Chapter 2
Tracking Persons: A Survey

Christine Leignel
Université Libre de Bruxelles, Belgium

Jean-Michel Jolion
Université de Lyon, France

ABSTRACT
This chapter presents a survey of methods used for tracking in video sequence. We mainly focus this survey on tracking persons. We introduce three main approaches. First, we present the graph based tracking approach where the sequence of tracked objects are embodied in a graph structure. Then we introduce the features (extracted from the images) based tracking and matching with a model. We survey the main primitives and emphasize the approaches based on 2D and 3D body model. We present the particular case of tracking in a network of cameras with the particle filtering method. Finally, As a generalization, we focus on the single vs. stereo approaches.

INTRODUCTION
Computer vision reflects a growing interest because of lower cost of new technology whose skills are growing. The video flow, traditionally processed by a human operator, is gradually being replaced by an automatic processing either to detect abnormal events, or to track a person into a scene for teleconferencing applications.

The pioneers in the field of tracking people are (Siebel, 2003), (O’Rourke & Badler, 1980) and (Hogg, 1983). In the area of surveillance, there are many tracking algorithms. The first step in any images sequence processing system for tracking people is to detect the movement of mobile regions in the image. Such regions are classified as individuals, groups and other classes of objects, and are grouped in a graph tracking to facilitate the tracking of individuals over a long period (case for persons who join or leave a group). We can classify the tracking methods in six categories:
Tracking Persons

- **Category 1**: methods, sometimes without a model, based region or « blobs » (set of pixels connected and grouped according to a criterion) tracker, based on color, texture, punctual primitives and contours ((Bremond, 1997), (Cai, Mitiche, & Aggarwal, 1995), (Khan, Javed, Rasheed, & Shah, 2001), (Lipton, Fujiyoshi, & Patil, 1998), (Wren, Azarbayejani, Darrell, & Pentland, 1997));

- **Category 2**: methods using a human body 2D appearance model ((Baumberg, 1995), (Haritaoglu, Harwood, & Davis, 2000), (Johnson, 1998)), with or without explicit model of the shape;

- **Category 3**: methods with a 3D articulated model ((Gavrila & Davis, 1996), (Sidenbladh, Black, & Fleet, 2000));

- **Category 4**: methods by background removal ((Haritaoglu, Harwood, & Davis, 1998), (Wren, Azarbayejani, Darrell, & Pentland, 1997)). The system can be more robust in textured environments by combining color, texture and movement to segment the foreground;

- **Category 5**: The temporal difference (two or three images) (Anderson, Burt, & Van Der Wal, 1985) yielding a binary map of motion (such as category 4) where motion pixels are grouped into « blobs » ((Haritaoglu, Harwood, & Davis, 2000), (Jabri, Duric, Wechsler, & Rosenfeld, 2000), (Zhao, Nevatia, & Lv, 2001)). The movements and interactions between individuals are obtained by the tracking of the « blobs »;

- **Category 6**: Another complementary approach to that of category 5 is the differential approach based on estimation of the velocity field at all points of the image, also known by motion detection. It calculates the velocity vector in the scene, making the invariance assumption between t and t+d. It defines an error function called DFD « Difference Deplaced Frames ». It seeks to minimize the DFD for all points of the image at time t. This family includes the method by « optical flow » (Barron, Fleet, & Beauchemin, 1994). The motion estimation by « optical flow » in terms of spatial and temporal variation of the function of intensity is a way of understanding the movement in a scene. Motion detection highlights mobile regions in the current image.

Background subtraction can extract moving objects but the background must be well modeled (by Gaussian or mixtures of Gaussian). This method, faster than other methods, is well suited for indoor environments where brightness is stable and the movements of the background low. The temporal difference is well suited to dynamic environments but suffers from the « problem of aperture » due to the uniform colors of objects in motion and does a poor extraction of primitives. The optical flow is a very robust technique to textured environments with movements in the background or camera movement, but is very expensive in computational cost and thus little used for real-time applications.

In the outdoor scenes, usually a single camera is enough to track an object for a long time. Objects can be in occlusion by the external elements: trees and buildings. A promising solution is to use a network of cameras for tracking objects in a cooperative and coordinated manner from a camera to another. (Matsuyama, 1998) presented such an approach in indoor environment where four cameras were tracking a moving object on the ground. (Chleq & Thonnat, 1996) have made a decision support system for the monitoring of human activity, triggering an alarm in a risky situation. For the detection of dangerous behavior in subways, (Cupillard, Avanzi, Bremond, & Thonnat, 2004) proposed an approach with multiple cameras to recognize individuals, groups or the crowd. Other applications such as teleconfer-