Image and Vision Computing 27 (2009) 1445–1458

Contents lists available at ScienceDirect

Image and Vision Computing

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

Robust pedestrian detection and tracking in crowded scenes

Philip Kelly *, Noel E. O'Connor, Alan F. Smeaton

Centre for Digital Video Processing, Electronic Engineering, Adaptive Information Cluster, Dublin City University, Dublin, Ireland

ARTICLE INFO

Article history:
Received 1 March 2007
Received in revised form 30 November 2007
Accepted 10 April 2008

PACS:
42.30.Tz

Keywords:
Pedestrian detection
Pedestrian tracking
Stereo
Crowds

ABSTRACT

In this paper, a robust computer vision approach to detecting and tracking pedestrians in unconstrained crowded scenes is presented. Pedestrian detection is performed via a 3D clustering process within a region-growing framework. The clustering process avoids using hard thresholds by using bio-metrically inspired constraints and a number of plan-view statistics. Pedestrian tracking is achieved by formulating the track matching process as a weighted bipartite graph and using a *Weighted Maximum Cardinality Matching* scheme. The approach is evaluated using both indoor and outdoor sequences, captured using a variety of different camera placements and orientations, that feature significant challenges in terms of the number of pedestrians present, their interactions and scene lighting conditions. The evaluation is performed against a manually generated groundtruth for all sequences. Results point to the extremely accurate performance of the proposed approach in all cases.

1. Introduction

The vision of Ambient Intelligence (AmI) [1] depicts environments that are able to adapt intelligently to facilitate the requirements of the people present. AmI leverages a networked system of smart devices and sensors, which have been smoothly integrated into the environment to act as a global interface between users and information systems [2]. In this way, the control of the augmented environment becomes action oriented, responding appropriately to the behaviour of the human users present. This promises many benefits for both single individuals and larger groups of people in a variety of application scenarios.

In order for AmI to become a reality, a number of key technologies are required from a variety of disciplines [1]. These include unobtrusive sensor hardware, wireless and fixed communication systems, software design, information fusion, intelligent agents, to cite but a few. In this paper, a focus is made on the requirement for robust detection and tracking of humans in unconstrained scenes. This is a key enabling technology since knowing who is where in a scene *and* what their actions have been allows other layers in an AmI framework to infer beliefs about those people. Consider the example of an automated pedestrian traffic light system. An embedded intelligent system should be able to determine the number of people waiting to cross, whether any special assistance should be flagged for any individual pedestrian (e.g. wheelchair, children or elderly pedestrians), estimate the time needed for everyone to cross, determine the state of traffic flow on the road and ensure each person crosses the road successfully before allowing vehicular traffic to flow. Clearly detecting and tracking the pedestrians is a necessary pre-processing step. However, this poses significant challenges when pedestrian detection and tracking in unconstrained real-world crowded environments is considered. For example, just because a person is in the scene does not mean that they want to cross the road, however, if the person walks towards the crossroads, stops and waits, then they probably do. RFID tagging is a possible solution for determining this in constrained environments, but cannot help in scenarios where there is no contact with people in a scene until they enter the environment.

Many of the person detection techniques described so far in the literature – see Section 2 – make assumptions about the environmental conditions, pedestrian and background colour intensity information, the pedestrian flow, that a person will exist in the scene for a given number of frames, or that a person enters the scene un-occluded. In this paper, a robust pedestrian detection and tracking system for a single stereo camera is presented, which attempts to minimise such constraining assumptions. It is able to robustly handle:

(1) occlusion, even when multiple people enter the scene in a crowd;
(2) lack of variability in colour intensity between pedestrians and background;
(3) rapidly changing and unconstrained illumination conditions;
(4) pedestrians appearing for only a small number of frames;

* Corresponding author. Tel.: +353 863391546.
E-mail address: kellyp@eeng.dcu.ie (P. Kelly).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.04.006

(5) relatively unconstrained pedestrian movement;
(6) relatively unconstrained pedestrian pose, appearance and position with respect to the camera;
(7) varying camera heights, rotations and orientations;
(8) static pedestrians.

In addition, as the proposed pedestrian detection algorithm uses a simple biometric person model that is defined with respect to the groundplane, the system requires *no* external training to detect and track pedestrians. However, although the proposed system was designed to minimise constraining assumptions, a small number of inherent assumptions still exist within the system framework. They include;

(1) that pedestrians in the scene are standing upright with respect to the groundplane;
(2) that all moving objects in the scene (within the volume of interest) are caused by foreground pedestrians;
(3) that pedestrians in the scene are moving at a velocity of less than 3 m/s.

In addition to this, the system does have a small number of drawbacks on the type of scenario it can survey. These include; (a) that a relatively flat groundplane is present within the scene, where no object of interest is located below this groundplane; (b) the camera must be orientated so that the groundplane is visible in the image plane; and (c) the system is only able to reliably detect pedestrians for a short-medium range, up to a maximum distance of 8 m from the camera. An area of future work envisioned by the authors includes the investigation of techniques to further reduce these assumptions and limiting constraints.

The main areas of contribution of this paper are twofold. The first lies in the introduction of a novel, non-quantised, plan-view statistic (an overview of such statistics is given in Section 2) which incorporates global features into the pedestrian clustering framework of the authors' previous work [3]. The use of this plan-view statistic within this framework significantly improves robustness to both over- and under-segmentation of pedestrians in comparison to [3]. The second main contribution area lies in the robust pedestrian tracking technique that has been developed. Within this area a number of contributions can be identified, which include; (a) a matching technique that incorporates a novel weighting scheme for matching pedestrians to previous tracks; (b) a series of kinematic constraints that model possible pedestrian movement through the scene and that can be used to remove implausible matches of pedestrians to previous tracks; and (c) rollback loops and post-processing steps to increase track robustness to both over-/under-segmentation.

This paper is organised as follows: Section 2 gives an overview of the related work in the area of pedestrian detection and tracking techniques and outlines the benefits of stereo information within this area. Section 3 gives an overview of the key components to the overall pedestrian detection and tracking system. Sections 3.1 and 3.2 discuss the details of the proposed approach to pedestrian detection and tracking, respectively. In Section 4 experimental results (evaluated against a groundtruth) are provided for indoor and outdoor situations at various orientations containing multiple pedestrians at various depths, some with severe occlusion and displaying a large variability in both local and global appearance. Finally, Section 5 details conclusions and future work.

2. Related work

Robust segmentation and tracking of pedestrians within an unconstrained scene is one of the most challenging problems in computer vision. A few of the complicating factors to segmenting people include; the large variability in a person's local and global appearance and orientation [4]; occlusion of an individual by one or several other persons, or objects, especially if the person is located within a crowd; lack of visual contrast between a person and background regions. In addition, unconstrained real-world outdoor environments tend to create further challenges, such as rapidly changing lighting conditions due to varying cloud cover, shadows, reflections on windows, and moving backgrounds.

A significant amount of research literature exists on person detection and tracking. Various techniques for segmenting individual pedestrians have been investigated using traditional 2D computer vision techniques. Unfortunately, few of these, if any, produce reliable results for long periods of time in unconstrained environments [5]. Reasons for this stem from various assumptions regarding the environmental conditions and type of pedestrian flow being violated. For example, techniques, such as [6–10], depend on accurate segmentation of moving foreground objects from a background colour intensity model as a first step in their algorithmic process. This relies on an inherent assumption that there will be significant difference in colour intensity information between people and the background. Other techniques [11–14] use rhythmic features obtained from a temporal set of frames for pedestrian detection, such as the periodic leg movement of a walking human, or motion patterns unique to human beings, such as gait. However, the assumption that a person will be moving (and normally in a pre-defined direction), means that people standing still, or performing unconstrained and complex movement, or in crowded scenes when legs are occluded, will not be detected. Other techniques, such as [6,7], make an assumption that a person will appear in the scene un-occluded for a given period of time allowing a model of the pedestrian to be built up while they are isolated. In addition, appearance-based techniques often fail when two people get close together, as the algorithm fails to allocate the pixels to the correct model because of similarities in appearance, and tracking is lost. To increase reliability, some systems, e.g. [15], integrate multiple cues such as skin colour, face and shape pattern to detect pedestrians. However, skin colour is very sensitive to illumination changes and face detection can identify only pedestrians facing the camera.

3D stereo information has been proposed as a technique to overcome some of these issues. The use of stereo information carries with it some distinct advantages over conventional 2D techniques [5,16]:

(1) It is a powerful cue for foreground–background segmentation [17];
(2) It is not significantly affected by sudden illumination changes and shadows [18];
(3) The real size of an object derived from the disparity map provides a more accurate classification metric than the image size of the object;
(4) Occlusions of people by each other or by background objects can be detected and handled more explicitly;
(5) It permits new types of features for matching person descriptions in tracking;
(6) It provides a third, disambiguating dimension for matching temporal pedestrian positions in tracking.

However, range information also has its disadvantages; (a) it can be a noisy modality where the standard deviation of the depth value at a pixel over time is commonly of the order of 10% of the mean [5]; (b) it cannot segment foreground objects at the same depth as background regions; and (c) no technique has been developed that returns correct range information in all scenarios, all of the time. However, despite these drawbacks, the authors consider