods are implemented in real-time. We then track each group as a single entity using a tracker based on an extended Kalman filter.

The methods presented here are most effective for scenes in which people travel as individuals or in social groups. Examples of places where this might be applicable include shopping malls, college campuses, and public transportation areas such as train stations and airports. This system is not intended for use in environments with immense crowds of people that cannot be segmented into groups, such as mob scenes, political rallies, or the like.

The paper is organized as follows. In Section 2, we discuss some of the previous work in the field of crowd monitoring and counting. We provide a brief overview of the entire system in Section 3. Section 4 explains the projective geometry techniques used in the group population estimators. In Section 5, we introduce the tracking mechanism and its extensions. We present the two counting methods in Sections 6 and 7. An alternative bounds-based formulation is presented in Section 8. Section 9 describes how the system handles groups merging and splitting. We then present and analyze the results obtained from experiments in Section 10, and some concluding remarks and future direction are given in Section 11.

2. Related work

There have been various approaches to the problem of counting people in crowded environments. Early work involved locating people by looking for heads in the vertical histograms of the blobs, where the number of peaks was assumed to correspond to the number of heads. Davies et al. [1], in 1995, proposed a system which provides an estimate based on the number of foreground pixels and edge pixels. It uses Fourier transform techniques to identify motion of the crowd. It is restricted to motion only in the vertical direction in the image. Crowd density estimation is done using elementary techniques such as estimating the area of the image segments which correspond to moving crowds, or using the perimeter of the region occupied by the crowd. Although the system works well for scenes with few people, it is incapable of accurately determining density in case of occlusions and illumination changes as mentioned by the authors. Since occlusions occur frequently in cases with large groups, the system also fails when there is a large number of people.

Pfinder [2], which used a statistical model for color and shape to segment a person, tracked heads and hands, and identified gestures, mainly dealt with individuals. W4 [3] was another system for detecting and tracking individuals based on shape models. However, detection of individuals becomes less feasible in the case of crowds, where people are typically severely occluded. Hydra [4], an extension of W4, proposed a method based on silhouettes to discern people moving as groups. It used heads to count people in groups. However, heads may not be a very reliable cue when they occupy only a few pixels in the image, and hence require a reasonably close view of the scene.

Texture was used as a cue by Marana et al. [5] in order to estimate crowd density using a Kohonen neural network and Haralick's gray level dependence matrix. One drawback is that the method requires the background to consist of low frequency variations only. Since texture is estimated using frequency variations in the image, high frequency changes in the background adversely affect this method. This limits applicability to controlled, indoor scenes only. Also, since the method uses two neural networks, extensive training is required for good performance. The authors mention training using 151 images in the paper. They were able to provide an estimate of the density but were not able to estimate the exact count of people.

Zhao and Nevatia [6] proposed a Bayesian model-based segmentation algorithm using shape models which segmented each individual from a scene. It was able to count individuals in groups of people. This method was based on Markov chain Monte Carlo sampling and was prohibitively slow for large crowds. They later proposed a method for tracking people [7], which used 3D ellipsoid models to track individuals, which, the authors claim, performs better than the original method.

The major drawback of most of the methods mentioned above is that they assume there is a distinct visual separation between individuals, so that the motion-segmented image contains enough visual information to separate individuals moving as a group. However, this is not always true in dense groups, when people are severely occluded and visually inseparable. See, for example, Fig. 1.

In [8], Marana et al. describe a technique for estimating the density of crowds using the Minkowski fractal dimension as a characterization of image texture. Kettnaker and Zabih [9] outline a people counting method using multiple cameras. They use visual appearance matching and mutual content constraints together in order to count people. Using multiple cameras has the obvious advantage of alleviating problems due to occlusions. However, registration needs a significant amount of work and prior calibration.

In [10], an MRF-based approach is used first to do foreground extraction, and then a density estimate of the crowd is obtained using calibration information on the extracted foreground. This method suffers when there is occlusion as well, and the authors propose using the time-median statistic for better average results.