Egyptian Informatics Journal (2015) **16**, 187–198

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Human action recognition using trajectory-based representation

Haiam A. Abdul-Azim [a,*], **Elsayed E. Hemayed** [b]

[a] *Physics Department, Faculty of Women for Arts, Science and Education, Ain Shams University, Egypt*
[b] *Computer Engineering Department, Faculty of Engineering, Cairo University, Egypt*

Received 16 January 2015; accepted 27 May 2015

KEYWORDS

Human action recognition;
Spatio-temporal features;
Cuboid detector;
Trajectory-based feature description;
Bag-of-Words

Abstract Recognizing human actions in video sequences has been a challenging problem in the last few years due to its real-world applications. A lot of action representation approaches have been proposed to improve the action recognition performance. Despite the popularity of local features-based approaches together with "Bag-of-Words" model for action representation, it fails to capture adequate spatial or temporal relationships. In an attempt to overcome this problem, a trajectory-based local representation approaches have been proposed to capture the temporal information. This paper introduces an improvement of trajectory-based human action recognition approaches to capture discriminative temporal relationships. In our approach, we extract trajectories by tracking the detected spatio-temporal interest points named "cuboid features" with matching its SIFT descriptors over the consecutive frames. We, also, propose a linking and exploring method to obtain efficient trajectories for motion representation in realistic conditions. Then the volumes around the trajectories' points are described to represent human actions based on the Bag-of-Words (BOW) model. Finally, a support vector machine is used to classify human actions. The effectiveness of the proposed approach was evaluated on three popular datasets (KTH, Weizmann and UCF sports). Experimental results showed that the proposed approach yields considerable performance improvement over the state-of-the-art approaches.

1. Introduction

Over the past years, human action recognition in videos has been a growing field of research in computer vision with many real-world applications, such as video surveillance, video indexing/browsing, recognizing gestures, human–computer interfacing and analysis of sport-events. However, it is still a challenging problem because of cluttered backgrounds, illumination changes, different physiques of humans, variety of

* Corresponding author. Tel.: +20 1272588373; fax: +20 224157804.
E-mail addresses: haiamadel@yahoo.com (H.A. Abdul-Azim), hemayed@ieee.org (E.E. Hemayed).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.05.002

clothing, camera motion, partial occlusions, viewpoint changes, scale variation of video screen, etc.

Generally, human action recognition procedure consists of two main steps: action representation, and action learning and classification. Existing action recognition approaches are classified by Weinland et al. [1] based on action representation into two main approaches: global and local representations. Global representation approaches focus on detecting the whole body of the person by using background subtraction or tracking. Silhouettes, contours or optical flow are usually used for representing the localized person. These representations are more sensitive to viewpoint changes, personal appearance variations and partial occlusions.

In local representation approaches, videos are represented as a collection of small independent patches. These patches involve the regions of high variations in spatial and time domains. Centers of the patches are called spatio-temporal interest points (STIPs). The detected points are described by capturing the appearance and/or motion information from their patches and clustered to form a dictionary of prototypes or visual-words. Each action sequence is then represented by Bag of Words model (BOW) [2]. Recently, these approaches have become very successful approaches for human action recognition. They overcome some limitations of global representation such as sensitivity to noise and partial occlusion and the necessity of accurate localization by background subtraction and tracking.

Several STIPs detectors have been proposed to determine the spatio-temporal interest locations in videos. For example, Laptev [3] extended Harris corner detector for the spatio-temporal case and propose Harris3D detector, Dollar et al. [4] proposed the Cuboid detector by applying 1-D Gabor filters temporally, Willems et al. [5] proposed Hessian detector which measures the saliency with the determinant of the 3D Hessian matrix, and Wang et al. [6] introduced Dense sampling detector that extracts STIPs at regular positions and scales in space and time. Also, various descriptors for STIPs have been proposed such as Gradient descriptor [4], Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors [2], 3D Scale-Invariant Feature Transform (3D SIFT) [7], 3D Gradients descriptor (HOG3D) [8] and the extended Speeded Up Robust Features descriptor (ESURF) [5].

Despite the popularity of local representation approaches, it has some drawbacks. One of the main drawbacks is the ignorance of spatial and temporal relationships between local features. This may be a major problem in human action recognition. The spatial and/or temporal connections between the detected low-level action parts are necessary to introduce intrinsic characteristics of actions. A lot of attempts have been made to weaken this limitation of the local representation approaches based on BOW model. To capture these relationships early work was introduced such as Laptev et al. [2], Liu and Shah [9], Gilbert et al. [10], Zhang et al. [11] and Bregonzio et al. [12].

This paper introduces an enhancement of the recently proposed approaches which called trajectory-based local representation approaches [13–19]. These approaches capture some temporal relationships between the detected interest points by tracking them throughout the video. They differ in the trajectory generation and representation methods used. In this framework, we track the STIPs detected by Cuboid detector using SIFT-matching, then after some refinement we use the tracked points to form the action trajectories and then we describe the volumes around these trajectories' points. These features are represented with a Bag-of-Words (BOW) model. Finally, human actions are classified using a Support Vector Machine. In order to evaluate the proposed approach, we train and recognize action models on three popular datasets, KTH [20], Weizmann [21] and UCF Sports [22].

This paper is organized as follows. Section 2 reviews the previous related work. Section 3 describes the proposed trajectory-based approach for video representation. Section 4 presents the experimental setup, datasets and discusses the obtained results. Finally, Section 5 concludes the paper.

2. Related work

Recent works [13–19] show good results for action recognition in which the local spatio-temporal volumes are determined by using the trajectories of the interest points through video sequences. These trajectory-based approaches leverage the motion information extracted from the spatio-temporal volumes and utilize different methods for representation. Messing et al. [13] allowed a bag-of-features model to capture a significant amount of non-local structure by tracking Harris3D interest points [3] with the Pyramid Lucas–Kanad e–Tomasi (KLT) tracker [23]. For action classification, Trajectories are represented by computing quantized velocities over time which called "velocity history". Matikainen et al. [14] introduced a trajectory-based motion features which called "trajectons". Trajectories are produced using a standard KLT tracker. For trajectories representation, clustering trajectories is performed using K-means and then an affine transformation matrix is computed for each cluster center. Sun et al. [15] generated their trajectories by matching SIFT descriptors between consecutive frames based on a unique-match constraint that yields good motion trajectories. Actions are then described in a hierarchical way where three levels of context information are exploited. Sun et al. [16] also extracted long-duration trajectories by combining both KLT tracker and SIFT descriptor matching. Moreover, random points are sampled for tracking within the region of existing trajectories to capture more salient image structures. For action representation, spatio-temporal statistics of the trajectories are used. Raptis and Soatto [17] proposed spatio-temporal feature descriptors that capture the local structure of the image around trajectories. These descriptors are a computation of HOG or HOF descriptors along the trajectories. The final descriptor is applied in action modeling and video analysis. Bregonzio et al. [18] presented an action representation based on fusing the trajectories generated by both KLT tracker and SIFT matching with the extracted spatio-temporal local features. This fusion enhanced the trajectory-based action representation approaches to be able to recognize actions under realistic conditions such as small camera movement, camera zooming, and shadows. Wang et al. [19] introduced dense-trajectories and used the motion boundary histograms (MBH) [24] as a trajectory descriptor. Points are detected by dense sampling detector and checked by Shi and Tomasi criterion [25] then tracked using a dense optical flow field. Trajectories are described with four different descriptors (trajectory shape, HOG, HOF and MBH).