Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Vision and Image Understanding 102 (2006) 1–21

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Simultaneous tracking of multiple body parts of interacting persons

Sangho Park *, J.K. Aggarwal

Computer and Vision Research Center, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA

Received 18 October 2003; accepted 18 July 2005
Available online 14 November 2005

Abstract

This paper presents a framework to simultaneously segment and track multiple body parts of interacting humans in the presence of mutual occlusion and shadow. The framework uses multiple free-form blobs and a coarse model of the human body. The color image sequence is processed at three levels: pixel level, blob level, and object level. A Gaussian mixture model is used at the pixel level to train and classify individual pixel based on color. Relaxation labeling in an attribute relational graph (ARG) is used at the blob level to merge the pixels into coherent blobs and to represent inter-blob relations. A twofold tracking scheme is used that consists of blob-to-blob matching in consecutive frames and blob-to-body-part association within a frame. The tracking scheme resembles multi-target, multi-association tracking (MMT). A coarse model of the human body is applied at the object level as empirical domain knowledge to resolve ambiguity due to occlusion and to recover from intermittent tracking failures. The result is 'ARG–MMT': 'attribute relational graph based multi-target, multi-association tracker.' The tracking results are demonstrated for various sequences including 'punching,' 'hand-shaking,' 'pushing,' and 'hugging' interactions between two people. This ARG–MMT system may be used as a segmentation and tracking unit for a recognition system for human interactions.

Keywords: Tracking; Body part; Human interaction; Occlusion; ARG; MMT

1. Introduction

Video surveillance of human activity requires reliable tracking of moving human bodies. Tracking non-rigid objects such as moving humans presents several difficulties for computer analysis. Problems include segmentation of the human body into meaningful body parts, handling the occlusion of body parts, and tracking the body parts along a sequence of images. Many approaches have been proposed for tracking a human body (see [1–3] for reviews). The approaches for tracking a human body may be classified into two broad groups: model-based approaches and appearance-based approaches.

Model-based approaches use a priori models explicitly defined in terms of kinematics and dynamics. The body model is fitted to an actual shape in an input image. Various fitting algorithms are used with motion constraints of the body model. Examples include 2D models such as the stick-figure model [4] and cardboard model [5], and 3D models such as the cylinder model [6] and super-ellipsoid model [7]. 3D models can be acquired with either multiple cameras or a single camera [8,9]. Difficulties with model-based approaches lie in model initialization, efficient fitting to image data, occlusion, and singularity involved in inverse kinematics.

Appearance-based approaches use heuristic assumptions on image properties when no a priori model is available. Image properties include pixel-based properties such as color, intensity, and motion, or area-based properties such as texture, gradient, edge, and neighborhood areas. Appearance-based approaches aim at maintaining and tracking those image properties along the image sequence. Examples include edge-based methods such as energy minimization [10], sampling-based methods such as Markov chain Monte Carlo estimation [11], area-based methods [12,13], and template-based methods [14]. Some

* Corresponding author. Fax: +1 512 471 5532.
E-mail addresses: sanghopark@alumni.utexas.net (S. Park), aggarwaljk@mail.utexas.edu (J.K. Aggarwal).

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2005.07.011

approaches may combine model-based methods with appearance information [15,16].

Most of the methods that use a single camera assume explicitly or implicitly that there is no significant occlusion between tracked objects. To date, research has focused on tracking a single person in isolation [17,13], or on tracking only a subset of the body parts such as head, torso, hands, etc. [18]. Research on segmentation or tracking of multiple people has focused on the analysis of the whole body in terms of the silhouettes [19,14], contours [20,21], color [22], or blob [13,23].

The objective of this paper is to present a method for segmentation and tracking of multiple body parts in a bottom-up fashion. The method is a bottom-up approach in the sense that individual pixels are grouped into homogeneous blobs and then into body parts. The tracks of the homogeneous blobs are automatically generated and multiple tracks are maintained across the video sequence. Domain-knowledge about the human body is introduced at the high-level processing stage.

We propose an appearance-based method for combining the attribute relational graph and data association among multiple free-form blobs in color video sequences. The proposed method can be effectively used to segment and track multiple body parts of interacting humans in the presence of mutual occlusion and shadow. In this paper, we address the problem of segmenting multiple humans into semantically meaningful body parts and tracking them under the conditions of occlusion and shadow in indoor environments. This is a difficult task for several reasons. First, the human body is a non-rigid articulated object that has many degrees of freedom (DOF) in its articulation. Precise modeling of the human body would require expensive computation. Model-based approaches often require manual initialization of the body model. Second, loose clothing introduces irregular shape deformation. Silhouette- or contour-based approaches are sensitive to noise in shape deformation. Third, occlusion and shadow are inevitable in situations that involve multiple humans. Self-occlusion occurs between different body parts of a person, while mutual occlusion occurs between different persons in the scene. Image data is severely hampered by occlusion and shadows, making it difficult to segment and track body parts. Multiple-view approaches are often introduced to overcome the occlusion and shadow effects. But multiple-view approaches are not applicable in widely available single-camera video data. High-level domain knowledge may also be used to infer the body-part relations under occlusion.

The proposed system processes the input image sequence at three levels: pixel level, blob level, and semantic object level. A Gaussian mixture model is used to classify individual pixels into several color classes. Relaxation labeling with attribute relational graph (ARG) is used to merge the color-classified pixels into coherent blobs of arbitrary shape according to similarity features of the pixels. The multiple blobs are then tracked by data association using a variant of the multi-target, multi-association tracking (MMT) algorithm used by Bar-Shalom et al. [24]. Unmatched residual blobs are tracked by inference at the object level using a body model as domain knowledge. A coarse body model is applied as empirical domain knowledge at the object level to assign the blobs to appropriate body parts. The blobs are then grouped to form the meaningful body parts by the simple body model. Using the simple human-body model as a priori knowledge helps to resolve ambiguity due to occlusion and to recover from intermittent tracking failure. The result is 'ARG–MMT': 'attribute relational graph based multi-target, multi-association tracker.'

Fig. 1 shows the overall system diagram of the ARG–MMT. At each frame, a new input image is compared with a Gaussian background model. The background subtraction module produces the foreground image. Pixel-color clustering produces initial blobs according to pixel color. Relaxation labeling merges the initial blobs on a frame-by-frame basis. Multiblob tracking associates the merged blobs in the current frame with the track history of the previous frame and update the history for the current frame. Body-part assignment assigns the tracked blobs to the appropriate human body parts. The body-pose history of the previous frame is incorporated as domain knowledge about the human body. The assigned body parts are recursively updated for the current frame.

The rest of the paper is organized as follows. Section 2 describes the procedure at the pixel level, Section 3 describes the blob formation, Section4 presents a method to track multiple blobs, while Section 5 describes the segmentation and tracking of semantic human body parts. Experiments and conclusions follow in Sections 6 and 7, respectively.

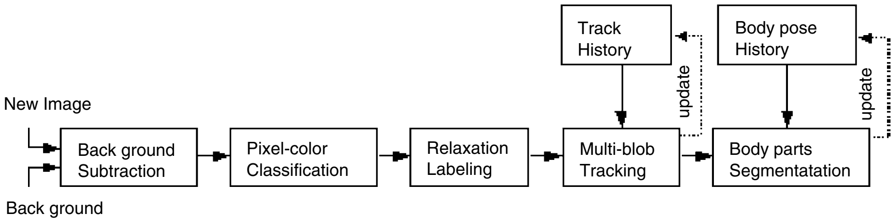

Fig. 1. System diagram.