Computer Vision and Image Understanding 149 (2016) 229–248

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Robust multi-dimensional motion features for first-person vision activity recognition

Girmaw Abebe[a,b,*], Andrea Cavallaro[b], Xavier Parra[a]

[a] CETpD, UPC-BarcelonaTech, Rambla de l'Exposicio, Vilanova i la Geltru, Spain
[b] Centre for Intelligent Sensing, Queen Mary University of London, London, UK

ARTICLE INFO

Article history:
Received 17 April 2015
Accepted 23 October 2015

Keywords:
Human activity recognition
First-person vision
Grid optical flow
Inertial data
Wearable camera

ABSTRACT

We propose robust multi-dimensional motion features for human activity recognition from first-person videos. The proposed features encode information about motion magnitude, direction and variation, and combine them with virtual inertial data generated from the video itself. The use of grid flow representation, per-frame normalization and temporal feature accumulation enhances the robustness of our new representation. Results on multiple datasets demonstrate that the proposed feature representation outperforms existing motion features, and importantly it does so independently of the classifier. Moreover, the proposed multi-dimensional motion features are general enough to make them suitable for vision tasks beyond those related to wearable cameras.

1. Introduction

Advances in wearable technologies are facilitating the understanding of human activities using first-person vision (FPV) for a wide range of assistive applications [1,2]. Application domains that employ wearable cameras (Fig. 1) include life-logging and video summarization [3–7], activity recognition [8–21], and eye-tracking and gaze detection [22–25]. Human activities can be categorized as ambulatory (e.g., walk) [8–15]; person-to-object interactions (e.g., cook) [16–19]; and person-to-person interactions (e.g., handshake) [20,21]. In particular, the recognition of ambulatory activities [26] involving a full-body motion (Fig. 2) is of interest in a range of tasks from (self-) monitoring of the elderly to performance analysis of athletes.

Ambulatory activity recognition systems can be modeled as a cascade of three main blocks, namely data acquisition and preprocessing, motion estimation and feature extraction, and classification (Fig. 3). Wearable cameras are often employed jointly with other sensors, more commonly with inertial sensors [8–10], in order to leverage the merits of the latter. However, using multiple wearable sensors results in obtrusiveness of the system, complexity of the preprocessing stage (e.g., need for synchronization), and higher computational cost for feature extraction. The main contribution of this work is on the extraction of a robust feature vector from motion data only of a first-person video, while providing the type of information that is usually generated by the combination of a wearable camera with inertial sensors.

In this paper, we propose a robust motion-feature (RMF) that combines grid optical flow-based features (GOFF) and video-based inertial features (VIF). We concatenate features extracted from discriminative motion patterns in the optical flow data such as magnitude, direction and frequency; and also include features extracted from virtual inertial data derived from the movement of intensity centroid across frames in a video without physically using inertial sensors. Intensity centroid [27] is analogue to a center of mass in physics where a rigid body experiences a zero-sum of weighted relative location of its distributed mass. The centroid is computed from weighted averages of intensity values (image moments [28,29]). The proposed RMF is generic and can be employed with any classifier. In particular, for validation we use support vector machines (SVM) and k-nearest neighborhood (KNN) to test the flexibility of the proposed RMF and compare it with three state-of-the-art motion features, experimented across different activities and environments on four different datasets. The first dataset is used to experiment indoor ambulatory recognition (IAR) task of eight activities and the second is related to basketball activity recognition (BAR) of eleven activities recorded in an outdoor court. IAR and BAR datasets are recorded by ourselves; and to the best of our knowledge, BAR dataset is the first of its type[1]. In addition to IAR and BAR, we also validate the experiments

* Corresponding author at: Centre for Intelligent Sensing, Queen Mary University of London, London, UK.
E-mail addresses: g.abebe@qmul.ac.uk, girmaw.abebe@upc.edu (G. Abebe), a.cavallaro@qmul.ac.uk (A. Cavallaro), xavier.parra@upc.edu (X. Parra).

[1] The datasets and the annotation are available at http://www.eecs.qmul.ac.uk/~andrea/FPV.html.

http://dx.doi.org/10.1016/j.cviu.2015.10.015

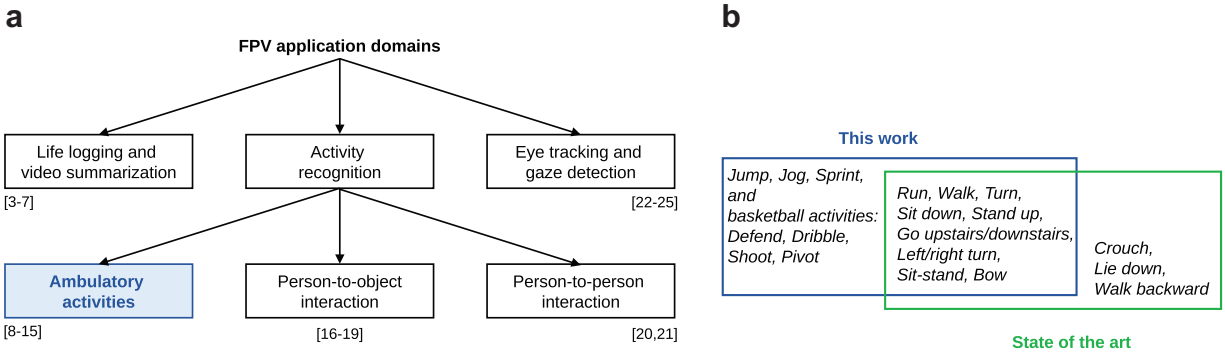

Fig. 1. The focus of the proposed work is ambulatory activities involving the whole body. (a) Classification of existing First-Person Video (FPV) application domains. (b) Comparison of the activities covered in the proposed work and in the state of the art.

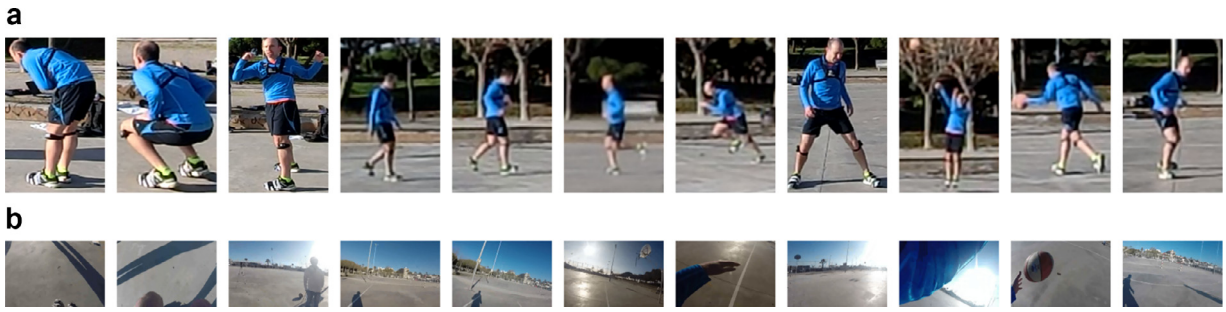

Fig. 2. Sample ambulatory activities considered in this work: (a) activities viewed from an external camera; (b) frames from the first-person vision acquired by a wearable camera while a user performs the corresponding activity in the top row. The activities from left to right are *Bow*, *Sit-Stand*, *Left-right turn*, *Walk*, *Jog*, *Run*, *Sprint*, *Pivot*, *Shoot*, *Dribble* and *Defend*.

on two more publicly available datasets: JPL-interaction dataset [20] of seven activities and DogCentric [30] dataset of ten activities.

The remaining of the paper is organized as follows. Section 2 reviews the related work. Section 3 formulates the problem and presents the proposed method along with the analysis of parameter settings and computation time. In Section 4, we describe the details of the datasets, the experimental set-up, and the baseline method developed as a reference for comparisons. Section 5 focuses on the results of experiments and discusses significant findings, and Section 6 concludes the paper.

2. Related works

Ambulatory activities such as *Walk*, *Turn*, *Run*, *Sit*, *Stand*, *Go upstairs*, *Go downstairs* and *Left-right turn* involve full-body motions. Therefore, motion in FPV of an ambulatory activity is generally dominated by a global motion on which discriminant features are extracted. Existing motion-features use either raw grid optical flow [8,11] or limited directional and/or magnitude information [12–14]. Motion patterns of activities can vary in their magnitude, direction and frequency characteristics [14]. For example, on the one hand, *Walk* and *Run* have similar direction but different magnitudes and frequency patterns, on the other hand, *Sit-down* and *Stand-up* possess similar motion magnitudes but in opposite directions. Generally, existing works employ either interest point-based [12,13] or optical flow-based [8–11,14] methods in order to estimate motion and then extract features.

Interest point-based methods involve the detection, description and matching of interest points on subsequent frame pairs [31–33]. Detection refers to the localization of key-points in the image (e.g., corners), whereas a descriptor represents the neighborhood of a key-point with invariant characteristics (e.g., SURF [34]); then the matching of descriptors is performed on each subsequent pair of frames. Matched descriptors are further refined (e.g., smoothing and outliers rejection) to achieve precise motion estimation. Zhang et al. [13] employed Shi and Tomasi [35] features in order to recognize *Sitting*, *Walking*, *Bowing*, *Crouching* and *Left-right turning* activities using a chest-mounted camera and a SVM classifier. The work was later extended to include the following: a multi-scale detection of interest points, *Sitting-up* activity, and KNN and Naive Bayes (NB) classifiers [12]. Motion was computed as pixel-wise displacement between two matched key-points. The displacement was computed as the difference of the key-points' locations in the corresponding frames. Then outliers were rejected using Random Sample Consensus [36], and discarding small motion vectors. Histogram computation on motion-direction resulted in low-dimensional motion representation. The final motion-feature was built from the sum of direction histograms in a video segment. Average standard deviation [13] and combined standard deviation [12] of direction histogram were utilized to reflect temporal variation. However, interest point-based methods generally fail when there is not enough texture to detect interest points, or when the activities (e.g., *Dribble*) involve complex ego-motion, motion blur and parallax. Moreover, these features are not appropriate to discriminate activities such as *Jog* and *Run* as they do not include specific motion characteristics other than direction (e.g., magnitude) [12,13].

Optical flow-based methods (OFM) use direct motion estimation [37]. Direct methods, also known as appearance-based methods [32], do not involve the detection, description and matching procedures used by interest point-based methods. Direct methods can achieve sub-pixel accuracy and determination of global motion in the presence of multiple local motions and motion parallax [32,33,37,38]. When an ambulatory activity is dominated by a global motion, in absence of major occlusions, the use of optical flow vector