International Journal of Approximate Reasoning 52 (2011) 526–540

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Approximate reasoning and finite state machines to the detection of actions in video sequences ☆

L. Rodriguez-Benitez [a,*], C. Solana-Cipres [a], J. Moreno-Garcia [b], L. Jimenez-Linares [a]

[a] Oreto Research Group, Universidad de Castilla-la Mancha, Escuela Superior de Informatica, Paseo de la Universidad s/n, 13071 Ciudad Real, Spain
[b] Escuela Ingenieria Tecnica Industrial, Avda. Carlos III s/n, 45071 Toledo, Spain

ARTICLE INFO

Article history:
Received 30 April 2010
Revised 21 October 2010
Accepted 14 December 2010
Available online 25 December 2010

Keywords:
Approximate reasoning
Action recognition
Video analysis

ABSTRACT

In this paper a novel approach for recognizing actions in video sequences is presented, where the information obtained from the segmentation and tracking algorithms is used as input data. First of all, the fuzzification of input data is done and this process allows to successfully manage the uncertainty inherent to the information obtained from low-level and medium-level vision tasks, to unify the information obtained from different vision algorithms into a homogeneous representation and to aggregate the characteristics of the analyzed scenario and the objects in motion. Another contribution is the novelty of representing actions by means of an automaton and the generation of input symbols for the finite automaton depending on the comparison process between objects and actions, i.e., the main reasoning process is based on the operation of automata with capability to manage fuzzy representations of all video data. The experiments on several real traffic video sequences demonstrate encouraging results, especially when no training algorithms to obtain predefined actions to be identified are required.

1. Introduction

In recent years there has been a considerable growth in the development of a great number of systems for security, home automation, zone traffic regulation, and others that are based on automatic video analysis. This area of computer vision involves event recognition, behaviour understanding, action representation, and natural language description of behaviours. These sub-areas are usually linked and worked within this context.

The understanding of behaviours can be seen as the classification of time varying feature data, i.e., matching unknown sequences with reference sequences to represent normal behaviours. The feature data to be classified is achieved through motion detection and tracking processes to transform the pixel-level data into the higher-level data needed in behaviour understanding. This fact implies that both areas use similar techniques to solve the problem, as pointed out by Hu and Tan [8]: dynamic time warping (DTW), finite-state machines (FSMs), hidden Markov models (HMMs) and variations (CHMMs and PHMMs), neural networks (NNs), syntactic techniques, and non-deterministic finite automatons (NFAs). A long dissertation about analysis of human movement, mentioning techniques and related work can be found in [4]. In this concrete field of vision-based human action recognition, Poppe [20] presents a detailed overview of current advances in the field and proposes a classification of techniques depending on image representation and action classification. Related to image representation, most of current methods are designed for limited view variations. Nevertheless, Junejo et al. [11] address recognition of human actions under view changes.

☆ This work has been funded by the Regional Government of Castilla-la Mancha and the Spanish Ministry of Science under projects PII1C09-0137-6488 and TIN2009-14538-C02-02.

* Corresponding author.
E-mail address: luis.rodriguez@uclm.es (L. Rodriguez-Benitez).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.12.003

According to Van-thinh [28], techniques to represent and recognize temporal scenarios for automatic video interpretation can be classified into three different categories: (1) probabilistic and stochastic: Bayesian Networks [6] and Hidden Markov Models (HMMs). The main feature of these techniques is the ability to explicitly model uncertainty through probabilistic reasoning. HMMs are non-deterministic state machines which, given an input, move from state to state according to various transition probabilities. HMMs can deal with unsegmented data. However, one of the limitations of using these techniques is that they are not suitable for encoding the dynamic of long-term activities. Another limitation is the long data sets needed to train the systems. (2) symbolic: action classification, automata, constraint satisfaction problem. These techniques aim at transforming numerical observations into symbolic scenarios. (3) symbolic temporal techniques: temporal constraint satisfaction problem, plan recognition, event calculus and Petri nets, chronicle recognition and temporal constraint propagation. These techniques try to model temporal relations at a symbolic level.

Another classification of works for event detection in video sequences is proposed in [5]: (1) approaches that detect events in videos based on predefined models like templates, rules or constraints; (2) approaches that automatically learn event models using training data and (3) approaches that do not model events and uses low-level information of the video combining with clustering methods.

The main problems to be solved in activity recognition are knowledge representation about objects and scenarios and how to carry out the reasoning process. Another important problem is that the results of this techniques depends on low-level vision tasks as segmentation [27] and tracking [12,19]. The motivation of this paper is to develop a technique that helps to interpret video sequences by using Fuzzy Logic as a tool of knowledge representation and approximate reasoning techniques supported by a finite state automaton. In general, most of the current activity recognition approaches are composed of designed models for specific activity types that suit the goal in a particular domain developing procedural recognition methods. In the field of the approximate reasoning, rule-based methods are used to approximate the belief of the occurrence of activities [7].

There is a rich literature on behaviour understanding; some works closely related to our research are then summarised. For example, Hongeng et al. considered an activity is composed of action threads [7]. Each single-thread action is executed by a single actor and is represented by a stochastic finite automaton of event states. Each state represents features of the trajectory and shape of moving blobs. They introduced a hierarchical activity representation that allows the recognition of a series of actions performed by a single mobile object. Bobick and Ivanov presented an article [2] inspired by work in speech recognition where the inference problem is divided into two levels. The lower one obtains candidate detections of low level features and the higher one uses these values to provide an input stream for a stochastic context-free grammar parsing mechanism. The grammar and parser allow the inclusion of a priori knowledge about the structure of temporal events in a given domain. A simpler and faster pattern recognition approach is proposed in Campbell and Bobick [3], where a phase-space representation in which the velocity dimensions are projected out is presented. Ivanov and Bobick [10] also presented a remarkable work at a higher level behaviour, employing a two-layer event abstraction. Event primitives are modeled by HMM and a stochastic context free grammar (SCFG) is developed for the problem domain. Robertson and Reid [23] also researched a method for human activity recognition in video using HMM. They describe on behaviours by combining a data-driven non-parametric learning, plus a classification technique for actions, plus a HMM representation of action sequences. Lin et al. [14] proposed a semantic event representation and recognition using syntactic attribute graph grammar. Their approach defines a graph grammar that allows syntactic representation of complex events; the grammar model decomposes a semantic event into a composition of actions using a dictionary of spatio-temporal relations. Kollnig et al. [13] investigated ways to describe scene motion in terms of natural language by using a logic-based framework.

Finally, in recent years, there have been relevant works toward the fusion of multi-modal information like color, motion, acoustic, speech, and text for event and action recognition. However, these approaches rely on contextual knowledge and are limited to specific domains such as football games [9], athletic jumps [21], hockey [15], classrooms, or stationary traffic monitoring [22]. Other recent works try to avoid the high cost of the training stages and the expert implication on the configuration of the system variables because most of these methods often involve a careful hand-tuning of parameters such as threshold values.

1.1. Contributions of this work

A new approach to action recognition performed by different actors in a video sequence is presented. This work differs from classical techniques of computer vision in two aspects. On the one hand, the data relative to detected objects in the video sequence is modeled as a set of linguistic elements. On the other hand, the analysis technique is based on a Mealy machine used to represent predefined actions (behaviors to be detected). This automaton does not obtain a relation (*object*, *action*) as result: "the object performs action number i" but instead: "the membership value of the object to action number i is Z", where Z is the final output of the sequential machine.

1.2. Paper structure

In Fig. 1, a graphical representation of the overall process described in this paper is shown. The numbers in the figure represent a number of section or subsection. So, this paper is organized as follows. In Section 2, the transformation of data into fuzzy domain is justified and the definitions of linguistic elements used to represent the data obtained from segmentation