Available online at www.sciencedirect.com
ScienceDirect
J. Vis. Commun. Image R. 19 (2008) 75–91
JOURNAL OF VISUAL Communication & IMAGE Representation
www.elsevier.com/locate/jvci

Adaptive multi-modal stereo people tracking without background modelling

Rafael Muñoz-Salinas [a,*], Miguel García-Silvente [b], Rafael Medina Carnicer [a]

[a] Department of Computing and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain
[b] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

Received 21 September 2006; accepted 10 July 2007
Available online 28 July 2007

Abstract

Detecting and tracking persons in the sequences of monocular images are the important and difficult problems in computer vision and have been well studied in these two decades. Recently, the methods based on stereo vision have attracted great attentions since 3D information can be exploited. This paper presents an approach for multiple-people detection and tracking using stereo vision. Tracking is carried out using a multiple particle filtering approach that combines depth, colour and gradient information. We modify the degree of confidence assigned to depth information, according to the amount of it found in the disparity map, using a novel confidence measure. The greater the amount of disparity information found, the higher the degree of confidence assigned to depth information in the final particles weights is. In the worst case (total absence of disparity), the proposed algorithm makes use of the information available (colour and gradient) to track, thus performing as a pure colour-based tracking algorithm. People are detected combining an adaboost classifier with stereo information. In order to test the validity of our proposal, it is evaluated in several sequences of colour and disparity images where people interact in complex situations: walk at different distances, shake hands, cross their paths, jump, run, embrace each other and even swap their positions quickly trying to confuse the system. The experimental results show that the proposal is able to deal with occlusions and to effectively determine both the 3D position of the people being tracked and their 2D head locations in the camera image, and everything is realized in real time. Besides, as the proposed method does not require the use of a background model, it can be considered particularly appropriate for applications that must run on mobile devices.

Keywords: People tracking; Stereo vision; Particle filtering; Real-time imaging; Confidence measure; Colour processing

1. Introduction

In the last decade, there is a growing interest in the development of systems and techniques for people detection and tracking both in indoor and outdoor environments. The potential applications of such technology include: ambient intelligent systems [14,27,48,56], visual servoing applications [2,37], augmented reality and human–computer interaction [12,15,23,31,62], video compression [39,59] or robotics [8,31,43,44].

People tracking based on monocular images is a well explored topic, approached in the most of the cases by the integration of multiple visual cues. However, nowadays, the use of stereo vision for these purposes is an active research area. The availability of commercial hardware to solve the low-level problems of stereo processing, as well as the lower prices for these devices, turns them into an appealing sensor to be used in intelligent systems. Stereo vision brings several advantages over monocular systems. First, all the methods designed for tracking in monocular images can be applied, but with much richer per-pixel information (colour or luminance plus depth). Depth information can be employed to achieve a better tracking of people environments and a better understanding of their gestures. Second, disparity information makes the systems

* Corresponding author.
E-mail addresses: rmsalinas@uco.es (R. Muñoz-Salinas), M.Garcia-Silvente@decsai.ugr.es (M. García-Silvente), rmedina@uco.es (R. Medina Carnicer).

1047-3203/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2007.07.004

more robust against illumination changes. Therefore, systems that employ stereo vision are expected to be more robust in real scenarios where illumination changes might occur.

This paper presents a solution to the people detection and tracking problems by the integration of multiple visual cues using a particle filtering approach [30]. People are detected combining an adaboost classifier [47] with stereo information. Each person detected is tracked using a body model comprised by two ellipsoidal surfaces (one for tracking their head and another one for their torso). The proposed observation model combines a colour technique [13], with depth information, and with a gradient based technique [6]. However, depth information cannot always be extracted because the surfaces could not be lambertians, of depth discontinuities or due to the lack of texture. To overcome this problem, the relevance assigned to depth information depends on the amount of it encountered in the disparity map. When the amount of disparity obtained is high, depth information has an important degree of contribution to the final particle weight. However, as the amount of disparity decreases, so does the relevance assigned to depth information. In the worst case (total absence of disparity), the algorithm takes advantage only of colour and gradient information.

The multiple target tracking problem (MTT) has been approached using a multiple particle filter [16,33,46,57]. Independent trackers are employed for each person and an interaction factor is employed to avoid the coalescence problem. Our proposal estimates not only the 3D position of the persons being tracked but also the 2D location of their head in the camera image. Determining the head position of the person being tracked is an important piece of information that may be used for human–computer and human–robot interaction tasks (e.g., face pose estimation, expression analysis). As the method proposed does not utilize background estimation techniques, it can be considered particularly appropriate for applications that must run on mobile devices (e.g., mobile robots).

The remainder of this paper is structured as follows: Section 2 provides an overview about stereo people detection and tracking-related works. Sections 3 and 4 explain the basis of stereo processing and colour modelling, respectively. Section 5 shows how people detection has been carried out in this work. Section 6 explains our people tracking approach while Section 7 shows the experimentation performed. Finally, Section 8 draws some conclusions.

2. Related work

People detection and tracking are interesting topics that have attracted the interest of researches in many areas. These problems have been generally tackled by exploiting the morphological characteristics of the human body [28] and the colour characteristics of human skin [21,51,54]. Considering that static colour models usually lead to a performance degradation [38] (due to illumination changes), some authors have opted to employ dynamic skin colour models [55]. Another option chosen by some authors is to provide extra information to the tracker such as the colour of the clothes worn by the user. All such information is used to overcome problems as illumination changes and the confusion due to the presence of different subjects [35,42,52]. Finally, another ways of dealing with the detection problem include the use of multiple sensors [20], multiple visual cues [8] or specific hardware [61] to perform a more robust detection.

Several authors have employed stereo vision for developing more sophisticated tracking methods taking advantage of stereo vision. When using stereo vision, most of the monocular techniques available in the literature are also directly applicable. Therefore, a key aspect consists in the proper combination of the existing methods with depth information. Among the first works in that area we found the one by Cipolla and Yamamoto [11]. They present a stereo-based method for tracking objects (which are just moving along the horizontal plane) using the *visualised locus* method. Eklundh et al. [17] study the uses of active vision and they argue that the integration of multiple cues is necessary in order to accomplish real-life tasks. They present an active vision system able to decide where to focus its attention based on the integration of motion and stereo. In [15], Darrel et al. present an interactive display system capable of detecting and tracking several people. Person detection is based on the integration of the information provided by three modules: a skin detector, a face detector and the disparity map provided by a stereo camera. In their work, people segmentation was performed in the disparity map. Perseus [31] constitutes a remarkable example of using stereo vision for interacting with human users. The Perseus system is able to detect and track a single user and to determine when and where they point to. The Perseus architecture is also employed by Franklin et al. [19] to develop a robotic waiter controlled by simple hand gestures. As in the work of Darrel et al. [15], people detection and tracking is based on determining the person position in the camera image. Grest and Koch [23] develop a system able to detect and track a single-person using stereo vision for augmented reality applications. Once the user is located, colour histograms of the face and chest regions are created and employed by a particle filter to estimate their position in the next image. Then, stereo information assists to determine the real position of the person in the room. It is employed to calculate the corresponding position in the virtual environment. However, stereo and colour were not integrated in the tracking process. In that work, the stereo processing is performed using the information gathered by cameras placed at different locations in the room. Beymer and Konolige [5] use a Kalman filtering and a template system for people detection and tracking. Their proposal uses an interesting pyramid method to speed up the system. However, their system does not employ colour information so that the identity of the people being tracked might be confused. Moreno et al.