Signal Processing 93 (2013) 2151–2168

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

STV-based video feature processing for action recognition

Jing Wang, Zhijie Xu*

School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, United Kingdom

ARTICLE INFO

Article history:
Received 29 October 2011
Received in revised form
1 June 2012
Accepted 7 June 2012
Available online 16 June 2012

Keywords:
Video events
Spatio-temporal volume
3D segmentation
Region intersection
Action recognition

ABSTRACT

In comparison to still image-based processes, video features can provide rich and intuitive information about dynamic events occurred over a period of time, such as human actions, crowd behaviours, and other subject pattern changes. Although substantial progresses have been made in the last decade on image processing and seen its successful applications in face matching and object recognition, video-based event detection still remains one of the most difficult challenges in computer vision research due to its complex continuous or discrete input signals, arbitrary dynamic feature definitions, and the often ambiguous analytical methods. In this paper, a Spatio-Temporal Volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human actions recorded in videos. The distinctive characteristics and the performance gain of the devised approach stemmed from a coefficient factor-boosted 3D region intersection and matching mechanism developed in this research. This paper also reported the investigation into techniques for efficient STV data filtering to reduce the amount of voxels (volumetric-pixels) that need to be processed in each operational cycle in the implemented system. The encouraging features and improvements on the operational performance registered in the experiments have been discussed at the end.

1. Introduction

Computer vision theories and practices have been experiencing an accelerated development period over the last 2 decades. Demands from applications such as intelligent surveillance systems, machine vision, robotics and automated guided vehicle (AGV), innovative human computer interface (HCI), and even digital entertainment and computer games have pushed this trend. As an important application of video event detection technologies, action recognition is one of the hotly-pursued areas in computer vision research for automatically detecting and interpreting real-world events and activities such as human gestures, crowd actions, or other object patterns through extracting and analysing video features denoted in the spatial-temporal frame of reference.

As widely recognised, raw video data often suffer from high signal noise ratio and low resolution that require tedious and time-consuming processes to clean up at a frame-by-frame (FBF) basis. This research aims at investigating 3D spatial- and temporal-volume based event descriptors for video content analysis and event detection. The rationale behind this research is to develop robust and intuitive video processing methodologies and techniques for tackling challenging tasks such as large video database indexing and querying, automated surveillance system design, and Internet-based online video management.

Generally speaking, the large variations on conditions where videos being made and the complicated nature of human gestures and postures are still posing great challenges to the conventional pixel-based human action recognition strategies. The complexities exposed to video event detection tasks can be classified into three categories. Firstly, the

* Corresponding author. Tel.: +44 1484 472156; fax: +44 1484 421106.
E-mail address: z.xu@hud.ac.uk (Z. Xu).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.009

semantic meaning of an "event" in a video is often ambiguous since the variations of potential "event makers" defined by a particular application or system operators. Secondly, due to the limitations of today's videoing equipment and storage restrictions, real-life video data, especially those from surveillance systems and Closed-Circuit Television Systems (CCTV), inherit great technical difficulties to process. For example, the boundaries between an "event" signal and its "background" noise is often either inexplicit or occluded, which renders a complete separation of the two signals almost impossible. In many recent pilot researches, a background is often simplified into static sections in continuous video frames. However, this presumption is not always applicable in complex real-life scenarios, i.e., multiple dynamic objects or illumination changes can all cause confusions and blurriness. The third difficulty can be caused by the uncertainty of video event durations. The time-elapsed factor for encapsulating a discrete event is closely coupled with the nature of the event and the videoing sampling rates which might be substantially varied for different application systems.

Compared with the often non-generic local and 2-dimensional (2D) feature-based approaches such as head-shape detection, and torso-limb spatial relations, the emerging concept of the spatial and temporal feature space and the so-called Spatio-Temporal Volume (STV) data structure – first introduced by Adelson and Bergen [1] – have shown their promising global feature representation potentials for 3D and dynamic video event feature representation and pattern recognition. An STV model is capable of encapsulating static and dynamic video content features, hence simplifying an event recognition task into corresponding 3D geometric feature extraction and matching operations. As illustrated in Fig. 1, the STV is a volume space confined by a 3D coordinates system denoted by *X*, *Y* (spatial) and *T* (temporal) axes, in which a STV model can be represented by a 3D shape formed by a stack of 2D arrays of pixels, referred as voxels standing for volumetric-pixels.

In this research, an STV and a region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human action events recorded in videos. An innovative pattern recognition algorithm has been developed to harness the promising characteristics of the STV event models. The core of this approach is built upon the RI method to compare the STV "shapes" extracted from video inputs to the predefined 3D event templates.

It has long been understood that many 2D digital image processing (DIP) and pattern recognition algorithms can be readily extended to 3D domains using higher dimensional vectors for defining complex features such as 3D curves, shapes, and volumes. The methodology applied in this research has started with an optimised 3D over-segmentation operation to detect the boundaries and their feature distributions of the studied STV model, which tolerates the often confusing semantic differences between an event "actor" and its "background". This "background-independent" operation reduces the false-positive rate in differentiating signals from noises under complex real-life video settings.

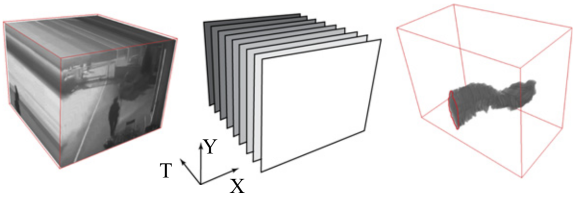

Fig. 1. A "falling down" video event defined by a STV shape.

After the feature extraction phase, a calibration mechanism is activated to measure the so-called normalised RI distances. This process evaluates the distribution of the segmented regions and then calibrates the matching distance by using a coefficient factor to record the linear corrections required. Compared with other conventional RI-based matching approaches, this extra step further improves the operational robustness at event recognition stage.

This paper will focus on two main contributions from the research to the human action recognition domain:

- An innovative segmentation algorithm for extracting voxel-level features has been developed based on a hybrid discontinuity-and-similarity-based segmentation model through combining Mean Shift (MS) clustering and the graph-based region description method.
- A novel extension of the Region Intersection technique has been realised for human action recognition by using a coefficient factor-boosted template matching method that is capable of superior performances under complex and real-life videoing conditions.

The rest of this paper is organised in the following order: Section 2 reviews the state-of-the-art for video event detection and STV-based feature processes. Section 3 introduces an innovative 3D STV segmentation strategy developed in this research. Section 4 focusses on the region-based template matching using the event shapes extracted from video inputs and a set of predefined single human action events. Section 5 highlights the adopted implementation strategies for system prototyping. The experimental designs are reported in Section 6. Section 7 concludes the research with discussions on results acquired and the planned future works.

2. Literature review

Video event detection researches encompass a wide spectrum of studies from basic DIP technologies and video processing to pattern analysis and even biological vision systems. After over 30 years of intensive research since its birth, progresses have been made on many fronts with extensive applications found in industry, such as traffic monitoring systems, CCTV-based security and surveillance networks, and robotic control. This section will focus on the prominent works to date on video event definition and STV-based template matching.

2.1. Action event definition

Generally speaking, a video event can be classified as single-human-based, such as gestures and postures, and