Robotics and Autonomous Systems 56 (2008) 857–864

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Human behavior recognition using unconscious cameras and a visible robot in a network robot system

Keiichi Kemmotsu *, Yoshihiro Koketsu, Masato Iehara

Takasago Research and Development Center, Mitsubishi Heavy Industries, Ltd., 2-1-1 Shinhama, Arai-cho, Takasago 676-8686, Japan

ARTICLE INFO

Article history:
Available online 26 June 2008

Keywords:
Behavior recognition
Network robot
Sensor fusion
Ubiquitous network

ABSTRACT

We developed a network robot system integrating various types of robots via ubiquitous networks that introduce an interactive robot into areas in which people are located. In this paper, we present a human behavior recognition method necessary for providing guidance in a public space, which uses a tangible network robot system composed of a mobile robot and vision sensors embedded in an environment. We define some basic features that a person typically exhibits when in need of guidance. Various human behaviors were successfully interpreted by first recognizing the basic features and then forming their different combinations.

1. Introduction

Building a ubiquitous network infrastructure is a key to realizing an invigorated, safe, secure, exciting, and convenient society in the 21st century. Interactive robots living with people in non-industrial application areas have begun to appear. Their market size is expected to increase dramatically in the next ten years, despite the fact that the size of the conventional industrial robot market has not grown in the past ten years [1]. The convergence of ubiquitous network and robot technologies will produce an innovative "network robot system" [2]. The basic concept of network robots is that various types of robots (called "visible", "unconscious", and "virtual robots") are embedded in the ubiquitous network, and that diverse services are realized through collaboration and interaction among those robots.

The network robots need to identify human behavior to provide a desired service at the appropriate time. We developed a tangible network robot system consisting of a visible (mobile) robot and unconscious robots (environmentally embedded vision sensors) connected via a network [3]. Human behavior is described through data fusion processes, both spatially and temporally, using the vision sensors of the mobile robot and those embedded in the environment. The advantage of our system is that it is applicable to a large space, where human behaviors cannot be recognized by an individual sensor.

We intend to use the system to provide guidance and assistance services in a public space, as shown in Fig. 1. People usually exhibit typical behaviors when they are in need of guidance or assistance. The proposed system identifies such people by detecting certain basic behavioral features and then forming their different combinations. Recognition of human behavior will greatly enhance communication between people and the system, and help the system determine the type of service desired by the person.

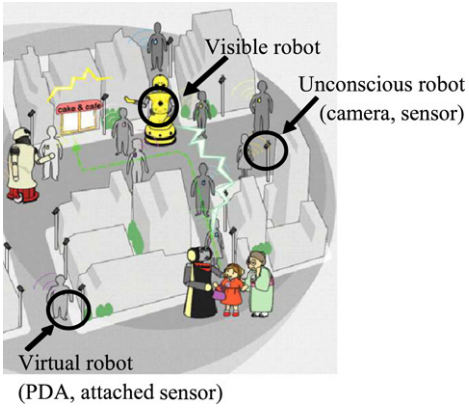

Fig. 1. Guidance and assistance services by a network robot system.

The features of our system are as follows:

- Our motion detection and tracking method uses a single camera; however, we can easily add more cameras to cover a larger area. When multiple cameras are used to observe the same person, position estimation accuracy can be improved without the need for solving any stereo correspondence problem. Furthermore, collaboration between unconscious

* Corresponding author. Tel.: +81 79 445 6752.
E-mail addresses: keiichi_kenmotsu@mhi.co.jp (K. Kemmotsu), yoshihiro_koketsu@mhi.co.jp (Y. Koketsu), masato_iehara@mhi.co.jp (M. Iehara).

0921-8890/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2008.06.004

robots and a visible robot improves system performance as a result of the mobility of the visible robot.
- Some human behaviors are identified by a two-stage approach. First, short-term motion is estimated using a stochastic model-based technique. Next, long-term motion recognition is performed using a histogram of short-term motions.
- Human behavior is recognized by combining five basic features of a person: the current position of the person, the path traversed by the person, the person's pose, the position and motion of the person's hands, and the position and direction of the person's face.

Related work

In recent years, there have been several research projects on detecting and tracking people, and recognizing human behavior [4]. Hu et al. [5] conducted a detailed survey on visual surveillance of object motion and behavior. Studies on recognizing human behaviors with vision sensors have typically involved two basic steps: (1) motion detection and object tracking, and (2) behavior description.

The first step mainly consists of using image processing and stochastic methods for data analysis, and the second step involves structural analysis of symbolic data gathered during the first step [6]. Several famous visual surveillance systems have been developed for motion detection and object tracking. The real-time visual surveillance system W^4 [7] uses a combination of shape analysis and tracking to locate people and their body parts, and tracks them using appearance models. This system uses a single grayscale camera. The Pfinder system, developed by Wren et al. [8], generates a 2D description of a person in a large room. It solves the problem of tracking in complex scenes, in which there is a single unoccluded person and a fixed camera.

Zhao and Nevatia [9] presented a single-camera system using an ellipsoidal human shape-model for segmentation and tracking of multiple humans. The shape-model enables segmentation of multiple humans with persistent occlusion (e.g., walking together) and provides a representation for tracking. A 3D model used in combination with a camera model and the assumption that people move on a ground plane, makes the approach suitable for a wide variety of viewpoints, automatically scales the model as people move, facilitates occlusion reasoning, and provides 3D trajectories. Human body postures are inferred in a 3D locomotion model over a period of time.

There are several single-camera detection, tracking, and behavior recognition algorithms, all of which face the same difficulties in tracking 3D objects using only 2D information in case of occluded objects as well as appearance changes [10]. Multiple-camera systems offer efficient and promising methods for dealing with occlusion.

Utsumi et al. [11] utilized multiple cameras to detect human motions, and deal with occlusion problems by using a viewpoint selection mechanism. Cai and Aggarwal [12] extended a single-camera tracking system by starting with tracking in a single-camera view, and then switching to another camera when the system predicts that the current camera will no longer have a good view of the subject. They used location, intensity, and geometry as parameters for matching among images taken by different cameras. Tsutsui et al. [13] developed an optical flow-based human tracking method, in which if an object occludes another object in a particular camera's view, the system predicts the 3D position and speed of motion of the occluded object using other cameras. Mittal et al. [10] developed a system that is capable of segmenting, detecting, and tracking multiple people in a cluttered scene using multiple synchronized surveillance cameras located far away from each other. The image regions of multiple cameras are fused to estimate the positions of the persons on the ground plane. Collins et al. [14] developed a system for video surveillance in an outdoor environment using multiple pan/tilt/zoom cameras. It can detect and track multiple persons and vehicles within cluttered scenes and monitor their activities over long periods of time.

These methods determine 3D positions of moving objects using stereo matching, or obtain an integrated representation by mapping 2D information from every camera into 3D space. Matching with incorrect correspondences or mapping with incorrect 2D image regions results in outliers.

Our system uses a single camera and can also handle multiple cameras; it does not assume stereo vision while obtaining 3D information. It uses an ellipsoidal human model with the assumption that a person moves on a known ground plane. When multiple cameras observe the same object, position estimation accuracy can be improved using a particle filter, rather than solving a stereo correspondence problem. Therefore, it is easier to add more cameras to cover a larger space.

After successfully tracking the moving objects in an image sequence, the problem of understanding object behavior from image sequences follows naturally. Behavior understanding involves analysis and recognition of motion patterns, and the production of high-level description of actions and interactions [5]. Bobick and Davis [15] proposed an appearance-based approach for recognizing human movement. They developed a view-based approach for representing and recognizing movements that is designed to support direct recognition of the motion itself.

A hidden Markov model (HMM) is a stochastic state model [16–18]; it implicitly handles data with spatio-temporal variability. The use of HMMs involves two stages: (1) training and (2) classification. In the training stage, the number of states of an HMM must be specified, and the corresponding state transition and output probabilities are optimized, so that the generated symbols can correspond to the observed image features of the examples within a specific movement class. In the matching stage, the probability with which a particular HMM generates the test symbol sequence corresponding to the observed image features is computed [5].

Yamato et al. [16] presented a method for recognizing human action in an HMM framework. Mesh features of binary moving human blobs were used as the low-level feature for learning and recognition. Learning was implemented by training the HMMs to generate symbol patterns for each class. Optimization of the model parameters was achieved using the Baum–Welch algorithm. Recognition was based on the output of the given image sequence using forward calculation [4].

An HMM framework covers short-term gesture recognition, such as "lifting the right arm", "nodding the head", and "waving a hand." However, the human behaviors on which we focus, include long-term motions lasting for at least several tens of seconds, and often up several minutes. Our idea is to hierarchize human motions into two motion classes, short-term and long-term, according to their duration, and to recognize them, step by step, through a two-stage recognition method.

A video understanding platform, called Video Surveillance Interpretation Platform (VSIP), has been proposed to automatically recognize behavior of individuals, groups of people, crowds, and vehicles, by detecting visual invariants (a visual invariant is a visual property or clue that characterizes a behavior) [6]. Different methods have been suggested for recognizing specific behavior types under different scene configurations.

We do not handle general behaviors but focus on certain typical behaviors observed when persons look for a guidance and assistance service in a public space. Therefore, a strong context restricts variation in the typical behaviors. Our idea is to recognize the typical behaviors based on five basic features for practical use. In this paper, we present a human behavior recognition method based on these basic features detected by our network robot system [3].