Contents lists available at ScienceDirect

Pattern Recognition

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

Trajectory aligned features for first person action recognition

Suriya Singh [a,*], Chetan Arora [b] , C.V. Jawahar [a]

[a] CVIT, IIIT Hyderabad, India
[b] IIIT Delhi, India

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form
17 April 2016
Accepted 23 July 2016
Available online 26 August 2016

Keywords:
Action and activity recognition
Egocentric vision
Video indexing and analysis
Video segmentation

ABSTRACT

Egocentric videos are characterized by their ability to have the first person view. With the popularity of Google Glass and GoPro, use of egocentric videos is on the rise. With the substantial increase in the number of egocentric videos, the value and utility of recognizing actions of the wearer in such videos has also thus increased. Unstructured movement of the camera due to natural head motion of the wearer causes sharp changes in the visual field of the egocentric camera causing many standard third person action recognition techniques to perform poorly on such videos. Objects present in the scene and hand gestures of the wearer are the most important cues for first person action recognition but are difficult to segment and recognize in an egocentric video. We propose a novel representation of the first person actions derived from feature trajectories. The features are simple to compute using standard point tracking and do not assume segmentation of hand/objects or recognizing object or hand pose unlike in many previous approaches. We train a bag of words classifier with the proposed features and report a performance improvement of more than 11% on publicly available datasets. Although not designed for the particular case, we show that our technique can also recognize wearer's actions when hands or objects are not visible.

1. Introduction

Advances in camera sensors and other related technologies have led to the rise of wearable cameras which are comfortable to use. In the past few years, the use of Google glass [1] and GoPro [2] has become increasingly popular. Such cameras are typically worn on the head or along with the eyeglasses and have the advantage of capturing from a similar point of view as that of the person wearing the camera. We refer to such cameras with first person view as egocentric cameras.

Excitement of sharing one's actions with friends and the community have made egocentric cameras like GoPro a de facto standard in extreme sports. Egocentric cameras can be used to capture visual logs for law enforcement officers leading to a significant decrease in complaints against the officers [3]. Daily logs from egocentric cameras are also useful in a video sharing application or simply as a memory aid for the wearer. For the visually challenged, researchers are trying to augment egocentric videos with meta data such as facial identity, place, text, etc. [4]. Even for people with regular vision, the promise of giving context aware suggestions is compelling. In spite of their popularity, egocentric videos can be difficult to watch from start to end because of the constant and extreme shake present due to natural motion of wearer's head.

Our focus in this paper is on recognizing wearer's actions from an egocentric video. Owing to their shakiness, egocentric videos are significantly more challenging to analyze than third person videos. Action recognition gives structure to such ‘wild’ videos which can then be used to search, index or browse. Action recognition is also usually a first step in many other egocentric applications, for example, video summarization, augmented reality, real time suggestions, etc. We follow the popular notation in the field to differentiate between ‘activity’ and ‘actions’. *Activity* is a high level description of what a person is doing at a particular point of time. An activity is usually composed of many short *actions*, which are perceptually closer to the gestures performed by the person. For example, while making tea is an activity, picking the jar, opening the lid and taking sugar are the actions. Other types of actions popular in computer vision are sitting, standing, jumping, etc.

Egocentric videos are different from their third person counterparts, not only because of the change in camera perspective but also because of change in camera motion profile. Many of the accepted techniques for third person video analysis do not work as is for egocentric videos, and the community has been trying to adapt or develop from scratch solutions to these problems in the new context. Works done in the last few years have ranged from

[*] Corresponding author.
E-mail addresses: suriya.singh@research.iiit.ac.in (S. Singh), chetan@iiitd.ac.in (C. Arora), jawahar@iiit.ac.in (C.V. Jawahar).

http://dx.doi.org/10.1016/j.patcog.2016.07.031
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

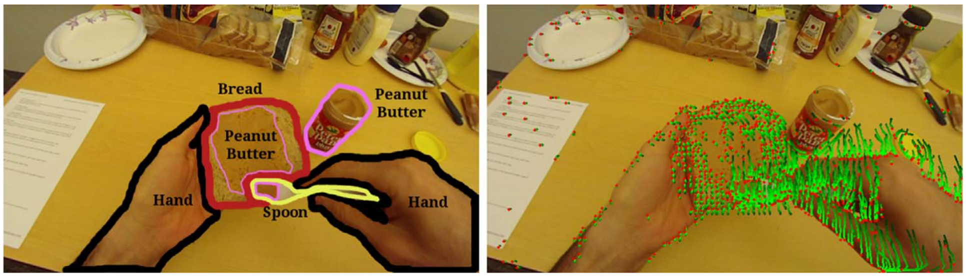

Fig. 1. The focus of this paper is on recognizing wearer's actions from egocentric videos. Earlier work in this area has suggested complicated image segmentation followed by hand or object recognition (left image). We observe that salient objects (hands or handled objects) in such actions are also the objects moving dominantly with respect to the background and can be captured easily using trajectory aligned features (right image) without any prior image segmentation or hand or object recognition. The example images shown here are from GTEA database [5].

tackling simpler problems like object recognition [5–7] and activity recognition [8–15], to more complex problems like video summarization [16–18], and understanding social interactions [19]. Interesting ideas which exploit special properties of egocentric videos have also been proposed for problems like temporal segmentation [20,21], frame sampling [22,23], hyperlapse [24], gaze detection [25] and camera wearer identification [26,27].

Wearer's action recognition from egocentric video is harder compared to regular third person action recognition due to associated unstructured and wild motion of the camera caused by wearer's natural head movement. Different speeds of performing actions and widely varying operating environment also cause difficulties. Fig. 2 gives some examples of the actions we are interested in recognizing.

Given the unique perspective of the egocentric camera, which makes unavailable, the view of the actor or his/her pose, standard action recognition techniques from third person actions are not applicable as is. Also quickly changing view field in typical egocentric videos makes it hard to develop models from foreground or background objects. Therefore, the techniques developed for wearer's action recognition have so far remained independent of work done in third person actions. The earliest work in wearer's action recognition used global features (GIST) for the task [11]. Later works focussed on objects present in the scene for recognition [12,28]. Position and pose of hand are important cues for action recognition involving object handling and have been explored by the researchers as well [5]. In action categories which do not involve any handled object, researchers have typically exploited the optical flow observed in the video, which for an egocentric video is indicative of head motion and is highly correlated with the kind of action being performed by the wearer [21,20]. Eye-motion and ego-motion have also been used to recognize indoor desktop actions [14].

Object or hand pose is an important cue for wearer's action but detecting them in an egocentric video is a difficult task and the dependence of the action recognition on such explicit detection/recognition affects the overall action recognition accuracy, besides making the system more complex and inefficient. We show in this paper that such prior information is not necessary. We observe that in any egocentric action scenario involving handled objects, the dominantly moving objects in the scene are typically hands and handled objects only (Fig. 1). Optical flow observed for the background is due to motion of the wearer's head. Such motion causes three dimensional rotation of the camera and can be easily compensated by cancelling frame to frame homography. This leads to a simple algorithm for extraction of hands and objects. We further show that complicated models of hand pose or object recognition are not necessary for the action recognition task, and instead, simple trajectory based features, combining motion profile and the visual features around these trajectories alone are

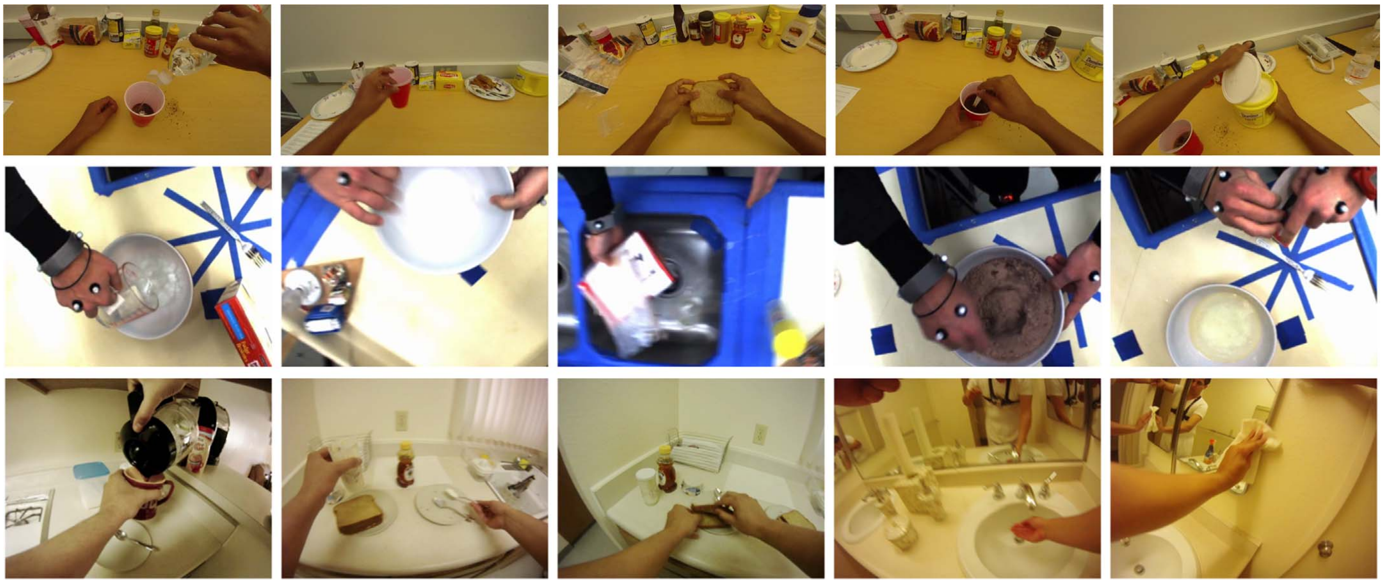

Fig. 2. Examples of wearer's action categories we propose to recognize in this paper from different datasets: GTEA [5] (top row), Kitchen [11] (middle row) and ADL [12] (bottom row). First, second and third columns across all rows are 'pour', 'take' and 'put' actions respectively. Fourth and fifth columns are 'stir' and 'open' actions for top and middle rows, and 'wash' and 'wipe' actions for bottom row. The actions vary widely across datasets in terms of appearance and speed of action. Features and technique we suggest in this paper is able to successfully recognize wearer's actions across different presented scenarios, showing robustness of our method.