Pattern Recognition 41 (2008) 3078 – 3091

Contents lists available at ScienceDirect

Pattern Recognition

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

Boosted string representation and its application to video surveillance

Jun-Wei Hsieh*, Yung-Tai Hsu

Department of Electrical Engineering, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li 320, Taiwan

ARTICLE INFO

Article history:
Received 26 February 2007
Received in revised form 1 February 2008
Accepted 15 March 2008

Keywords:
Behavior analysis
Centroid contexts
String matching
Boosting algorithm

ABSTRACT

This paper presents a new behavior analysis system for analyzing human movements via a boosted string representation. First of all, we propose a triangulation-based method to transform each action sequence into a set of symbols. Then, an action sequence can be interpreted and analyzed using this string representation. To analyze action sequences with this string representation, three practical problems should be tackled. Usually, an action sequence has different temporal scaling changes, different initial states, and symbol converting errors. Traditional methods (like hidden Markov models and finite state machines) have limited abilities to deal with the above problems since many unknown states should be constructed and initialized. To tackle the problems, a novel string hypothesis generator is then proposed for generating a bank of string features from which different invariant features can be learned for classifying behaviors more accurately. To learn the invariant features, the Adaboost algorithm is used and modified to train a strong classifier from the set of string hypotheses so that multiple human action events can be well classified. In addition, a forward classification scheme is proposed to classify all input action sequences more accurately even though they have various scaling changes and coding errors. Experimental results prove that the proposed method is a robust, accurate, and powerful tool for human movement analysis.

1. Introduction

Human action analysis [1,2] is an important task in various application domains like video surveillance [1–12], video retrieval [13], human–computer interaction systems, and so on. Characterization of human action is equivalent to dealing with a sequence of video frames that contain both spatial and temporal informations. The challenge in human action analysis is how to properly characterize spatial–temporal informations and then facilitate subsequent comparison/recognition tasks. There have been many approaches proposed for tackling different problems in video-based human action analysis. For example, Aggarwal et al. [14,13] used multi-layer finite state automata (FSA) to model and recognize human interactions from videos. Here, a low-level FSA is used for analyzing body parts and a high-level FSA is used for analyzing human interactions. The important feature in FSA is its state transition function which can be deterministic or non-deterministic for different object classifications. For the deterministic approach, Hareing et al. [6] used a state transition graph to detect hunting events in wildlife videos. In addition, Cucchiara et al. [10] used a probabilistic projection map to model postures and then performed frame-by-frame posture classification to recognize human behaviors. For the non-deterministic one, Hongeng et al. [7] proposed a probabilistic FSA and used it and a set of user-defined hierarchical multiple scenarios to analyze various human interactions like handshaking. Wada et al. [15] used a non-deterministic FSA to analyze image sequences obtained from multiple views for multi-object behavior recognition. The advantage of the FAS approach is that it does not need a large set of training data for behavior modeling. However, the number of states and the transitions between states often needs manual efforts to be properly initialized and labeled.

In addition to FSA, context-free grammar is another good tool to analyze semantic events from videos. For example, in Ref. [4], Ivanov et al. used a context-free grammar parsing scheme to analyze video targets like persons or cars. In Ref. [16], Ogale et al. used multi-view training videos to automatically create a view-independent probabilistic context-free grammar for analyzing human actions. In Ref. [17], Brand used a simple non-probabilistic grammar to recognize human behaviors in videos. In addition, Kojima et al. [18] used a concept hierarchy to recognize various behaviors of single person by translating them to natural language-based descriptions. The difficulty in the context-free grammar approach is how to transform video sequences into semantic descriptors.

Hidden Markov model (HMM) [19–21] is another commonly used stochastic method for human action analysis. It uses two processing phases (training and recognition) to give different sophisticated probabilistic analyses on uncertainty targets. At the training phase, this approach models each possible scenario with a HMM whose parameters are trained a priori with the so-called Baum–Welch algorithm. During the recognition phase, each body posture is

* Corresponding author. Tel.: +886 3 463 8800x2430; fax: +886 3 463 9355.
E-mail address: shieh@saturn.yzu.edu.tw (J.-W. Hsieh).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.03.026

recognized according to the HMM producing the highest probability. For example, Oliver et al. [3] used HMMs for classifying the interactions between humans into different types such as meeting, approaching, walking, and so on. In addition, Nguyen et al. [8] developed a HHM-based surveillance system for recognizing human behaviors in multiple camera environments like the corridor, staff room, and vision lab. The system uses an abstract hidden Markov memory model and object trajectories to recognize complex human behaviors. In Ref. [9], Navaratnam et al. used HMMs and a set of 2D templates for recovering 3D human body poses. In Ref. [22], Jojic et al. embedded a template matching technique in a transformed hidden Markov model (THMM) for the framework of object tracking. A serious problem related to HMMs involves how to specify and learn the HMM model structure. Usually, human actions have different spatial and temporal scaling changes. The changes will create many unknown states and make the HMM framework become difficult and complicated in finding a proper state transition graph and its model parameters for human behavior analysis. In addition, human actions have various unexpected conditions which will make a HMM have many unexpected initial states and thus lead to its failure to recognize a correct behavior type.

This paper presents a novel string-based method for modeling human behaviors and recognizing them directly from videos. First of all, we use a triangulation-based method [23] to convert a human action sequence to a set of symbols. Then, a novel string hypothesis generator is proposed for generating a bank of string features to represent a human action. The representation can tackle three problems when analyzing human actions; that is, (a) the uncertainty in human actions, (b) spurious detections, and (c) temporal inconsistency. For (a), a human action usually will not begin with the same status. The uncertainty will increase different difficulties and complexities in behavior modeling. Since our method does not require any state transition graph (often created in HMM and FSA), it can avoid the failure of state errors if some state conditions are not correctly set. For (b), noise or different practical conditions will make some action frames be wrongly detected or recognized. The spurious detection or recognition will cause wrong state transitions in HMM or FSA. However, in this paper, the problem can be easily tackled if a boosting technique is used. For (c), a human cannot often perform an action with the same speed. The temporal inconsistency will increase many difficulties in behavior modeling. This paper proposes a novel hypothesis generator to generate different string features. Then, from the set of string features, we use the Adaboost algorithm to train a strong classifier for solving this problem in behavior analysis. Since this algorithm is originally designed for solving two-class problems, we further use an error correction concept and a forward scheme to improve its ability so that multiple action events can be well classified. The trained classifier has higher tolerances to different converting errors when frames are converted to symbols. In addition, it can classify any action sequences even though they have large temporal scaling changes. Experimental results demonstrate the feasibility and superiority of our proposed approach in human behavior analysis.

The remainder of the paper is organized as follows. Section 2 describes the flowchart of our proposed system. Then, in Section 3, the triangulation-based technique for frame-to-symbol converting is discussed. Section 4 describes details of our string hypothesis generator. Details of the Adaboost algorithm are discussed in Section 5. Section 6 describes the experimental results. Finally, some conclusions are made in Section 7.

2. Overview of the proposed system

The flowchart of the system is shown in Fig. 1. Firstly, we apply background subtraction to extract body postures from video sequences and then derive their boundaries by contour tracing. Next, a Delaunay triangulation technique [24] is used to divide a posture into different triangular meshes. To extract the posture feature, we propose a graph search method that builds a spanning tree from the triangulation result. The spanning tree corresponds to the skeletal structure of the analyzed body posture from which we construct a new posture descriptor, namely, the centroid context descriptor, for recognizing postures. After posture classification, each posture will be assigned a semantic symbol so that each human movement can be converted and represented by a set of symbols. Based on this representation, we first use the Adaboost algorithm to train a strong string classifier for behavior analysis. At the recognition stage, according to the learned string classifier, the proposed system will extract a set of important string hypotheses from the input video for recognizing different human movements. Even though they have different temporal and spatial variations, our system still performs well to recognize them. In what follows, details of the triangulation technique for posture classification will be described.

3. Deformable triangulation technique for frame-to-symbol converting

In this paper, it is assumed that all video sequences are captured by a stationary camera. When the camera is static, the background of the captured video sequence can be reconstructed using a mixture of Gaussian models [25]. This allows us to detect foreground objects by subtracting the background. We then apply some simple morphological operations to remove noise (like holes). After that, different postures in a video sequence can be well extracted. Although there are some approaches [12] which can analyze human postures directly from a moving background, the ill-posed problem in body segmentation makes their performances be limited in real cases. After subtraction, to better convert a sequence into a set of symbols, we use the constrained Delaunay triangulation technique [24] to make each posture into triangular meshes. Then, different features are extracted from the triangulation result to characterize each posture. In what follows, details of posture classification using our proposed centroid contexts are first discussed. Then, in Section 3.2, a clustering scheme is proposed for extracting a set of key postures for converting each frame (or posture) to symbols.

3.1. Posture classification using centroid contexts

Assume P is a posture which is a binary map of a person extracted through a background subtraction technique and V is a set of control points sampled along the boundary of P. We adopt the constrained Delaunay triangulation technique proposed by Chew [24] to divide V into triangular meshes. As shown in Fig. 2, Φ is the set of interior points of V in R^2. In addition, $e(u, v)$ denotes the edge between two vertices u and v (the straight line connecting u and v). Given three vertices v_i, v_j, and v_k on V, the triangle $\Delta(v_i, v_j, v_k)$ belongs to the constrained Delaunay triangulation if and only if

(i) $v_k \in U_{ij}$, where $U_{ij} = \{v \in V | e(v_i, v) \subset \Phi, e(v_j, v) \subset \Phi\}$; and
(ii) $C(v_i, v_j, v_k) \cap U_{ij} = \emptyset$, where C is a circum-circle of v_i, v_j, and v_k. That is, the interior of $C(v_i, v_j, v_k)$ does not have a vertex $v \in U_{ij}$.

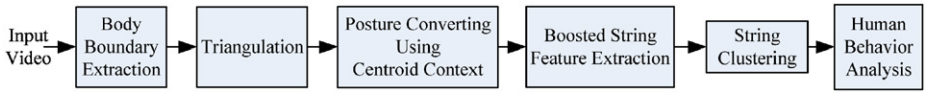

Fig. 1. Flowchart of the proposed system.