Robust Real-Time Human Activity Recognition from Tracked Face Displacements*

Paul E. Rybski and Manuela M. Veloso

The Robotics Institute and Computer Science Department,
Carnegie Mellon University, USA
{prybski, mmv}@cs.cmu.edu

Abstract. We are interested in the challenging scientific pursuit of how to characterize human activities in any formal meeting situation by tracking people's positions with a computer vision system. We present a human activity recognition algorithm that works within the framework of CAMEO (the Camera Assisted Meeting Event Observer), a panoramic vision system designed to operate in real-time and in uncalibrated environments. Human activity is difficult to characterize within the constraints that the CAMEO must operate, including uncalibrated deployment and unmodeled occlusions. This paper describes these challenges and how we address them by identifying invariant features and robust activity models. We present experimental results of our recognizer correctly classifying person data.

1 Introduction

Recognizing human activity is a very challenging task, ranging from low-level sensing and feature extraction from sensory data to high-level inference algorithms used to infer the state of the subject from the dataset. We are interested in the scientific challenges of modeling simple activities of people who are participating in formal meeting situations. We are also interesting in recognizing activities of people as classified by a computer vision system.

In order to address these challenges, our group is developing a physical awareness system for an agent-based electronic assistant called CAMEO (Camera Assisted Meeting Event Observer) [1]. CAMEO is an omni-directional camera system consisting of four FireWire cameras mounted in a 360° configuration, as shown in Figure 1. The individual data streams extracted from each of the cameras are merged into a single panoramic image of the world. The cameras are connected to a Small Form-Factor 3.0GHz Pentium 4 PC that captures the video data and performs image processing.

* This research was supported by the National Business Center (NBC) of the Department of the Interior (DOI) under a subcontract from SRI International. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by the NBC, DOI, SRI, or the US Government.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 87–98, 2005.

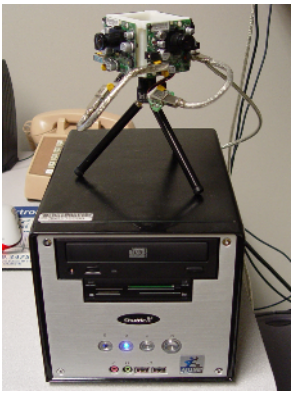

Fig. 1. The CAMEO system consists of a set of FireWire cameras arranged in a panoramic fashion and a small-form-factor PC

The panoramic video stream is scanned for human activity by identifying the positions of human faces found in the image. We make the assumption that people can be recognized in the image based on the location of their face. To be able to recognize people's activities within any meeting scenario, the challenge is to find appropriate features in terms of face positions without a global coordinate system. Low-level features are extracted from the raw dataset and are modeled as the observations for the recognition methods. The range of all possible human activities is reduced to a small discrete set.

We successfully solve this problem by two main contributions: (i) the identification of robust meeting features in terms of relative displacements; and (ii) the application of Dynamic Bayesian Networks (DBNs) to this problem, extending their use from other signal-understanding tasks.

2 Related Work

We use DBNs [2] to model the activities of people in the meetings. DBNs are directed acyclic graphs that model stochastic time series processes. They are a generalization of both Hidden Markov Models (HMM) [3] and linear dynamical systems such as Kalman Filters. DBNs are used by [4] to recognize gestures such as writing in different languages on a whiteboard, as well as activities such as using a Glucose monitor. Our system infers body stance and motion by tracking the user's face in a cluttered and general background rather than attempting to track the hands, which is difficult to do in general.

In [5], finite state machine models of gestures are constructed by learning spatial and temporal information of the gestures separately from each other. However, it is assumed that the gestures are performed directly in front of the camera and that the individual features of the face and hands can be recognized and observed without error.