Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

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

A unified spatio-temporal human body region tracking approach to action recognition

Nouf Al Harbi *, Yoshihiko Gotoh

Department of Computer Science, University of Sheffield, United Kingdom

ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form
5 November 2014
Accepted 27 November 2014

Keywords:
Spatio-temporal segmentation
Human body volume
Object tracking
Regions of interest
Action recognition

ABSTRACT

There are numerous instances in which, in addition to the direct observation of a human body in motion, the characteristics of related objects can also contribute to the identification of human actions. The aim of the present paper is to address this issue and suggest a multi-feature method of determining human actions. This study addresses the matter by applying a sturdy region tracking method, instead of the conventional space–time interest point feature based techniques, demonstrating that region descriptors can be attained for the action classification task. A cutting-edge human detection method is applied to generate a model incorporating generic object foreground segments. These segments have been extended to include non-human objects which interact with a human in a video scene to capture the action semantically. Extracted segments are subsequently expressed using HOG/HOF descriptors in order to delineate their appearance and movement. The LLC coding is employed to optimise the codebook, the coding scheme projecting every one of the spatio-temporal descriptors into a local coordinate representation developed via max pooling. Human action classification tasks were used to assess the performance of this model. Experiments using KTH, UCF sports and Hollywood2 dataset show that this approach achieves the state-of-the-art performance.

1. Introduction

It is not a difficult task for us to comprehend the actions occurring in a video clip, regardless of the scene context, individuals in the scene or the camera angles with which the scene is presented. Furthermore, viewers can follow an extensive series of actions no matter how complex they are. From the computational point of view however, action representation poses considerable challenges. To provide a solution to this problem, most existing approaches are geared towards the expression of motion information within a scene. Descriptors for motion information are highly important; in recent years methods used to garner space–time interest point (STIP) features have been greatly improved [1]. It has been demonstrated that high-level models, which functioned on representations developed based on tracked objects, their features and/or interaction, were capable of identifying complex actions [2,3]. Such a high-level model, relying on interaction primitives, was proven to be highly effective when extracting appearance and STIP primitives.

STIPs are extricated from video data using the Bag of Words (BoW) model. They provide the basis for current action recognition research, which depends mostly on the ability that differentiates unique local space–time descriptors. STIP primitives are outlined on person and object trajectories, and attained via a flexible part-model detector [4]. However, despite their efficiency, such models remain incapable of tracking all object types or of functioning in a variety of observation conditions. It disregards information about the spatio-temporal organisation of the interest points which could be important for various computer vision tasks. This paper moves away from point-feature-based approaches, instead examines a spatio-temporal 'region-based' approach to interpret motion extracted from a video stream. In order to process complex actions which are challenging to track efficiently using conventional descriptors, this paper investigates a new model for action representation that relies on detecting spatio-temporal person–object interaction regions [5,6].

The argument brought forth is that video ought to be perceived as an assemblage of three-dimensional volumes. The integrated analysis of the temporal and spatial dimensions of a video presents a number of advantages. First of all, it facilitates the preservation of spatial and temporal consistency. Secondly, higher-level algorithms are able to concentrate on extensive, sparse regions rather than undertaking a multiple frame analysis of pixels, thus enhancing efficiency. Thirdly,

* Corresponding author.
E-mail addresses: nmalharbi1@sheffield.ac.uk (N. Al Harbi), y.gotoh@dcs.shef.ac.uk (Y. Gotoh).

http://dx.doi.org/10.1016/j.neucom.2014.11.072
0925-2312/

joint modelling of object appearance and movement leads to improved recognition results.

To this end the approach builds on a segmentation of human and non-human objects, where a human body volume is detected along a video stream [7]. These segments are extended to accommodate non-human objects to form up final key-segment regions. They formulate a descriptor that encompasses the static and dynamic features of detected key segments. The KTH, UCF sports and Hollywood2 dataset are employed to assess these representations [1]. It is demonstrated that, in comparison to conventional methods, action representation is substantially optimised, and that the codebook enhancement based on the locality constrained linear coding (LLC) technique [8] conveys the highest performance. The contributions of this work can be summarised as follows:

- Extraction of a spatio-temporal human body volume by incorporating generic object foreground segments guided by the state-of-the-art of human detection and segmentation approaches as well as extension of these regions to accommodate an interacted non-human objects regions.
- Extension of the existing two-dimensional (2D) image LLC scheme to a spatio-temporal video signal.
- Development of an efficient and robust schema to represent a human action signal.
- Application of the spatio-temporal region-based approach to the action classification task with Hollywood2, one of the most challenging real-world datasets, demonstrating that the approach outperforms the state-of-the-art, interest point-based techniques by a clear margin.

The idea of using objects to improve the recognition rate for actions was proposed by [9]. However that work used a traditional approach to detecting humans and objects [10], suffering from the large size of 'space' not belonging to a human body or an object. On the contrary, as illustrated in Fig. 1, this paper presents an approach that segments a human body and object regions at a frame level and tracks them over a sequence of video frames, thus creating a carefully trimmed spatio-temporal human body volume. Outcomes from the experiment indicate that the availability of an exact object region results in more accurate action representation.

2. Related work

Action representations incorporating low-level track point features in a video have been embraced by a large number of research works [11–15]. However these types of representations can incur tracking errors, particularly when there is background clutter present. On the other hand such representations circumvent the onerous task of object and person identification.

The use of representations for human motion to identify human actions has received a fair amount of attention. One of the first to investigate this phenomenon was by Bobick and Davis [16], who managed to capture view-dependent motion, as well as Yacoob and Black [17], who developed parameter-based motion models. For action recognition, Ali et al. [18] suggested the use of kinematic flow features. A different approach proposed by Efros et al. [19], and later by Zelnik-Manor and Irani [20], was the correlation-based categorisation of human motion. Schechtman and Irani [21] have implemented this approach to associate correspondences and self-similarities between images and videos.

Space–time interest point (STIP) methods have attracted an increasing amount of attention recently. By employing local STIPs, a number of studies have generated representations on the basis of visual vocabularies outlined with the help of gradient-based descriptors obtained either at determined points of interest [14,15,22–24] or from the actual point locations [11,25]. The positive implications of associating static and dynamic descriptors have been emphasised as well [1,12]. Compound neighbourhood-based features were originally developed for static images and object identification [26,27], but have been recently expanded to video processing [14,15,23]. A wide range of approaches are available, aiming to apply a coarse grid of histogram bins to subdivide the space–time volume globally [14,15,23] or else to position grids around the raw points of interest in order to generate new representations based on the location of the interest points which are included in the grid cells encompassing a central point [11].

A dataset derived from Hollywood films was presented by Laptev et al. [15]. Due to the fact that it consists of a wider variety of viewing angles, background clutter and scenarios, this dataset is considerably more difficult to process than the earlier datasets. This new dataset was referred to as 'Hollywood2: Human Actions and Scenes Dataset', and developed further by Marszalek et al. [1], who incorporated contextual information in their method. Recently another line of work was proposed by Bilen et al. [28] and assessed using Hollywood2 data. They attempted to describe actions by extracting salient local regions, applying motion segmentation, then tracking them with optical flow.

3. Human action representation

Despite relying on local information as well, the approach in this paper is different from existing works in that it focuses on a human body region-based feature representation. It is more concerned with the temporal continuity (or tracking) of regions than with isolated spatio-temporal regions. Fig. 2 illustrates the processing flow of the technique presented in this paper, which is later on referred to as the 'human body region tracking' approach with visual object recognition (or HBRT/VOC).

3.1. Detecting and tracking human body regions

Our goal is to segment human body volume in an unlabelled video. The approach consists of two main stages (Fig. 3). Firstly, human body objects are segmented at a frame level by combining low-level cues with a top-down part-based person detector, formulating grouped patches. Secondly, detected segments are propagated along the time line of video frames, exploiting the temporal consistency of detected foreground objects using colour

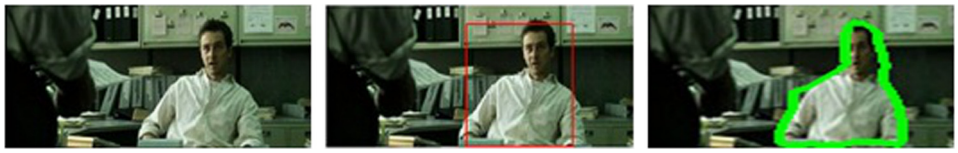

Fig. 1. A sample segmentation from the Hollywood2 dataset: the original frame from a video clip 'sceneclipautoautotrain00319' (left), a segmentation using Felzenszwalb et al. [10] (middle) and a segmentation using the approach presented in this paper (right).

Please cite this article as: N. Al Harbi, Y. Gotoh, A unified spatio-temporal human body region tracking approach to action recognition, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.072