ELSEVIER

Image and Vision Computing 19 (2001) 833–846

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Monitoring human behavior from video taken in an office environment

Douglas Ayers, Mubarak Shah*

Computer Vision Lab, School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

Received 11 May 1999; revised 18 December 2000; accepted 20 January 2001

Abstract

In this paper, we describe a system which automatically recognizes human actions from video sequences taken of a room. These actions include entering a room, using a computer terminal, opening a cabinet, picking up a phone, etc. Our system recognizes these actions by using prior knowledge about the layout of the room. In our system, action recognition is modeled by a state machine, which consists of 'states' and 'transitions' between states. The transitions from different states can be made based on a position of a person, scene change detection, or an object being tracked. In addition to generating textual description of recognized actions, the system is able to generate a set of key frames from video sequences, which is essentially content-based video compression. The system has been tested on several video sequences and has performed well. A representative set of results is presented in this paper. The ideas presented in this system are applicable to automated security.

Keywords: Video; Action recognition; Key frames; Context

1. Introduction

Human action recognition has become an important topic in computer vision. One of the most obvious applications of this technology is in security. In this paper we present a system for recognizing human actions, which is geared toward automated security applications. A human action recognition system is useful in security applications for two important reasons. The first is to detect the entrance of an unauthorized individual and monitor that individual's actions. The second is to monitor the behavior of people who do belong in an area. By recognizing actions a person performs and using context, the behavior of the person can be determined. Some behaviors are inappropriate for certain persons. For example, someone using another person's computer without them being in the room or taking objects they are not permitted to take. The ability to associate names with people in the scene would help achieve both of these goals.

The system described in this paper recognizes human action in an environment for which prior knowledge is available. Three low-level computer vision techniques are used in our system. These techniques are skin detection, tracking and scene change detection. All three techniques use color images. Our system is capable of recognizing the actions of multiple people in a room simultaneously.

The system can recognize several actions: entering the scene, picking up a phone, putting down a phone, using a computer terminal, standing up, sitting down, opening something, closing something, picking up an object (specified as interesting in advance), putting down an object (previously picked up), leaving the scene and leaving the scene with an object. Several of the actions are fairly generic in nature. Objects that could be picked up by a person include briefcases, computer equipment, etc. Objects that can be opened and closed include cabinets, overhead compartments, doors, etc.

In addition to generating a textual description of recognized actions, our system reduces a video sequence into a smaller series of key frames, which concisely describe the important actions that have taken place in a room. Reduction of a video sequence to a series of key frames facilitates further analysis of the scene by computers or humans (such as deciding the name of the person who performed certain actions). Another advantage of key frames is the reduction of space required to store and time required to transmit them.

The rest of this paper is organized into six sections. Section 2 deals with related work in this area. In Section 3, we discuss the role of prior knowledge in our approach. Section 4 describes our system. In this section, we describe three low-level computer vision techniques, the high-level

* Corresponding author. Tel.: +1-407-823-2341; fax: +1-407-823-5419.
E-mail addresses: ayers@cs.ucf.edu (D. Ayers), shah@cs.ucf.edu (M. Shah).

0262-8856/01/$ - see front matter
PII: S0262-8856(01)00047-6

method we use for action recognition and discuss strategies for determining key frames from video sequences. The low-level techniques are skin detection, tracking and scene change detection. All of the low-level techniques use color imagery and provide useful information for the action recognition, which is based on finite state machine model. In this section, we also discuss strategies for determining key frames from video sequences. Section 5 deals with the experimental results. We have tested our ideas with several video sequences and we provide a summary of our analysis. In Section 6, we comment on the limitations of our system. Finally, in Section 7 we provide conclusions and propose some future work in this area.

2. Related work

There is a large body of work on the analysis of human motion reported in the literature. Please see three excellent surveys: Cédras and Shah [16], Aggarwall and Cai [1] and Gavrila [3] for a detailed treatment of this subject. For a sample of recent work, refer to special section of IEEE PAMI on Video Surveillance [2]. In the following, we briefly describe some sample work in this area, which in no way is exhaustive and complete.

Bobick and Davis [17] described a method to recognize aerobic exercise from video sequences. First they apply change detection to identify moving pixels in each image of a sequence. Then Motion History Images (MHI) and Motion Recency Images (MRI) are generated. MRI is the union of all images after change detection has been applied, which represents all the pixels, which have changed in a whole sequence. MHI is a scalar-valued image where more recent moving pixels are brighter. In their system, MHI and MRI templates are used to recognize motion actions (18 aerobic exercises). Several moments of a region in these templates are employed in the recognition process. The templates for each exercise are generated using multiple views of a person performing the exercises. However, it is shown that during recognition only one or two views are sufficient to get reasonable results.

Stauffer and Grimson [4] presented a probabilistic approach for background subtraction to be used in a visual monitoring system. Their main contention is to use a mixture of Gaussian models as compared to a widely used single Gaussian to model the color values of each pixel. They claim that each pixel is an independent statistical process, which may be a combination of several processes. For example, swaying branches of a tree result in a bimodal behavior of pixel intensity. The authors also discuss simple classifications based on aspect ratio of tracked objects. Their method involves developing a codebook of representations using an on-line vector Quantization on the entire set of tracked objects.

Shershah et al. [18] presented a method for modeling and interpretation of multi-person human behavior in real-time to control video cameras for visually mediated interaction. They use implicit and explicit behaviors of people. Implicit behaviors are defined as body movement sequences that are performed subconsciously by the subjects. Explicit behaviors are performed consciously by the subjects and include pointing and waving gestures. Given a group behavior they introduce a high-level interpretation model to determine the areas where the cameras are focused.

Kojima et al. [14] proposed an approach to generate a natural language description of human behavior from real video images. First, a head region of a human is extracted from each frame. Then, using a model-based method, 3-D pose and position of the head are estimated. Next, the trajectory of the head is divided into segments and the most suitable verb is selected.

Davis and Shah [15] were the first ones to use a finite state machine approach to model different phases of gestures to avoid the time consuming step of warping. This paper demonstrated recognition of seven gestures which are representatives for actions of 'left', 'right', 'up', 'down', 'grab', 'rotate' and 'stop'. The system was able to recognize a sequence of multiple gestures. The vector representation for each of the seven gestures is unique. For the left gesture, the thumb and index fingers do not move, while the remaining fingers move from top to bottom. For the right gesture, all fingers move from right to left. The other five gestures are similar.

Intille and Bobick [12] and Intille et al. [13] discussed the use of context to enhance computer vision applications. In these articles, context is taken advantage of primarily to perform tracking.

The main goal of Rosin and Ellis's [10] system was to differentiate between humans and animals to reduce false alarms. This system is especially interesting because of its use of context to help improve recognition. To improve the performance of their intruder detection system, the authors include a map of the scene, which shows areas such as sky, fence and ground. This helps to differentiate between a person moving through the scene and an animal (such as a bird).

Nagai et al. [9] and Makarov et al. [11] have also written papers which involve intruder detection. Nagai et al. uses optical flow to find intruders in an outdoor environment. Makarov et al. focuses on intruder detection in an indoor scene. Comparison between edges in the current frame and edges in the background is used to perform intruder detection on image sequences with variant lighting.

Olson and Brill [7] developed an automated security system for use in an indoor or outdoor environment. Their system can detect events such as entering, exiting, loitering and depositing (an object). A map of the room which may have special regions labeled, such as tables, is used by the system to keep track of a person's movement through the room. A log of important events is kept by their system and alarms can be associated with certain events occurring at certain times in certain regions. The system also learns the