Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 24 (2006) 455–472
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Minimal-latency human action recognition using reliable-inference

James W. Davis *, Ambrish Tyagi

Department of Computer Science and Engineering, Ohio State University, 491 Dreese Lab, 2015 Neil Avenue, Columbus, OH 43210, USA

Received 26 October 2004; received in revised form 3 November 2005; accepted 31 January 2006

Abstract

We present a probabilistic reliable-inference framework to address the issue of rapid detection of human actions with low error rates. The approach determines the shortest video exposures needed for low-latency recognition by sequentially evaluating a series of posterior ratios for different action classes. If a subsequence is deemed unreliable or confusing, additional video frames are incorporated until a reliable classification to a particular action can be made. Results are presented for multiple action classes and subsequence durations, and are compared to alternative probabilistic approaches. The framework provides a means to accurately classify human actions using the least amount of temporal information.

Keywords: Action recognition; Reliable-inference; MAP; Video analysis

1. Introduction

Recognition of human actions is a challenging task that is applicable to various domains including surveillance, video annotation, human–computer interaction, and autonomous robotics. A fundamental question regarding action recognition is 'how much time (or video frames) is actually necessary to identify common human actions'? In the extreme case, people can easily recognize several different actions from looking at just a single picture—one need only flip through the pages of a newspaper or magazine to make this point. Other actions may require seeing additional video frames (perhaps the entire action sequence).

There are several instances of how biological perception is able to identify simple actions/gestures by observing just a small portion of the motion. For example, [22] showed that humans are capable of perceiving and distinguishing simple actions, such as walking, running, and jumping-jacks, from point-light exposures of only 200 ms. Clearly, no periodic information is being exploited. Other instances of such behavior can be seen with animal tricks/performances in response to the trainer's gesture commands. Often the animals are trained on a closed set of gestures and are able to identify their trainer's gesture from just an initial brief exposure of the movement.

A system capable of recognizing human actions from the *smallest* number of video frames would be particularly advantageous to automatic video-based surveillance systems. For instance, consider small unmanned aerial vehicles (UAVs) equipped with video cameras. The UAV's view area is constantly and rapidly changing, and therefore, immediate decisions about the activity in the scene are desirable within a few frames. This is also the case for multi-camera surveillance systems having limited computational processing time scheduled per camera. Even when longer duration video is available, rapid action detection may be particularly helpful in bootstrapping more sophisticated action-specific tracking or recognition approaches.

An important ability for a rapid (low-latency) action recognition system is to *automatically* determine when enough of a video sequence has been observed to permit a reliable classification of the action occurring, as opposed to making forced decisions on short (and potentially unreliable) video exposures or instead waiting for the entire (long) video sequence. This problem becomes even more challenging when the observation could start from any point (temporal offset) during the action sequence (i.e. the observed sequence does not always begin at the expected start pose of the action). For example, consider the case when a camera is switched on to detect actions and there is already an action in progress.

In this paper, we present an approach for quickly recognizing non-nested, temporally constrained human actions from the smallest video exposure possible. The method is

* Corresponding author. Tel.: +1 614 292 1553; fax: +1 614 292 2911.
E-mail addresses: jwdavis@cse.ohio-state.edu (J.W. Davis), tyagia@cse.ohio-state.edu (A. Tyagi).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.01.012

formulated in a probabilistic inference framework and is trained with examples. First, the reliability of an input frame/subsequence is examined with a series of *a posteriori* ratio comparisons to the possible action classes (using Hidden Markov Model (HMM) output likelihoods and action priors). These reliability ratios assess the 'goodness' of an input for inferring a particular action class with respect to the probability of committing an error. If the input is deemed unreliable (not strongly indicative of a particular action), then no action classification takes place. In this case, the input is expanded/extended with additional video frames and then re-examined with the posterior ratios for a reliable-inference. This process is continued until either a reliable action inference is made or there are no more video frames to include (e.g. the person exited the scene). The worst-case scenario is that the entire sequence is evaluated (the default for most other approaches). Therefore, the approach is useful for recognizing actions that have some distinct difference at some point prior to the end of the action sequence. The method keeps tight bounds on the classification error by incorporating additional information for brief and confusing video inputs rather than forcing an early, and likely erroneous, action classification with a fixed-length sequence.

The main advantage of the proposed method is that the system only makes classifications when it believes the input is 'good enough' for discrimination between the possible actions. This is particularly favorable when there is a high cost for making errors and low (or no) cost for passively waiting for more video frames to arrive (advantageous with real-time video). Other popular probabilistic approaches, such as *maximum likelihood* (ML) and *maximum a posteriori* (MAP), perform a *forced-choice* classification for a fixed-length input regardless of the saliency/reliability of the input. To our knowledge, no previous attempts have been made to accurately classify human actions using the least amount of temporal information.

We demonstrate our rapid-and-reliable action recognition approach with sets of common actions having different video exposures of the full action duration. First, we examine the reliability of only a single frame to discriminate different actions, and demonstrate how single frame action classification can be a potential source of high confusion. We then address the reliability of action subsequences (multiple frames). Initially, we examine our method with a set of actions that begin in the same starting pose but gradually diverge into the different actions over time. Our goal here is to determine the *minimal* video exposures from the start frame/pose needed to reliably discriminate the actions. Next, we consider an even more general and difficult scenario with subsequences that are not constrained to begin with the start frame of the original sequence. In each of the experiments, we compare our recognition results with competing classification techniques.

We begin with a review of related methods for detecting and recognizing actions in single frames and video sequences (Section 2). We then formulate our probabilistic reliable-inference approach (Section 3). Next, we compare the approach to posterior analysis and discuss similar techniques that are capable of rejecting uncertain inputs (Section 4). Then, we present experimental evaluations and results with a discussion (Section 5) and provide possible extensions to the framework (Section 6). Lastly, we provide a summary and conclusion of the work (Section 7).

2. Related work on action recognition

The importance of human action recognition is evident by its increased attention in recent years (see surveys in [1,17,37]). Previously, several action recognition techniques have been proposed, but here we briefly mention only a few representative approaches that are relevant to comparing the single- and multiple-frame analysis domains.

For identifying pedestrians in a *single image*, [29] used wavelets to learn a characteristic pedestrian template and employed support vector machines for classification. In [18], a hierarchical coarse-to-fine template approach with edge maps was used to detect pedestrians. In [32], a general unsupervised recognition framework using clustering was employed to recognize several silhouette poses. In the *two frame* category, [23] used two simple properties (dispersedness, area) for discriminating humans and vehicles from image difference results, and maintained consistency over time with a temporal classification histogram. A cascaded AdaBoost architecture was proposed in [35] that used rectangular filters on intensity and difference images for pedestrian detection. None of these approaches attempt to recognize extended temporal actions based on characteristic movement patterns.

Dynamic action recognition methods from *multiple frames* or sequences typically involve analysis of trajectories or templates. For periodic actions such as walking and running, trajectory-based approaches for single and multiple cycle exposures include frequency-based Fourier methods [24], feature-based properties (e.g. stride) [11], spatio-temporal patterns [28], and HMMs [9]. With no part tracking, a different approach is to use spatio-temporal templates derived from the image sequence directly, including generic layered templates [6] and other periodic template representations [10,25,26,30].

Our work here is focused on action recognition from the *shortest-duration* video exposures needed to achieve high recognition rates with low recognition latency. In previous work, we presented initial results of the proposed reliable-inference framework for discriminating actions from single frames (poses) [13] and also provided a hierarchical extension for multiple frames using Motion History Images (MHIs) [12]. In this paper, we extend our prior work to efficiently examine multiple subsequence durations for different actions and incorporate HMM temporal models (instead of hierarchical MHIs) for the probabilistic evaluations. Many of the action recognition approaches described above could be potentially incorporated into our framework, however, we are not aware of any other general approaches related to our goal of reliable, yet minimal-latency, action recognition.