



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'Exposició, 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.

© 2015 The Authors. Published by Elsevier Inc.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

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 type¹. 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>.

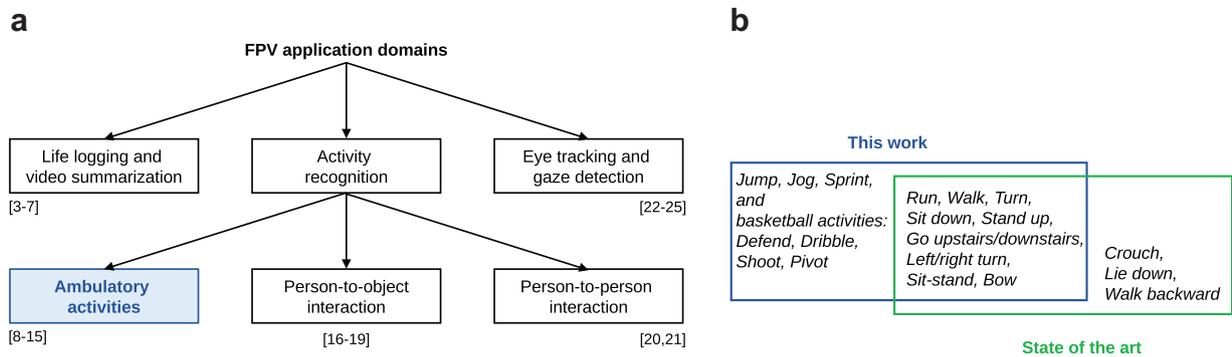


Fig. 1. The focus of the proposed work is ambulatory activities involving the whole body. (a) Classification of existing First-Person Video (FPV) application domains. (b) Comparison of the activities covered in the proposed work and in the state of the art.

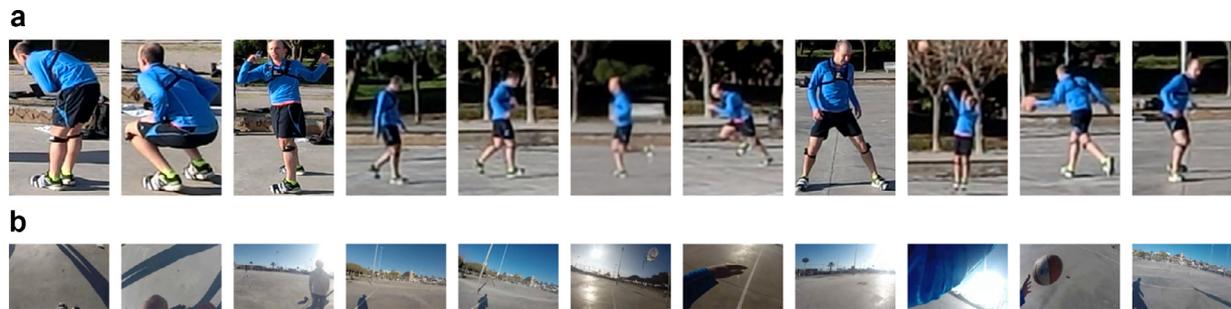


Fig. 2. Sample ambulatory activities considered in this work: (a) activities viewed from an external camera; (b) frames from the first-person vision acquired by a wearable camera while a user performs the corresponding activity in the top row. The activities from left to right are Bow, Sit-Stand, Left-right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

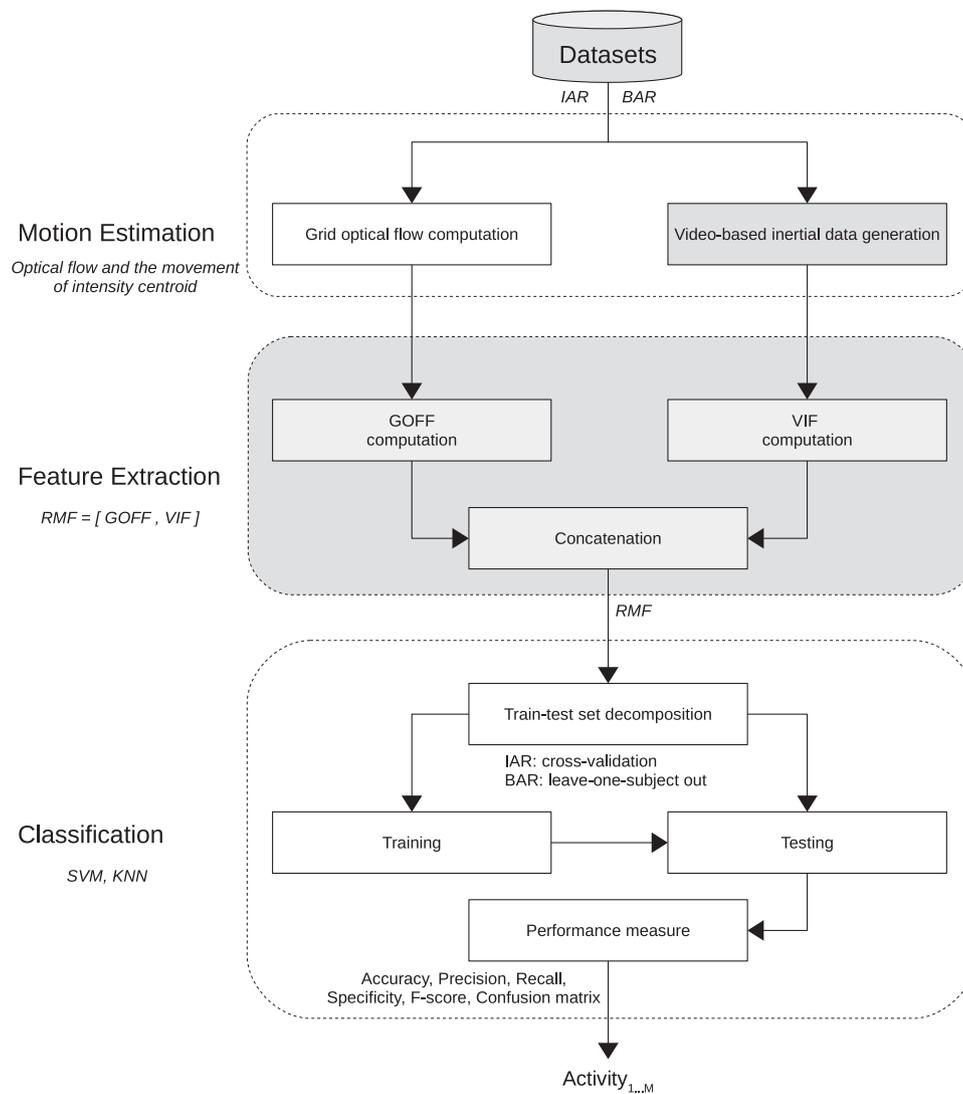


Fig. 3. The overview of the proposed ambulatory recognition system in which highlighted blocks show our contributions. IAR: indoor ambulatory recognition dataset; BAR: basketball ambulatory recognition dataset; GOFF: grid optical flow-based features; VIF: vision-based inertial features; RMF: robust motion feature; SVM: support vector machines; KNN: k-nearest neighborhood.

at each pixel (dense optical flow) results in redundant motion information. For this reason, a grid representation is usually preferred [8–11,14], on which histogram computation can be further applied. Kitani et al. [14] encoded magnitude, direction and periodical motion components with a joint space of three magnitude, four direction and three magnitude variance bins concatenated with sixteen histogram bins of the frequency component. Dirichlet process mixture modeling was applied to learn motion histogram codebook for unsupervised segmentation of ego-actions in sport videos.

Ambulatory activity recognition systems that employed OFM use visual data alone [11] or along with accelerometer data (*multi-modal methods*) [8–10]. Zhan et al. [11] concatenated the horizontal and vertical components of grid optical flow to represent the motion feature. SVM outperformed LogitBoost and KNN classifiers that were implemented independently to validate the classification of *Walk, Going upstairs, Going downstairs and Drink* activities. The performance was enhanced by the local smoothing of grid flow (average pooling) and a hidden Markov Model (HMM) based structural learning. The work was later extended in [8,9] in order to classify a larger set of activities by combining visual features from grid optical flow and inertial features from 3-axis accelerometer data. Zhan et al. [8,9] and Nam

et al. [10] leveraged vision and inertial-based classifications independently; then structural learning was applied using a multi-scale conditional random field (CRF) in [8,9] while knowledge-driven algorithm determined the final class label in [10,15]. Zhan et al. [8,9] adopted the visual features from [11], and the state-of-the-art time and frequency-domain features were combined to obtain the inertial features. The set of activities in [9] were extended from [11] to include *Stand-up, Sit-down, Sitting, Reading, Watching TV/monitor, Writing, Switch water-tap and Washing*. More activities, namely *Standing, Lying, Transfer, Open door, Lie-down and Sit-up* were also considered in [8]. Both camera and inertial sensors were mounted together on the forehead using safety goggles. The activities studied in [10,15] were *Walk forward, Walk backward, Turn, Run, Sitting, Standing, Going upstairs, Going downstairs and Taking elevator* using a waist-mounted embedded module of the two sensors. The magnitude and direction components of optical flow were partially used as knowledge-based information to discriminate activities. Grid optical flow vectors were the visual features while correlation between axes and energy constituted the acceleration features.

Existing works in *inertial sensor-based* ambulatory activity recognition involve one [39] or multiple [40] 2-axis [41] or 3-axis accelerometers [42], often along with a gyroscope [43], mounted on

Table 1

Summary of the state of the art in ambulatory activity recognition. M.P.: mounting position; Fu.: fusion with accelerometer sensor; Eenvt.: environment; ID: indoor; OD: outdoor; IO: indoor-outdoor; FPS: frame rate; N/A: not available; # Sub.: number of subjects; MEM: motion estimation method; IPM: interest point-based method; OFM: optical-flow based method; AMD: accumulated motion distribution; MRGF: multi-resolution good features; AP: average pooling; MBH: motion-based histograms; RMF: robust motion features; Dim.: feature dimension; Smo.: smoothing method; Accu.: accumulation; Ave.: averaging; Gau.: Gaussian filtering; Ra.: raw grid optical flow; Ma.: magnitude component; Di.: direction component; Fr.: frequency component; NB: naive Bayes; HMM: hidden Markov model; CRF: conditional random field; DPM: Dirichlet process mixture model; I = {Walk forward, turn, run, sitting, standing, going upstairs, going downstairs}; II = {Sitting, walking, bowing, crouching, left-right turning}; III = {Walk, going upstairs, going downstairs, drink}; IV = {stand-up, sit-down, sitting}; V = {reading, watching TV/monitor, writing, switch water-tap, hand washing}; VI = {standing, lying, transfer, open door, lie-down, write, sit-up}; VII = {Bow, defend, dribble, jog, left-right turning, pivot, run, shoot, sit-stand, sprint, walk}; Sport refers to ego-actions in two choreographed and six real world youtube sport videos.

Ref.	Dataset			Feature extraction method												
	M.P.	Fu.	Activities	Envt.	FPS	Resolution	# Sub.	MEM	Feature	Dim.	Smo.	Ra.	Ma.	Di.	Fr.	Classifier
[10,15]	Waist	✓	I+ walk backward + taking elevator	IO	N/A	640 × 480	N/A	OFM	N/A	N/A	N/A	✓	✓			SVM
[13]	Chest	x	II	IO	10	320 × 240	N/A	IPM	AMD	9	Accu.			✓		SVM
[12]	Chest	x	II + sitting-still	IO	10	320 × 240	1	IPM	MRGF	9	Accu.			✓		NB, KNN, SVM
[11]	Eyeglass	x	III	IO	25	640 × 480	N/A	OFM	AP	48	Ave.	✓				(KNN, LogitBoost, SVM)+HMM
[9]	Fore-head	✓	III + IV + V	IO	15	144 × 676	5	OFM	AP	N/A	Ave.	✓				(LogitBoost, SVM)+CRF
[8]	Fore-head	✓	III + IV + VI	IO	15	144 × 676	30	OFM	AP	N/A	Ave.	✓				(LogitBoost, SVM)+CRF
[14]	Head	x	Sport	OD	N/A	N/A	1	OFM	MBH	52	N/A		✓	✓	✓	DPM
Proposed	Chest	x	I + jump VII	ID	60	1080 × 1920	1	OFM	RMF	243	Gau.+Accu.	✓	✓	✓	✓	KNN, SVM
				OD	30	720 × 1280	4	OFM	RMF	243	Gau.+Accu.	✓	✓	✓	✓	KNN, SVM

waist [44], chest [45] or specific movement-sensitive body parts (e.g., arm [46] and ankle [47]). Both time and frequency-domain features are used to quantify discriminative characteristics in the inertial data. Examples of time-domain features include mean, standard deviation, variance, zero-crossing rate, energy, minimum, maximum, kurtosis and correlation between axes. Frequency-domain features are extracted from the Fourier response of the inertial data [26,48,49], and feature reduction techniques (e.g., principal component analysis [50]) are also commonly applied. As for classifiers, SVM is dominantly employed, followed by KNN and decision trees [39,51]. Threshold-based classifiers are also reported achieving competitive performance with high classification speed [42,49].

In summary, while optical flow-based methods are frequently employed in the state of the art due to their sub-pixel accuracy and flexibility to work under different motion models [37], the majority of existing works in this category do not exploit key motion characteristics, such as magnitude that helps to discriminate activities with similar direction patterns (e.g., *Jog*, *Run* and *Sprint*). Unlike to [8–11], which are only based on optical flow data, our method exploits magnitude, direction and frequency (periodicity) characteristics more effectively. The method proposed by Kitani et al. [14] is the closest to the proposed one but significantly differs for the following reasons: 1) less emphasis was given to motion direction that was encoded with few bins together with flow magnitude and its variance bins in a joint space, and 2) frequency features were extracted from amplitude values while we did it from both magnitude and direction components. In addition to this, we generate (virtual) inertial data from a video and extract inertial features similar to [8,9] in order to enhance the recognition performance of the optical flow-based system. A summary of the state-of-the-art and proposed motion feature extraction methods for ambulatory activity recognition from first-person vision is shown in Table 1.

3. Proposed method

Let $D = \{V_i\}_{i=1}^N$ be a dataset of N video segments captured from FPV. Each V_i is recorded using a fixed resolution and frame rate, and contains F_i frames $V_i = \{f_{ki}\}_{k=1}^{F_i}$. The number of frames might vary among video segments. Let $\mathcal{C} = \{A_j\}_{j=1}^M$ denote M different activities, while each V_i contains only one activity A_j . The aim is to extract robust motion-feature (RMF) in order to classify a sample in V_i into its corresponding activity class $A_j \in \mathcal{C}$. While allowing local motions due

to occlusions, we assume that a global motion is dominant over the majority of the frames in V_i .

3.1. Feature extraction

The types of extracted features are motivated by the nature of variations among ambulatory activities (see Figs. 4 and 5). Activities such as *Sit-down* and *Stand-up* vary in their direction components (Fig. 4c) while they possess similar magnitude values (Fig. 4a). Activities such as *Sprint* and *Walk* have similar direction information (Fig. 4d) but significantly different in their magnitude patterns (Fig. 4b). In addition to this, a spectrogram of motion direction in Fig. 5 shows that discriminative features can also be extracted in the frequency domain. We develop RMF by extracting features from the optical flow and inertial data of a video (Fig. 6).

3.1.1. Grid optical flow-based features (GOFF)

The proposed motion-features exploit optical flow data more effectively than existing optical flow-based features [8,10,11,14]. In order to encode the variation in motion magnitude, direction and dynamics among activities, we extract a set of feature subgroups, namely Motion Magnitude Histogram Feature (MMHF), Motion Direction Histogram Feature (MDHF), Motion Direction Histogram Standard-deviation Feature (MDHSF), Fourier Transform of Motion direction Across Frame (FTMAF), and Fourier Transform of grid Motion Per Frame (FTMPF).

Given a video segment $V_i = \{f_{ki}\}_{k=1}^{F_i}$ where each frame $f_{ki} \in \mathbb{R}^{R \times C}$ has a height of R pixels and a width of C pixels, we compute the Horn-Schunk optical flow [52] for each subsequent pair of frames. We select the Horn-Schunk method, rather than the Lucas-Kanade approach [53], because of its global smoothness assumption which is preferred in our scenario where a global motion is assumed to be dominant and reflects the ego-motion of a user wearing the camera. Because a dense optical flow representation of a frame $[E_k]_{R \times C}$ contains redundancy of motion information under the assumption of a dominant global motion, we apply a grid representation $[B_k]_{G^2 \times 1}$, where G refers to the number of grids in each dimension (see Section 3.2 for the analysis part). We build the grid representation as $B_k = E_k(r_A, c_A)$, where r_A and c_A are G -dimensional row and column vectors sampled as $r_A = (1, 1 + R/G, 1 + 2R/G, \dots, R)$ and $c_A = (1, 1 + C/G, 1 + 2C/G, \dots, C)$. The sampling in r_A and c_A is conducted periodically after every R/G and C/G pixels, respectively, so

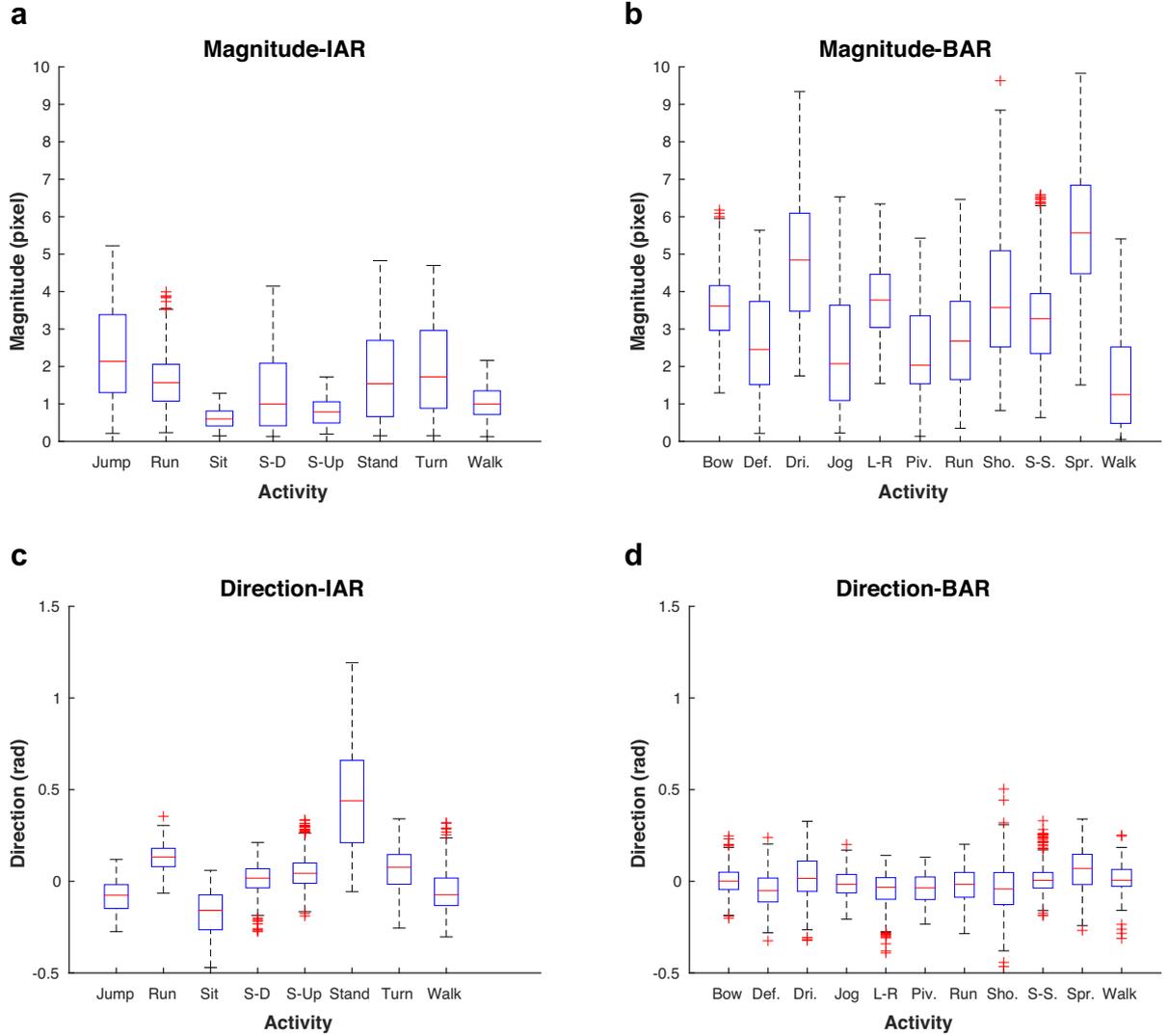


Fig. 4. Average magnitude and direction values for activities in (a,c) IAR and (b, d) BAR datasets. Comparatively, average magnitude of BAR activities is higher and more variant (3.35 ± 1.16 pixel) than that of IAR (1.42 ± 0.58 pixel). From direction point of view, IAR activities show more higher average variation (0.05 ± 0.19 rad) while BAR directions are restricted in -0.01 ± 0.03 rad. S-D: Stair-down; S-Up: Stair-up; Def.: Defend; Dri.: Dribble; L-R: Left-right turn; Piv.: Pivot; Sho.: Shoot; S-S: Sit-stand and Spr.: Sprint.

that B_k contains sample motions from all regions in a frame. Grid-optical flow representation of a frame includes horizontal and vertical components, $B_k = (B_{k_x}^g + jB_{k_y}^g)G^2$. By applying a grid representation, we reduce the dimension of motion-vector from $R \times C$ in E_k to G^2 in B_k .

We consider the grid motion-vectors of a set of L -frames as an activity sample that is assumed to contain adequate motion data to be classified as one of the activities in C . L represents the window length or temporal duration (in number of frames) to be found experimentally (see Section 3.2). The n^{th} activity sample of a video segment V_i is formulated as $[H_n] = \{B_{in}\}_{i=1}^L$. The number of activity samples in V_i depends on the temporal duration of V_i (F_i), the length (L) and overlapping percentage (ν) of the window technique applied. For example, a video segment with $F_i = 130$ frames, $L = 100$ frames and $\nu = 50\%$ has approximately three activity samples. The order of the L grid frames in H_n should be similar to B_k in order to keep the temporal relation across frames, which is later exploited to extract frequency-based features. We describe below each element of GOFF for an activity sample H_n . The discussion on the analysis of parameter values is given in Section 3.2.

MMHF is derived from the histogram representation of grid optical flow magnitude $[I_n]_{G^2 \times L}$. A generic example of MMHF

computation is shown in Fig. 7. The magnitude of each grid motion vector B_k^g is $\sqrt{(B_{k_x}^g)^2 + (B_{k_y}^g)^2}$ and we apply histogram computation on I_n using β_m magnitude bins to obtain the histogram representation $[O_n]_{\beta_m \times L}$. We apply non-uniform quantization since the majority in I_n are less than a single-pixel motion for most of the activities considered. In order to avoid unexpected high motion magnitude, which is not often a real ego-motion, we apply a Gaussian smoothing to I_n prior to the histogram computation. The histogram motion representation reduces the motion dimension from $G^2 \times L$ of H_n to $\beta_m \times L$ of O_n since $\beta_m < G^2$. Finally, the MMHF vector $[S_n^1]_{\beta_m \times 1}$ of an activity sample is computed from a normalization per frame in Eq. (1), followed by a summation along each bin in Eq. (2) (similarly to [12]):

$$\tilde{O}_n(:, \zeta) = O_n(:, \zeta) / \sum_{b=1}^{\beta_m} O_n(b, \zeta), \quad (1)$$

$$S_n^1(b) = \sum_{\zeta=1}^L \tilde{O}_n(b, \zeta). \quad (2)$$

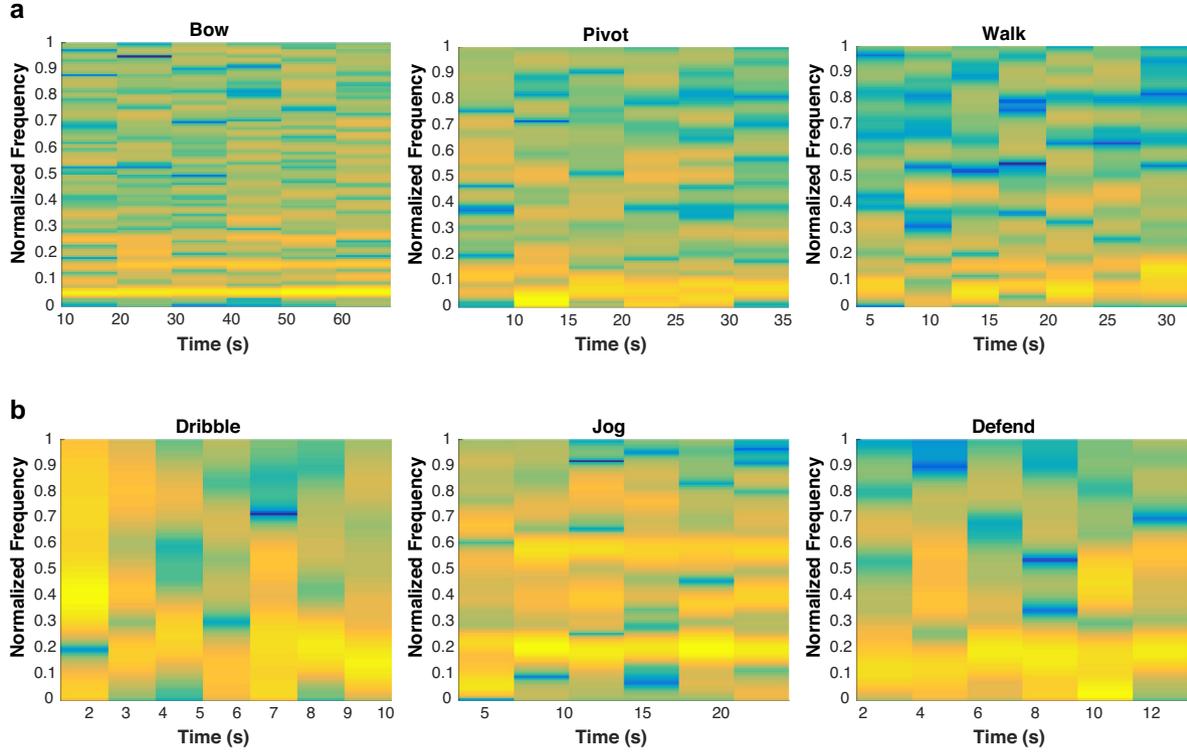


Fig. 5. Motion-direction spectrograms that reveal the discriminative power of frequency-based information. (a) low frequency activities (*Bow*, *Pivot* and *Walk*); (b): high frequency components (*Dribble*, *Jog* and *Defend*).

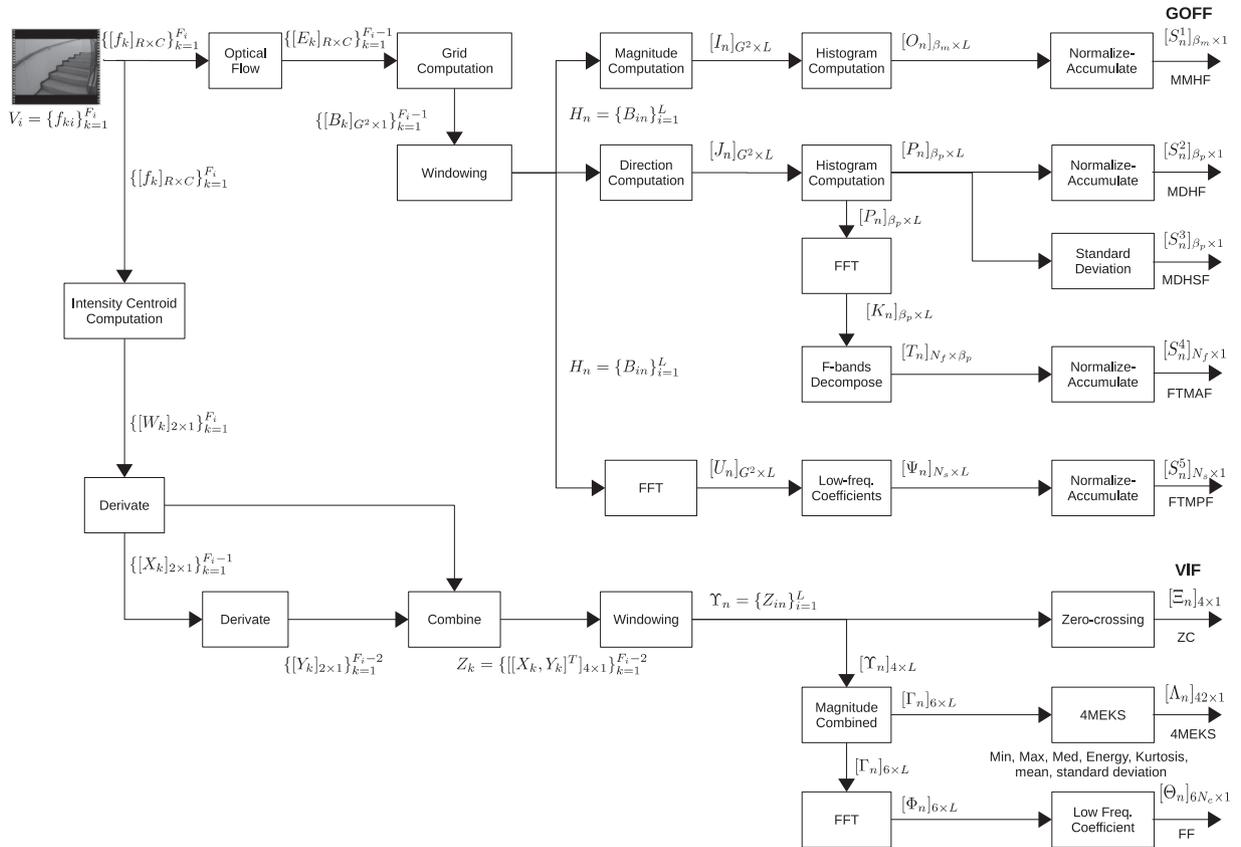


Fig. 6. Detailed block diagram of the extraction of the proposed multi-dimension motion features. GOFF: grid optical flow-based features; VIF: vision-based inertial features; FFT: fast Fourier transform; MMHF: motion magnitude histogram feature; MDHF: motion direction histogram feature; MDHSF: motion direction histogram standard deviation feature; FTMAF: Fourier transform of motion direction across frames; FTMPF: Fourier transform of grid motion per frame; ZC: zero-crossing; 4MEKS: minimum, maximum, median, energy, kurtosis, mean and standard deviation; FF: Frequency-based feature.

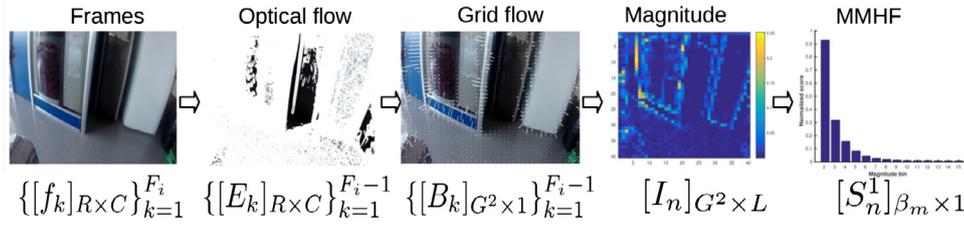


Fig. 7. A generic example to demonstrate the step-by-step computation of MMHF using $G = 40$ grids and $\beta_m = 15$ magnitude bins for a *Stand-up* activity.

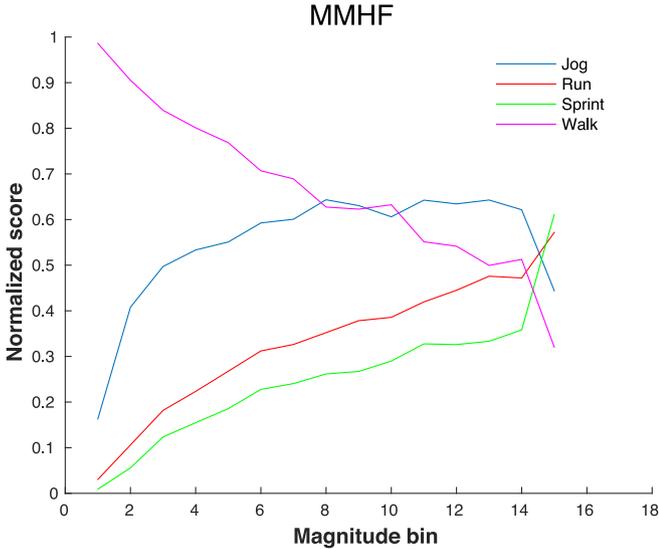


Fig. 8. MMHF vectors built from using $\beta_m = 15$ magnitude bins in the range of motion magnitudes $[0, 1]$ for *Walk*, *Jog*, *Run* and *Sprint* activities. The figure demonstrates that *Sprint* and *Run* contain higher magnitude values in the 15th bin and few motion grids of lower magnitude values, opposite to *Walk*, whereas *Jog* endures intermediate values as expected.

The summation in Eq. (2) accumulates the histogram representation of the motion magnitude I_n . In case some of the L frames contain noise or experience false ego-motion (e.g., due to a passerby), their effect on the final feature vector is minimized by the normalization (Eq. (1)) and the accumulation with other noise-free frames in the sample (Eq. (2)). MMHF is particularly advantageous to discriminate activities which involve similar direction patterns but different motion magnitudes. Examples include *Walk*, *Jog*, *Run* and *Sprint* for which the MMHF vectors are plotted in Fig. 8. MMHF values before normalization confirm the actual variation of motion magnitudes for these activities. Numerically, *Sprint*, *Run* and *Jog* video segments in BAR dataset (720×1280 resolution and $30fps$) are found to contain 87%, 81% and 62% of the frames with average magnitude greater than one pixel, respectively; while only the 45% of the frames have such magnitude value in a *Walk* segment.

MDHF is the histogram representation of the motion direction that is determined as $\tan^{-1}(B_k^{g,y}/B_k^{g,x})$ for a grid B_k^g . MDHF is computed similarly to the MMHF shown in Fig. 7, but using motion direction instead of magnitude; hence, it is vital to classify activities that might have similar motion magnitudes (Fig. 9). We develop the histogram representation $[P_n]_{\beta_p \times L}$ from motion-direction of an activity sample $[U_n]_{G^2 \times L}$ using β_p direction bins, where each bin covers a range of $(2\pi/\beta_p)^\circ$. The histogram representation reduces the motion dimension from $G^2 \times L$ of H_n to $(\beta_p \times L)$ of P_n since $(\beta_p < G^2)$. Then the normalization in Eq. (1) and the summation in Eq. (2) are performed on P_n in order to obtain the MDHF vector, S_n^2 .

MDHSF represents the standard deviation of each direction bin in MDHF across P_n , formally, $[S_n^3]_{\beta_p \times 1} = \sigma([P_n]^T)$, where σ represents

the standard deviation. Activities that involve high ego-motion (e.g., *Sprint* and *Run*) tend to possess higher variations, whereas slower activities (e.g., *Walk*) have minimal variations (Fig. 10). Different values of normalized score deviations: *Sprint* (0.11), *Run* (0.09), *Jog* (0.08) and *Walk* (0.06), reflect the level of dynamics in these activities. It is observed that *Sprint* and *Walk* relatively experience the highest and lowest dynamics, respectively.

FTMAF is a frequency-domain feature that contemplates the variation of direction bins in P_n ; and differently to MDHSF, it quantifies the detailed dynamics of motion direction. We compute the Fast Fourier Transform (FFT) of each bin in P_n to obtain $[K_n]_{\beta_p \times L}$, which is later decomposed into N_f frequency bands, $[T_n]_{N_f \times \beta_p}$. To do so, we consider only the half width ($L/2$) of K_n due to the symmetry property of the Fourier transform. The n_f^{th} band of the b^{th} bin in $[T_n]_{N_f \times \beta_p}$ is obtained as

$$T_n(n_f, b) = \sum_{l=1+\frac{(n_f-1)L}{2N_f}}^{\frac{n_f L}{2N_f}} K_n(b, l), \quad (3)$$

where each row of K_n is the FFT of the corresponding row in the direction histogram P_n . The FTMAF vector S_n^4 is derived from T_n using the normalization and summation operations in Eqs. (1) and (2), respectively. The majority of human ambulatory activities store much of their energy in the low frequency bands though significant variations can be depicted in Fig. 11. On the one hand, *Jog* and *Run* are found to have high values in the 10th and 11th frequency bands while *Sprint* possesses even higher frequency components (12th–14th bands). On the other hand, activities that have simple motion patterns (e.g., *Bow*, *Left-right turn* and *Sit-stand*) are shown to contain significant energy in the 3rd frequency band. The range of each band in the frequency response is defined in Eq. (3).

FTMPF is another frequency-based feature and measures the variation of grid optical flow in a frame. It is different from FTMAF since the FFT is performed on each frame in H_n smoothed by a Gaussian filter. FTMPF helps to discriminate complex activities with high dynamics of ego-motion (e.g., *Dribble*) from simple activities (e.g., *Walk*). The higher the ego-motion, the less likely the grid motion is to remain uniform. Since highly variant optical flow is not expected in a frame with the assumption of a uniform global motion, we select only the first N_s coefficients of the frequency response. The FTMPF vector S_n^5 is then calculated from $[\Psi_n]_{N_s \times L}$, which is the low frequency part of $[U_n]$, using Eqs. (1) and (2).

Finally, we combine MMHF, MDHF, MDHSF, FTMAF and FTMPF to obtain the GOFF descriptor for the n^{th} activity sample in a video V_i as,

$$[GOFF_n]_{N_g \times 1} = [S_n^1, S_n^2, S_n^3, S_n^4, S_n^5]^T, \quad (4)$$

where $N_g = \beta_m + \beta_p + \beta_p + N_f + N_s$ that sums the dimensions of MMHF, MDHF, MDHSF, FTMAF and FTMPF, respectively. The summary of GOFF is given in Table 2.

3.1.2. Vision-based inertial features (VIF)

The virtual inertial data generated from a video contain centroid velocity and acceleration values, both are derived from

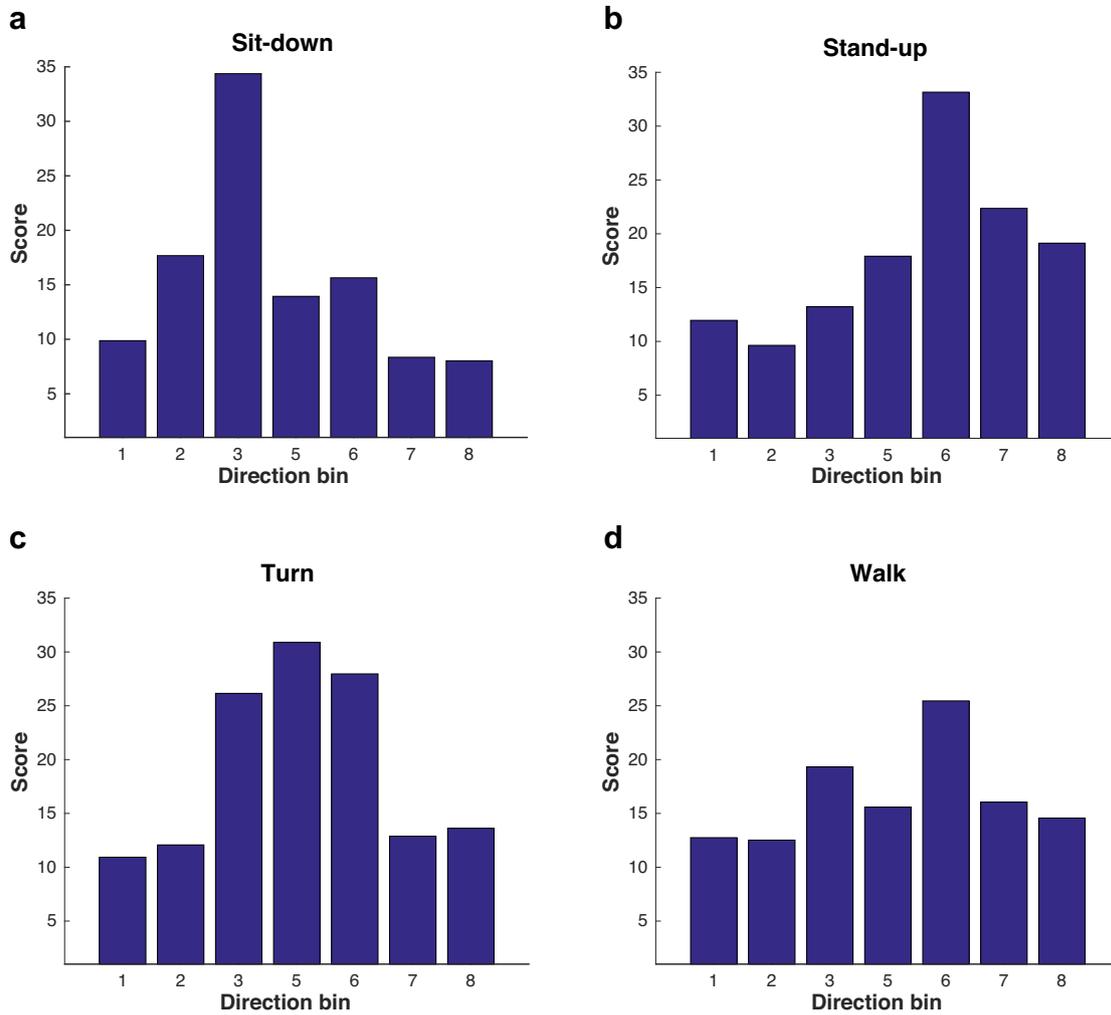


Fig. 9. The MDHF representations of (a) *Sit-down*, (b) *Stand-up*, (c) *Turn* and (d) *Walk* using $\beta_p = 8$ direction bins. Note that the 4th bin, which contains 0° , is not shown to achieve better visualization. It is clearly seen that MDHF vectors of *Sit-down* and *Stand-up* are mirror images to each other, reflecting the opposite motion directions they possess. *Sit-down* contains dominant motion direction of -1.35 ± 0.39 rad while *Stand-up* mainly lie in 1.35 ± 0.39 rad. On the other hand, the high score of the 5th direction bin centered at 80.79° in (c) shows the *Turning* angle in this particular video segment.

Table 2

Summary of GOFF and the motion characteristics each feature type describes. Variation refers to a difference among classes that a feature subgroup exploits; β_m : magnitude bins; β_p : direction bins; N_f : number of frequency bands in FTMAF; N_s : number of low frequency coefficients in FTMPF.

Subgroup	Measures	Variation	Symbol	Dimension
MMHF	Motion magnitudes using histogram bins	Average magnitude	S_n^1	β_m
MDHF	Motion direction using histogram bins	Average direction	S_n^2	β_p
MDHSF	Standard deviation of direction bins	Direction deviation	S_n^3	β_p
FTMAF	Variation of each direction bin in-detail	Periodicity (frequency)	S_n^4	N_f
FTMPF	Variation of grid optical flow in a frame	Ego-motion complexity	S_n^5	N_s

varying intensity centroid across frames in a video. Note that the inertial data are of the intensity centroid, and they are not equivalent to the actual values measured by real inertial sensors. It is not also our aim to replace the existing inertial sensor in the state of the art. In order to determine the centroid, we employ the procedure in Rublee et al. [28] that uses the first four image moments, \mathcal{M}_{pq} , where $p, q \in \{0, 1\}$. Each image moment of order $p + q$, \mathcal{M}_{pq} , is calculated as the weighted average of all intensity values in a frame (Algorithm 1). The velocity and acceleration values are computed by applying the first and second derivative, respectively, on the sequence of Gaussian-smoothed intensity centroids in a video.

The velocity $[X_k]_{2 \times 1}$ and acceleration $[Y_k]_{2 \times 1}$ vectors for each frame are concatenated $Z_k = \{[X_k, Y_k]^T\}$ before we apply L -frames long window, similarly to GOFF, to build an activity sample $[\Upsilon_n]_{4 \times L} = \{Z_{kn}\}_{k=1}^L$. Later, velocity and acceleration magnitudes of each frame are included $[\Gamma_n]_{6 \times L} = [\Upsilon_n^T, |X_n|^T, |Y_n|^T]^T$ in order to extract the inertial features-VIF, which contain time and frequency-domain features adopted from the state of the art [8,9,26,48,51]. Time-domain features are *minimum*, *maximum*, *median*, *energy*, *kurtosis*, *zero-crossing*, *mean* and *standard deviation* of each inertial signal. Kurtosis describes the peak of a signal distribution with respect to its mean. All time and frequency-domain features except zero-crossing are derived for each inertial vector across a window. A frequency-domain feature

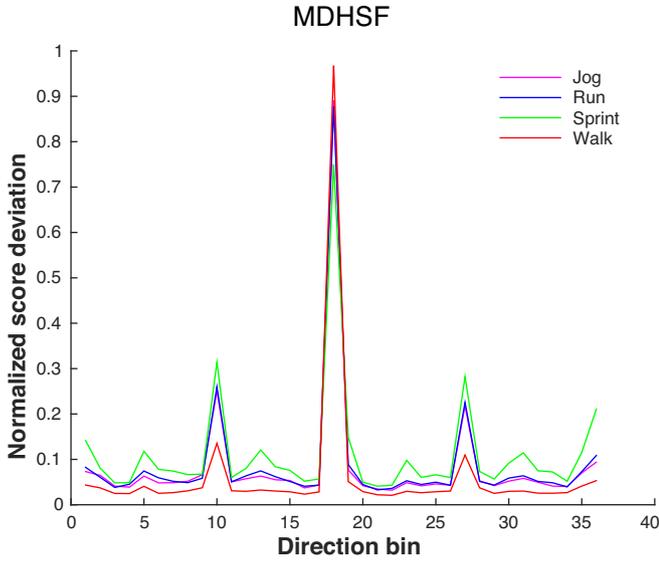


Fig. 10. MDHSF examples for *Jog*, *Run*, *Sprint* and *Walk* activities which are characterized by similar average direction in Fig. 4; but here, they are shown to have different variation of direction information (MDHSF) which reflects the level of dynamics in the activities.

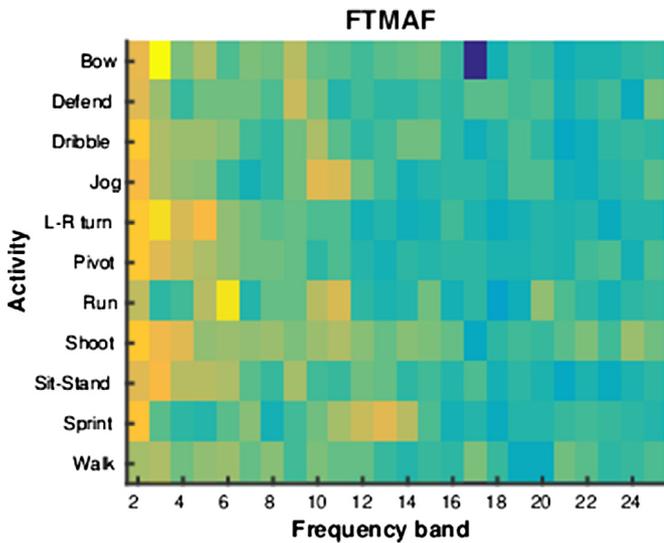


Fig. 11. FTMAF shows the distribution of the frequency response of direction histogram P_n across $N_f = 25$ frequency bands. All activities in the BAR dataset are seen to store much of their energy in the lowest frequency bands. The first band is not shown to visualize the distribution across the 24 frequency bands in-detail. *Jog*, *Run* and *Sprint* have higher frequency characteristics while *Bow*, *Left-right turn* and *Sit-stand* exhibit low frequency characteristics.

(FF) $[\Theta_n]_{6N_c \times 1}$ is generated from the FFT response $[\Phi_n]_{6 \times L}$ by selecting the first N_c low-frequency coefficients similarly to FTMPF in GOFF. VIF for an activity sample (VIF_n) is then obtained from the combination of the three feature subgroups as $[VIF_n]_{N_i \times 1} = S_n^6 = [\Xi_n, \Lambda_n, \Theta_n]$ where $N_i = 4 + 42 + 6N_c$.

Examples of inertial data generated from a *Left-right turn* video segment are shown in Fig. 12. The effect of a random appearance of a hand on the intensity centroid and optical flow is also shown. In general, if the duration of a clutter is long, it is considered as a part of background and false motion is less likely to be detected. However, if the duration is short enough, the remaining clutter-free frames in an activity sample help to reduce the error. Moreover, we also apply a Gaussian filtering across the inertial data in order to reduce noise effects.

Algorithm 1: Algorithm used to derive inertial data (velocity and acceleration) from a video in FPV.

Data: video (V_i)

Result: intensity centroid (W), velocity (X), and acceleration (Y)

% initialization

$F_i \leftarrow$ number of frames in V_i , $R \leftarrow$ row size, $C \leftarrow$ column size,

$\mathcal{M}_{pq} \leftarrow$ image moment,

$p, q \in \{0, 1\} \leftarrow$ moment orders,

for $k \leftarrow 1$ to F_i **do**

$\mathcal{M}_{pq}^k \leftarrow \sum_{r=1}^R \sum_{c=1}^C r^p c^q f_k(r, c)$ % image moments;

$W_k \leftarrow \begin{pmatrix} \mathcal{M}_{01}^k & \mathcal{M}_{10}^k \\ \mathcal{M}_{00}^k & \mathcal{M}_{00}^k \end{pmatrix}$ % intensity centroid

end

for $k \leftarrow 1$ to $F_i - 1$ **do**

$X_k \leftarrow W_{k+1} - W_k$ % velocity

end

for $k \leftarrow 1$ to $F_i - 2$ **do**

$Y_k \leftarrow X_{k+1} - X_k$ % acceleration

end

3.2. Parameters analysis

The extraction of both GOFF and VIF groups involve the appropriate setting of different parameters, namely grids G , window length L , overlapping ratio ν , direction bins β_p , magnitude bins β_m , frequency bands N_f and low-frequency coefficients N_s and N_c . The settings of the parameters are validated experimentally and discussed below.

An appropriate number of grids along each dimension of a motion frame is $G = 20$ as further increments of G do not tend to include new discriminative motion characteristic. This is because the motion, in general, is assumed to be dominantly global over the majority of pixels; so more grids are more likely to cause redundancy of the motion data contained in the twenty grids, which are selected from all parts of the frame using the periodical sampling (see Section 3.1.1). The window length L for the activities under analysis covers three seconds. This result is similar to the window length of ambulatory activity recognition using inertial data [48] and video data in FPV [8,9,11]. Higher values of L do not cause significant improvements in the system performance whereas the motion data become redundant and the number of activity samples from a video decreases. Similarly, the window overlapping (ν) experimented from 10% to 90% does not often affect the performance but reduces the number of activity samples. We therefore use $\nu = 50\%$.

We determine the number of bins for direction and magnitude histograms by experimentally optimizing the following trade-off. Very small values of β_p and β_m might not adequately quantize the direction and magnitude information of the optical flow data, whereas very high number of bins results in over-quantization and unnecessarily long feature dimension. Experiments reveal that $\beta_p = 36$ and $\beta_m = 15$ perform better. The number of frequency bands in FTMAF is $N_f = 25$ whereas the number of low frequency coefficients is $N_s = 25$ in FTMPF and $N_c = 10$ in VIF. We select fewer coefficients in VIF to minimize the length of the overall feature vector since the Fourier transform is applied on each inertial signal in Γ_n . N_f , N_s and N_c need to be long enough to include discriminative frequency information.

3.3. Computation time

The wall-clock computation time elapsed to accomplish each sub-task in the proposed method (Fig. 6) is given in Table 3, for an averaged video segment of approximately ≈ 150 frames. The grid-optical-flow and intensity centroid computations from raw video data took $B_k^t = 2.13s$ and $W_k^t = 3.38s$, respectively. Among the GOFF subgroups, the frequency-based features FTMAF and FTMPF needed

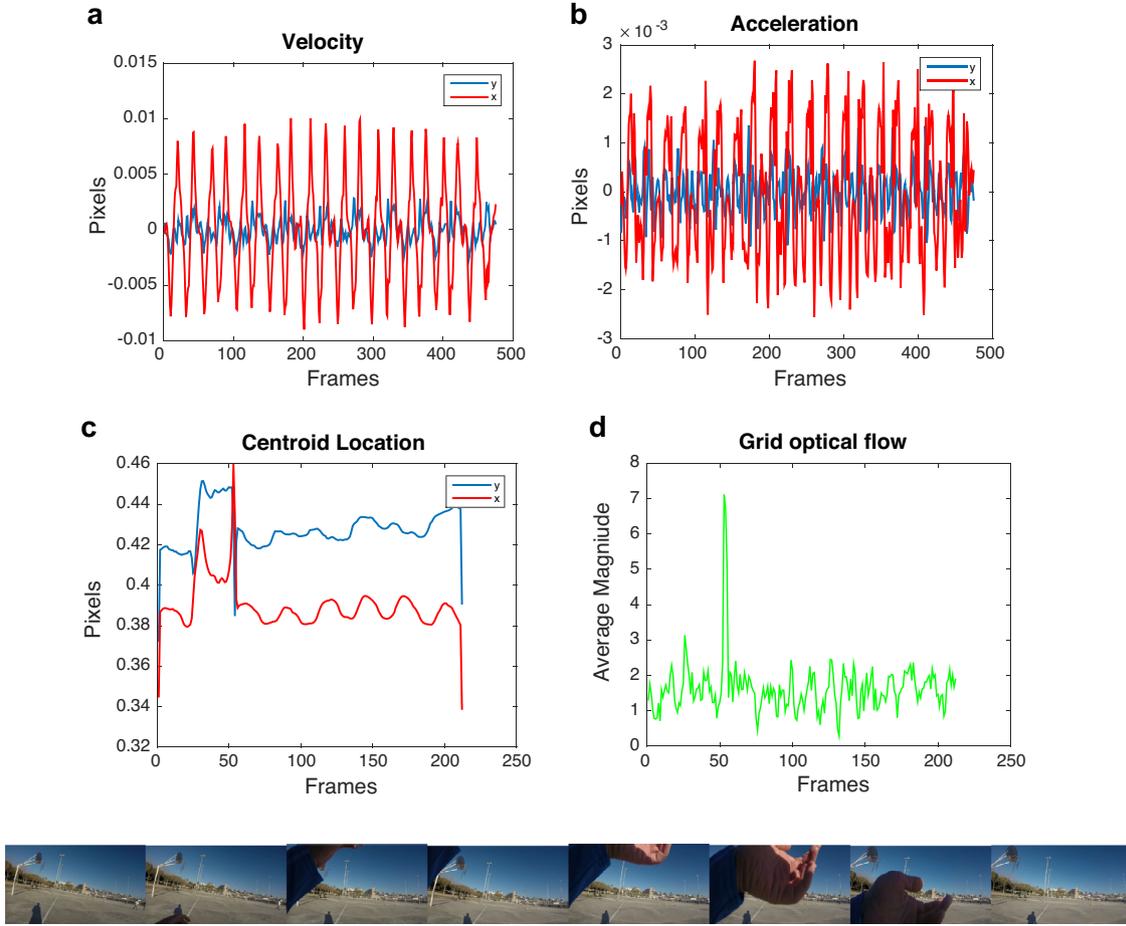


Fig. 12. Sample inertial data generated from the movement of the intensity centroid in *Left-right turn* video: (a) velocity, (b) acceleration; (c) and (d) demonstrate the effect of a randomly appeared user's hand (last row) during *Walking* on the intensity centroid and the optical flow data, respectively. Note that the pixel values of x and y in the intensity centroid are normalized by C and R , respectively.

longer time, $S_n^{4,t} = 6.80\text{ms}$ and $S_n^{5,t} = 18.37\text{ms}$, respectively. Overall, GOFF demanded 2.46s in relative to 3.39s of VIF; and hence, RMF is able to be computed in less than six seconds. All experiments were conducted using Matlab2014b, i7-3770 CPU @ 3.40GHz, Ubuntu 14.04 OS and 16GB RAM.

4. Datasets and validation protocol

We evaluate the performance of the proposed method and compare it against a baseline (see Section 4.3) and three state-of-the-art methods [8,12] and [14] across four datasets. The first state-of-the-art method is an interest point-based motion feature extraction approach presented in Zhang et al. [12,13], and referred to as multi-resolution good-feature (MRGF) implemented with SURF, which was reported to achieve better accuracy than Shin and Tomasi features [35]. The other two are from optical flow-based methods. The second method was employed in Zhan et al. [8,9,11], and we refer to it as average pooling (AP) because of the pooling procedure used to smooth the grid flow. The third method was proposed in [14] and we refer to it as motion-based histograms (MBH) because of the use of concatenated histograms to encode direction, magnitude and frequency components.

We use the following measures to assess performance of the recognition system: precision (\mathcal{P}), sensitivity or recall (\mathcal{R}), specificity (\mathcal{S}), accuracy (\mathcal{A}) and F_1 -score (\mathcal{F}):

$$\mathcal{P} = \frac{TP}{TP + FP} \quad \mathcal{R} = \frac{TP}{TP + FN} \quad \mathcal{S} = \frac{TN}{TN + FP}$$

$$\mathcal{A} = \frac{TP + TN}{TP + TN + FP + FN} \quad \mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \quad (5)$$

where TP : true positive, TN : true negative, FP : false positive and FN : false negative. GOFF and VIF are studied independently and the contributions of their feature elements to the overall performance are analyzed. In order to test the robustness of the methods, we introduce artificial Gaussian noise of different signal-to-noise ratio (SNR) values on the motion data. In addition to this, we test the proposed features for noisy data collected in previously unseen environment during training. We also analyze the sensitivity of our method by varying parameter settings of the feature subgroups, namely number of direction histogram bins (β_p), magnitude histogram bins (β_m), frequency bands (N_f) and low frequency coefficients (N_s and N_c).

4.1. Datasets

Among the four datasets utilized in this work, two are new and recently collected by ourselves. The other two are public datasets used in [20] and [30].

4.1.1. Our datasets

We collected two datasets with the aim of providing different environmental conditions and various activities: indoor ambulatory activity recognition (IAR) dataset and basketball activity recognition (BAR) dataset (Fig. 13). The IAR dataset contains the most frequently studied activities in the state of the art [8–10,12,13], namely *Walk*, *Run*, *Sit-down*, *Stand-up*, *Going upstairs*, *Going downstairs* and *Turn* in addition to *Jump*. Recording was conducted in three buildings with

Table 3

Summary of wall-clock computation time to process an averaged 150 frames long video segment. Elapsed time is measured for each sub-process in Fig. 6 and latter summarized for GOFF and VIF.

GOFF	$B_k^t = 2.13s$	$I_n^t = 0.32ms$	$O_n^t = 1.66ms$	$S_n^{1,t} = 2.11ms$	$GOFF_n^t = 2.16s$
		$J_n^t = 0.89ms$	$D_n^t = 1.37ms$	$S_n^{2,t} = 2.41ms$	
			$T_n^t = 5.69ms$	$S_n^{3,t} = 2.55ms$	
		$U_n^t = 18.19ms$		$S_n^{4,t} = 6.80ms$	
				$S_n^{5,t} = 18.37ms$	
VIF	$W_k^t = 3.38s$	$Z_k^t = 0.21ms$	$\Xi_n^t = 0.08ms$	$S_n^{6,t} = 2.12ms$	$VIF_n^t = 3.39s$
			$\Lambda_n^t = 1.08ms$		
			$\Theta_n^t = 0.96ms$		

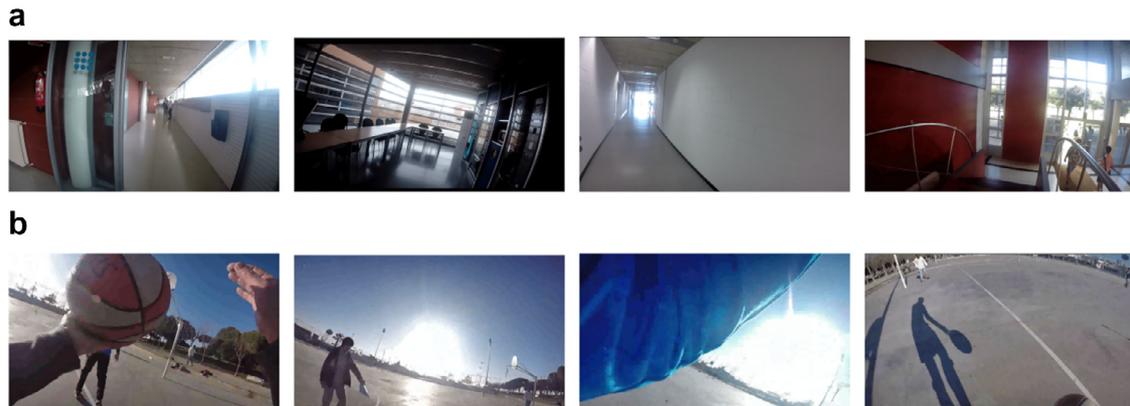


Fig. 13. Key-frames from (a) IAR and (b) BAR datasets depict some of the challenges in FPV-ambulatory activity recognition. The challenges in the IAR dataset include the effect of outdoor lighting, the lack of adequate indoor lighting and texture, and the mixing of outdoor scenes where the walls are made of glass. The challenges in BAR dataset include appearances of other subjects, body parts and shadows.



Fig. 14. Temporal sequence of frames from a *Pivot* video. The top frames are from the ground-truth video (external camera) while the bottom frames represent the corresponding instances in the first-person video. Depending on the relative position of the wearable camera to the external camera, similar or different scene contents might appear in the two videos.

different light conditions and indoor architectures such as staircases, corridors and wall textures. We assumed a separate occurrence of each activity, meaning that, activities like *Run* while *Going upstairs* were not considered. However, we included scenarios such as *Sit* or *Jump* on stair-cases. Note that even if the recordings were done in indoor locations, outdoor scenes and lighting were sometimes present (Fig. 13a).

The BAR dataset is composed of three warming-up exercises (adopted from [8,9,12,13]) and eight activities in a basketball game. This is the first dataset that includes basketball activities from FPV. The activities are *Bow*, *Sit-Stand*, *Left-right turn*, *Walk*, *Jog*, *Run*, *Sprint*, *Pivot*, *Shoot*, *Dribble* and *Defend*. Basketball activities were primarily defined by experts interviewed before the data collection, which was performed in an outdoor basketball court with four male subjects of different ages and playing experiences. Even if only a camera wearer was engaged in playing basketball during the recording, the scenes often contained the other subjects and/or shadows (Fig. 13b). Few frames of *Pivot* from both the external and wearable cameras are shown in Fig. 14.

In general, activities in the BAR dataset are more challenging as compared to the IAR dataset due to the following reasons. First, motion capture of sport activities usually involve motion parallax, blur and shutter effect along with high ego-motions [14]. Second, there is less inter-activity variation among few activities. Examples include *Left-right turn* and *Pivoting*; *Bow* and *Sit-Stand*; and *Jog* and *Run*. Third, the BAR dataset does also contain high intra-class variations in some activities. Examples include *Shoot*, which can be a *jump-shoot* or *layout-shoot*; *Pivot*, which can be performed in clockwise or counter clockwise directions; *Defend*, which can be *slide-defend* or *backward-defend*. Other challenges result from different age and playing experience of the subjects. An example is the similarity between *Sprint* and *Run* of older and younger subjects, respectively. A chest mounted GoPro Hero3+ Silver edition camera is used to record all the activities. Chest mounting is selected in order to maximize the quality of the data with respect to acquiring a full-body motion [12,13,19]. IAR was collected with a resolution of 1080×1920 and $60fps$, while 720×1280 with $30fps$ was set for the BAR dataset. A summary of our datasets is shown in Table 4.

Table 4

Summary of number of video segments in the IAR and BAR datasets; the top sub-table describes the IAR dataset and the number of video segments per activity in the three recordings (R1, R2 and R3). The bottom sub-table presents the contribution of the four subjects (S1, S2, S3 and S4) in the BAR dataset. Note that activities with shorter durations (e.g., *Shoot*) tend to have more video segments in order to achieve data balance. Reco.: Recording; Sub.: Subject; L-R: Left-right turn; S-S: Sit-Stand; Dur: Duration in minutes.

IAR												Dur.	
Reco.	Walk	Turn	Stand	Up-stair	Down-stair	Sit	Run	Jump	Total			(min)	
R1	14	16	13	15	13	13	13	14	111			12	
R2	21	21	24	18	17	23	22	20	166			11	
R3	21	23	21	3	3	23	9	14	117			17	
Total	56	60	58	36	33	59	44	48	394			40	
BAR												Dur.	
Sub.	Bow	Defend	Dribble	Jog	L-R	Pivot	Run	Shoot	S-S	Sprint	Walk	Total	(min)
S1	4	3	8	4	8	14	4	30	4	2	4	85	15
S2	4	6	8	4	4	6	4	30	4	4	4	78	15
S3	4	9	8	4	4	14	4	29	4	4	4	88	22
S4	4	6	6	4	5	12	4	26	5	4	4	80	20
Total	16	24	30	16	21	46	16	115	17	14	16	331	72



Fig. 15. Key-frames from JPL-interaction dataset [20]. Activities from left to right are *Hug*, *Pet*, *Shake*, *Point*, *Punch*, *Throw* and *Wave*. All videos were recorded indoors with the participation of eight subjects.

4.1.2. Public datasets

JPL-interaction [20] and DogCentric [30] are the two public datasets used in this work. JPL-interaction dataset was collected in five indoor locations of varied background conditions. A toy that emulated a robot was placed on a chair, on which a GoPro camera was mounted with a resolution of 320×240 and $30fps$. The set of activities (Fig. 15) include four friendly, one neutral and two hostile interactions between a participant and the toy. The friendly activities are *Hug*, *Pet*, *Shake* and *Wave*. The neutral interaction is *Point*, where two persons often point towards the toy while they are having a conversation. *Punch* and *Throw* are the hostile interactions. Eight participants were involved and a total of twelve video sets were produced (two subjects did more than one experiments). Most of the sets contain seven video segments, one per activity. Key-frames from the dataset are shown in Fig. 15 and the summary of segment durations in the video sets is presented in Table 5.

Due to the lack of public motion-oriented human-centric datasets, we also experimented on recently released DogCentric dataset [30]. Though the motion patterns of dogs are completely different from human motions; IAR, BAR and DogCentric datasets share similar guideline as the motion in an egocentric video infer the type of activity being performed by a subject in IAR and BAR datasets or by a dog in the DogCentric dataset. Four dogs were used while a GoPro camera (320×240 and $30fps$) was mounted on the back of each of the four dogs. The dog-centric activities considered are *Play with a ball*, *Car passing-by*, *Drink*, *Look-left*, *Look-right*, *Pet*, *Shake*, *Sniff* and *Walk*. Key-frames from the dataset are shown in Fig. 16 and the number and duration of video segments collected from each dog and per activity type are presented in Table 6.

4.2. Experimental setup

We validate the proposed motion-feature (RMF) using two geometrical classifiers [54] SVM and KNN which are, respectively, the most frequently employed parametric and non-parametric modeling techniques in the state of the art (see Section 2). We select one-versus-all approach for the SVM due to its smaller number of classi-

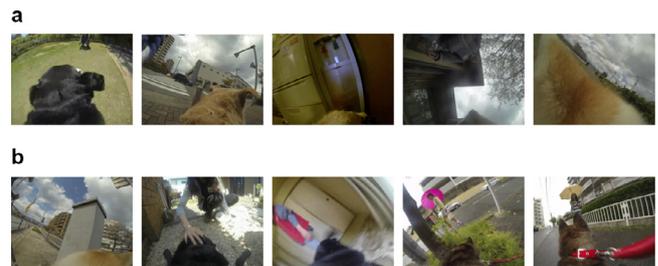


Fig. 16. Key-frames of the DogCentric dataset [30]. Activities from left to right are (a): *Ball*, *Car*, *Drink*, *Feed*, *Look-left*; (b): *Look-right*, *Pet*, *Shake*, *Sniff* and *Walk*. The dataset includes both indoor and outdoor environments and sometimes appearance of people.

fications with respect to one-versus-one approach. We assume that a test video segment V_i belongs to only one of the activity classes $A_j \in \mathcal{C}$, and we do not consider undefined class that represents none of the activities in \mathcal{C} . Experimental results reveal that polynomial kernel performs better than linear and Gaussian kernels in the SVM (Table 7). We set the number of KNN neighbors to be one since the performance is found to be less sensitive to the number of neighbors.

We set grid $G = 20$ and window length $L = 180$ frames for IAR and $L = 95$ frames for BAR with $\nu = 50\%$ overlapping (for the analysis see Section 3.2). The difference in window length comes from the different frame rate used in the two datasets, $60fps$ and $30fps$ respectively. We also found optimal parameter values for the state-of-the-art methods. MRGF was reported with a magnitude threshold of three pixels and eight direction bins in Zhang et al. [12,13]; however, higher performance is achieved with seven pixels threshold and thirty-six bins in our datasets, particularly in IAR. In general, we set $\beta_p = 36$, $\beta_m = 15$, $N_f = 25$, $N_s = 25$ and $N_c = 10$.

We applied a random decomposition in the IAR dataset to build train and test sets as 80% train and 20% test. The final accuracy was computed from the mean of results obtained from 100 iterations of train-test decomposition. In the BAR dataset, we employed a leave-one-subject-out approach and the final accuracy was derived from the mean of results obtained after each subject is left-out iteratively.

Table 5

Duration details of video segments in JPL-interaction dataset, measured in seconds, presented in accordance with the activity and video sets. The whole dataset is ≈ 10 min long in which activities *Point* and *Hug* account more than half of the overall dataset duration whereas *Wave* is the shortest activity.

	JPL dataset – video sets												Total (s)
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
<i>Hug</i>	9	11	10	9	17	10	11	6	11	15	13	10	132
<i>Pet</i>	8	8	6	11	16	6	6	5	7	10	7	7	97
<i>Point</i>	14	6	10	13	18	22	35	13	17	34	25	30	237
<i>Punch</i>	2	3	1	2	1	2	2	3	1	2	2	1	22
<i>Shake</i>	6	7	5	5	5	4	3	3	8	7	3	5	61
<i>Throw</i>	4	5	3	4	4	4	5	4	3	3	4	2	45
<i>Wave</i>	1	1	1	1	2	2	1	1	2	2	2	2	18
Total	44	41	36	45	63	50	63	35	49	73	56	57	612

Table 6

Details of DogCentric dataset. The number of video segments per activity is shown for each dog participated in the experiment. The overall duration of video segments recorded from each dog and for each activity is also presented. Dur.: duration in seconds.

	DogCentric- video segments per activity										Total (#)	Dur. (s)
	Ball	Car	Drink	Feed	Look-left	Look-right	Pet	Shake	Sniff	Walk		
<i>DogA</i>	6	7	5	7	8	7	8	8	8	7	71	248
<i>DogB</i>	5	1	2	3	4	2	4	2	7	4	34	139
<i>DogC</i>	3	14	2	8	3	4	8	3	7	7	59	313
<i>DogD</i>	0	4	1	7	6	5	5	5	5	7	45	142
Total	14	26	10	25	21	18	25	18	27	25	209	842

Table 7

F_1 -score measure (%) of different kernel types for the SVM classifier. **IAR**: indoor activity recognition; **BAR**: basketball activity recognition; **JPL**: JPL-interaction dataset; **DogC**: DogCentric dataset. Polynomial kernel achieves the highest accuracy for the majority of the methods in all datasets.

Dataset	Kernels	Baseline	AP	MRGF	MBH	RMF
IAR	Linear	53	51	72	66	83
	Gaussian	57	55	89	67	87
	Polynomial	58	63	84	65	88
BAR	Linear	24	14	25	60	75
	Gaussian	2	22	35	64	77
	Polynomial	19	17	37	65	80
JPL	Linear	3	1	42	63	83
	Gaussian	4	2	44	65	85
	Polynomial	2	12	62	63	86
DogC	Linear	34	41	30	40	55
	Gaussian	21	45	40	48	61
	Polynomial	25	41	39	59	61

We did not apply leave-one-out approach in the IAR dataset since the recordings (R1, R2 and R3) do not consist of equivalent number of video segments per activity (see Table 4).

However, for the JPL and DogCentric datasets, we adopted the corresponding approaches employed in [20] and [30], respectively. Ryoo et al. [20] used a repeated random sub-sampling validation to measure the classification accuracy. The video sets were decomposed into train and test randomly, each contained six video sets (42 segments). Experiments were repeated 100 times and the final accuracy was computed from the mean of the 100 iterations. For DogCentric dataset, video sequences of an activity were randomly decomposed into train and test sets, each containing half the number of total video sequences of the activity [30]. The mean final result was obtained by repeating this random train-test splits 100 times, as in the IAR and JPL datasets.

4.3. Baseline method

We develop a baseline method that estimates motion in a video by adopting the approach in Nagasaka et al. [55] (cited in Uehara et al. [56]) that utilizes the correlation of intensity projection to approxi-

mate pixel-wise displacement between subsequent pairs of frames. The aim of developing the baseline method is to make a comparison against the state-of-the-art and proposed methods using a simple motion-feature extraction approach.

The overview of the baseline method is shown in Fig. 17, which is similar to the VIF part of the proposed method in Fig. 6, but the centroid localization in VIF is replaced by the projections of intensity values $(f_k^j)_{j \in \{x,y\}}$ in the horizontal (x) and vertical (y) directions [55]. Given a current frame $[f_k]_{R \times C}$, the horizontal $[f_k^x]_{C \times 1}$ and vertical $[f_k^y]_{R \times 1}$ projections were computed as

$$f_k^x(c) = (1/R) \sum_{r=1}^R f_k(r, c), \quad (6)$$

$$f_k^y(r) = (1/C) \sum_{c=1}^C f_k(r, c).$$

We derive the projection velocity $X = [X^x, X^y]$ of the current frame f_k from the previous frame f_{k-1} using the following equation:

$$X^x = \underset{-\omega_x < \delta < \omega_x}{\operatorname{argmin}} (|f_k^x - (f_{k-1}^x < \delta >)|),$$

$$X^y = \underset{-\omega_y < \delta < \omega_y}{\operatorname{argmin}} (|f_k^y - (f_{k-1}^y < \delta >)|), \quad (7)$$

where ω_x and ω_y are, respectively, the maximum projection displacements that are assumed to exist between a pair of frames along the horizontal and vertical directions, and $(f_{k-1}^j < \delta >)$ is the circular shift of the projection f_{k-1}^j by δ pixels. Similarly to VIF, by applying a derivation on projection velocity X_k across frames, we obtain the corresponding acceleration vector Y_k . We extract kurtosis and frequency-domain features as in VIF, and magnitude histogram as in GOFF. We apply the same parameter settings as of the proposed method for the implementation, and set $\omega_x = \omega_y = 40$ pixels assuming that a true global-motion of higher displacement is less likely.

5. Results and discussions

Results show that the proposed feature representation (RMF) performs consistently higher than the state-of-the-art methods across

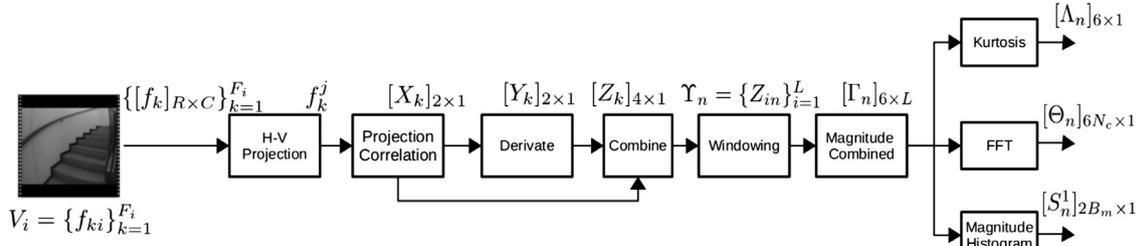


Fig. 17. Overview of the baseline method. H-V Projection: horizontal and vertical projections (f_k^j) _{$j \in \{x, y\}$} of intensity values. Projection displacement is computed from the correlation of intensity projections in each subsequent pair of frames. Acceleration, magnitude, kurtosis and FFT of VIF are applied and the final baseline feature vector consists of kurtosis (Λ_n), magnitude histogram (S_n^1) and frequency-domain features (Θ_n).

Table 8

Comparative performance (%) of the proposed (RMF) and state-of-the-art methods with respect to the baseline. **IAR**: indoor activity recognition; **BAR**: basketball activity recognition; **JPL**: JPL-interaction dataset [20]; **DogC**: dogcentric dataset [30]; SVM performance measures include \mathcal{A} : accuracy, \mathcal{P} : precision, \mathcal{R} : recall and \mathcal{F} : F_1 -score. Accuracy from the confusion matrix of KNN is also given. In IAR dataset, MRGF and RMF achieve similar performance and significantly higher than the other methods. However, all the methods, except RMF, find it difficult to achieve higher recognition rate consistently across the datasets. DogCentric dataset is proved to be more challenging for all the methods.

Datasets	Methods	\mathcal{A}	\mathcal{P}	\mathcal{R}	\mathcal{S}	\mathcal{F}	KNN
IAR	Baseline	91	69	46	98	55	53
	AP	92	85	42	99	56	49
	MRGF	97	90	87	98	88	79
	MBH	91	62	68	94	65	67
	RMF	97	91	85	99	88	78
BAR	Baseline	83	18	20	90	19	17
	AP	90	24	14	97	18	31
	MRGF	89	35	39	93	37	48
	MBH	95	63	67	97	64	71
	RMF	98	81	79	99	80	78
JPL	Baseline	84	5	1	98	2	13
	AP	76	5	16	86	7	34
	MRGF	85	55	72	87	62	55
	MBH	87	66	53	92	59	61
	RMF	96	87	85	97	86	82
DogC	Baseline	83	49	17	91	25	28
	AP	87	39	30	92	34	47
	MRGF	88	39	39	94	39	42
	MBH	86	38	27	92	32	51
	RMF	92	62	59	96	61	58

the four datasets considered. This highlights RMF's flexibility to work on a variety of activities and environmental conditions (Table 8). In IAR, MRGF and RMF achieve equivalent performance ($\mathcal{F} = 88\%$) followed by MBH ($\mathcal{F} = 65\%$), whereas AP and Baseline were found to be the least performing motion features ($\mathcal{F} = 56\%$). MRGF performed similarly to RMF (in IAR) and MBH (in JPL). This is partly because the scene is relatively closer to the camera in the two datasets since they were recorded indoors; and hence it is less challenging to detect interest points on which MRGF is built on. However, in the BAR and DogCentric datasets, where there are more complex ego-motions and activities with different motion magnitude patterns, both MRGF and MHB are found to have restricted discriminating potential. This is due to their lack of magnitude-based features and less effective encoding of direction information, respectively. In Fig. 18b and Table 9, we can see that MRGF has failed to discriminate *Jog*, *Run* and *Sprint* activities, which have similar direction patterns but different motion magnitude and frequency characteristics. The confusion between *Left-right turn* and *Pivot* in Fig. 18b do also share the same reason. MBH is also achieving lower accuracy in BAR dataset where activities like *Drizzle*, *Run* and *Sprint* are confused with each other. This is due to the lack of effective encoding of direction alternation (periodicity), a problem MDHSF and FTMAF are trying to solve in the RMF.

MBH also results in less accurate recognition of *Hug* than MRGF in the JPL interaction dataset (Fig. 18c and Table 9), which is also due to the difference in the encoding of the direction information in the two methods. In the JPL dataset, *Pet* is confused with *Shake* commonly across MRGF, MBH and RMF. This is because participants involve shake-type actions while petting the toy (e.g., holding the two hands of the toy and moving up and down). The DogCentric dataset is found to be more challenging for all the methods since the type of activities in the dataset (e.g., *Drink*, *Feed* and *Sniff*) contain salient local information in addition to the global motion they possess. *Look-left* and *Look-right* are also hardly recognized (Table 9) as they are misclassified to *Walk* (Fig. 18d). This is because the dogs were walking most of the time while they were performing these activities. In addition, the camera was mounted on the back of the dogs (not on the head), hence, crucial activity information was not recorded.

Generally, MBH and MRGF complement each other across the datasets, particularly in JPL and DogCentric datasets, since MRGF is based on motion direction while MBH exploits motion magnitude more than direction component. Table 8 demonstrates that the comparative performance of the methods in accordance with their KNN outputs reflect similar patterns with respect to the F_1 -score of their SVM outputs. Due to the one-versus-all approach of the SVM classifier, the normal accuracy and specificity of all the methods are very high as expected.

Not only does RMF outperform the state-of-the-art global motion features considered (Tables 8, 9 and Fig. 18), but also achieves competitive results reported in Ryoo et al. [20] (86% vs. 90%) and Dog-Centric [30] (61% vs. 60%) which leveraged structural matching and combination of multiple local features in addition to global motion features to achieve the reported results. This signals the potential of RMF to discriminate interaction-based activities (with people or objects) while coupled with other domain-specific features.

The per-class performance of the methods is presented in Table 9 where RMF is shown to be superior to other methods in precision, recall and F_1 measures across all datasets except IAR. This is because IAR is less challenging compared to others (see Section 4.1). The train-test decomposition scheme also plays a role in the improved performance since the cross-validation approach in IAR introduces correlation between train and test set activities; in comparison to the leave-one-subject-out approach in BAR and equal decomposition of train and test sets in JPL and DogCentric datasets.

RMF is also very rarely seen to achieve the second best recognition performance for few activities. Examples include: *Turn* and *Walk* in IAR dataset which are recognized better, $\mathcal{F} = 91\%$ and $\mathcal{F} = 85\%$, with purely directional motion feature (MRGF) in comparison with $\mathcal{F} = 83\%$ for both activities using RMF; *Sit-stand* in BAR dataset is recognized with higher F_1 -score using MBH ($\mathcal{F} = 83\%$) than RMF ($\mathcal{F} = 77\%$); *Feed* in DogCentric dataset is also recognized with highest precision using MBH ($\mathcal{P} = 34\%$) and recall using MRGF ($\mathcal{R} = 40\%$). However, RMF sustains superiority for the classes in BAR and JPL datasets. In BAR, RMF results in significantly higher recognition performance for simple activities that are characterized by dominant

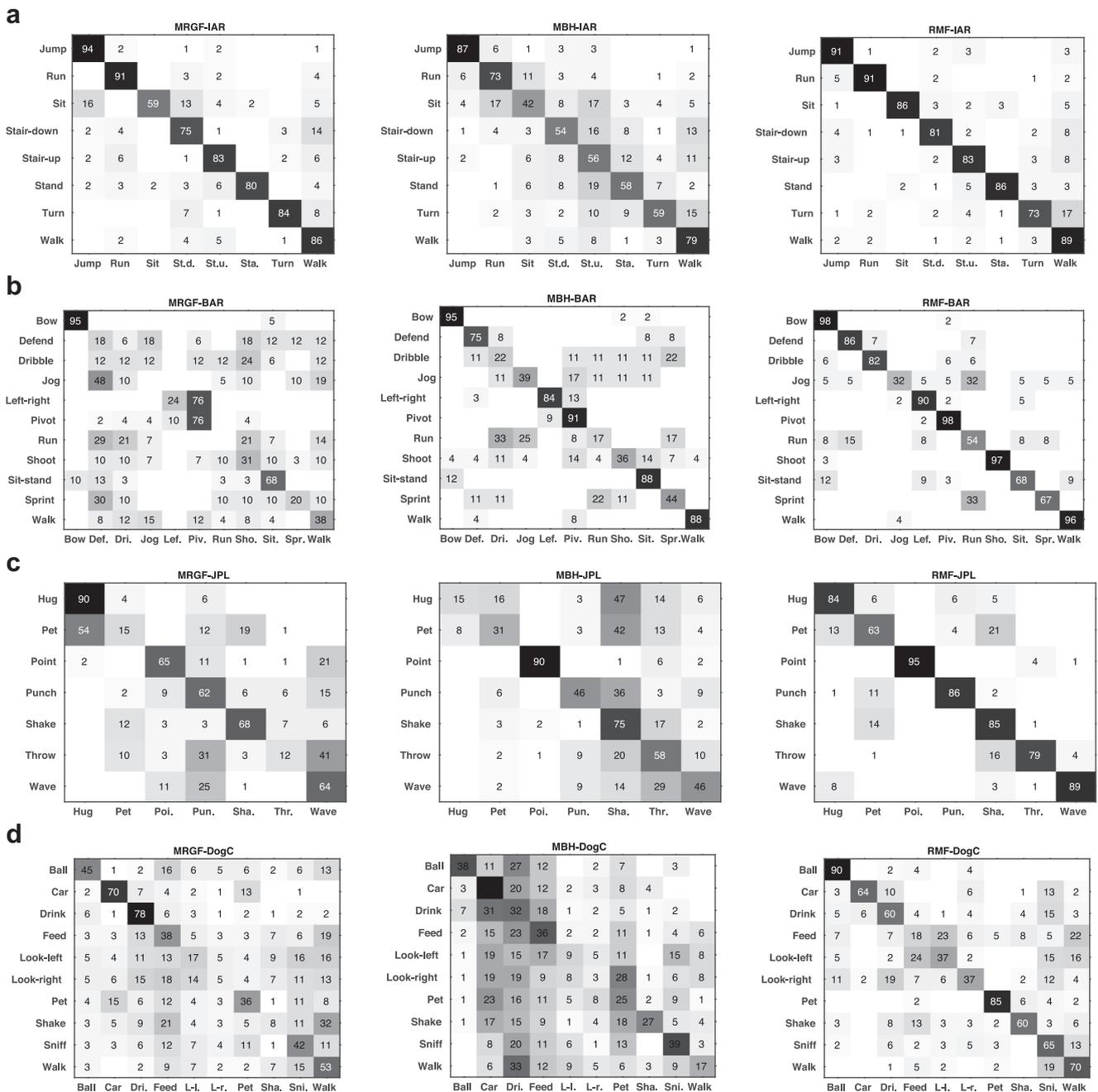


Fig. 18. SVM-validated confusion matrices of the methods in: (a) IAR, (b) BAR, (c) JPL and (d) DogCentric datasets. Baseline and AP methods are found to perform significantly inferior to MRGF, MBH and RMF (Tables 8 and 9); hence, their confusion matrices are discarded here for improved clarity. Though RMF achieves significantly higher performance in the majority of the datasets, it is possible to notice the difficulty posed by inter-class similarity between *Jog*, *Run* and *Sprint* in BAR dataset. In addition, weak recognition performances of RMF for *Feed*, *Look-left* and *Look-right* activities in DogCentric dataset signal the need of local descriptors, beside the limitation imposed by the mounting position of the camera.

motion along a single dimension while the player remains in a fixed position. Examples include *Bow* ($F = 93\%$), *Left-Right Turn* ($F = 88\%$), *Pivot* ($F = 93\%$), *Shoot* ($F = 97\%$) and *Sit-Stand* ($F = 77\%$). MBH follows RMF closely in BAR more than any of the other state-of-the-art methods except for *Dribble* where frequent changes of motion direction were not encoded effectively in the MBH.

Feature subgroups in GOFF are independently validated in all the datasets and compared against VIF as shown in Fig. 19. The results verify that the feature subgroups are ranked differently across the datasets, which signal the existence of different nature of variations among activities in the datasets. For example, due to directional variation of activities in IAR dataset (Fig. 4), direction-based feature subgroups (MDHF, MDHSF and FTMAF) show superiority

to magnitude-based feature MMHF (Fig. 19a). On the other hand, in BAR and JPL datasets where the activities involve different emotions and/or dynamics (e.g., *Sprint*, *Dribble* and *Defend*), MMHF and frequency-based feature FTMAF become significantly more important. In DogCentric dataset, none of the feature groups is found to dominantly surpass the others. The novel intensity centroid-based virtual inertial feature (VIF) is shown to excel more than any of the GOFF subgroups in the BAR, JPL and DogCentric datasets. As expected, FTMPF is the least performing subgroup of GOFF in the IAR and BAR datasets, where global motion is assumed to be dominant. Contrarily, FTMPF becomes more discriminative in JPL and DogCentric datasets, where local motion contains salient information (Fig. 19c and d).

Table 9

Per-class recognition performance (%) of RMF and the state-of-the-art methods. \mathcal{P} : precision; \mathcal{R} : recall; \mathcal{F} : F_1 -score of the SVM output. Apart from IAR dataset, RMF is shown to significantly outperform the state-of-the-art methods in the BAR, JPL and DogCentric datasets. Baseline and AP are shown to be less discriminant motion features since they did not encode magnitude and direction information effectively.

Dataset	Activity	Baseline			AP			MRGF			MBH			RMF		
		\mathcal{P}	\mathcal{R}	\mathcal{F}												
IAR	Jump	94	75	83	72	25	37	96	95	95	87	89	88	89	92	90
	Run	96	86	91	82	22	35	92	89	90	72	75	73	97	86	91
	Sit-down	38	25	30	94	41	57	91	91	91	46	52	49	91	85	88
	Stair-down	81	39	53	82	31	45	83	68	75	42	60	49	88	82	85
	Stair-up	49	12	19	87	61	72	89	83	86	45	58	51	92	85	88
	Stand-up	39	29	33	90	57	70	96	88	92	57	62	59	94	90	92
	Turn	88	52	65	97	34	50	93	90	91	72	68	70	90	77	83
BAR	Walk	64	54	59	78	66	72	83	87	85	72	81	76	84	83	83
	Bow	37	23	28	38	4	7	90	95	92	92	99	95	91	96	93
	Defend	7	49	12	0	0	0	3	7	4	59	66	62	82	88	85
	Dribble	10	40	16	0	0	0	11	12	11	8	16	11	87	85	86
	Jog	0	0	0	0	0	0	1	2	1	51	44	47	68	26	38
	Left-Right	6	10	8	52	12	20	67	30	41	89	91	90	90	86	88
	Pivot	25	6	10	23	34	27	49	83	62	68	92	78	89	97	93
	Run	3	6	4	0	0	0	2	2	2	23	24	23	40	42	41
	Shoot	97	16	27	75	3	6	37	45	41	75	51	61	99	97	98
	Sit-stand	8	23	12	19	10	13	56	64	60	75	92	83	80	75	77
JPL	Sprint	2	25	4	0	0	0	13	31	18	56	62	59	76	76	76
	Walk	4	18	7	51	91	65	59	54	56	91	95	93	91	96	93
	Hug	0	0	0	0	0	0	75	96	84	76	16	26	77	90	83
	Pet	0	0	0	7	1	2	57	68	62	64	32	43	80	68	74
	Point	0	0	0	10	46	16	83	54	65	98	89	93	100	92	96
	Punch	0	0	0	0	0	0	36	62	46	70	59	64	98	98	98
	Shake	0	0	0	2	1	1	50	84	63	41	77	54	72	92	81
	Throw	34	8	13	0	0	0	26	65	37	43	51	47	93	70	80
	Wave	0	0	0	15	62	24	54	77	63	68	50	58	92	88	90
	DogC	Play-Ball	24	14	18	28	87	42	55	48	51	72	36	48	79	91
Car		75	22	34	16	64	26	75	71	73	32	47	38	88	66	75
Drink		39	16	23	76	19	30	43	76	55	11	32	16	58	56	57
Feed		29	33	31	1	0	0	33	40	36	34	38	36	21	21	21
Look-Left		33	28	30	0	0	0	23	18	20	19	11	14	43	34	38
Look-Right		78	12	21	61	18	28	20	5	8	9	1	2	63	39	48
Pet		31	14	19	97	56	71	44	35	39	25	27	26	90	85	87
Shake		65	9	16	2	0	0	19	9	12	74	27	40	68	58	63
Sniff		60	6	11	38	33	35	43	40	41	54	39	45	53	66	59
Walk		52	13	21	68	21	32	39	51	44	53	17	26	60	72	65

Table 10

Evaluation of the combination of features in the RMF. +: a feature subgroup is concatenated to the above set; -: a feature subgroup is removed from the above set; MDHF: motion direction histogram feature; MDHSF: motion direction histogram standard deviation feature; FTMAF: Fourier transform of motion across frames; MMHF: motion magnitude histogram feature; FTMPF: Fourier transform of grid motion per frame; VIF: vision-based inertial feature; \mathcal{F} : F_1 -score of the SVM classifier (%); Generally, improved performance is achieved when we combine GOFF subgroups (S.No 1-5) and VIF (S.No 6), and the performance starts to decline slowly when we remove features (S.No 7-11). This delucidates that all feature subgroups are necessary though they have different discriminative levels across the datasets.

S.No	IAR			BAR			JPL			DogC		
	Feature	\mathcal{F}	KNN									
1	MDHF	82	72	FTMAF	52	53	MMHF	62	63	FTMAF	42	46
2	+ MDHSF	85	74	+ MDHF	66	66	+FTMAF	67	65	+MDHSF	45	47
3	+ FTMAF	87	75	+ MMHF	71	69	+MDHF	72	67	+MDHF	46	47
4	+ MMHF	88	78	+ MDHSF	71	72	+MDHSF	78	68	+FTMPF	48	48
5	+ FTMPF	88	79	+ FTMPF	72	73	+FTMPF	79	68	+MMHF	51	50
6	+ VIF	88	79	+ VIF	80	78	+VIF	86	82	+VIF	61	59
7	- FTMPF	88	77	- FTMPF	79	77	-FTMPF	85	82	-MMHF	60	58
8	- MMHF	87	76	- MDHSF	79	74	-MDHSF	85	82	-FTMPF	58	58
9	- FTMAF	86	76	- MMHF	76	72	-MDHF	84	81	-MDHF	58	59
10	- MDHSF	84	72	- MDHF	72	66	-FTMAF	81	80	-MDHSF	57	57
11	- MDHF	57	48	- FTMAF	62	60	-MMHF	80	78	-FTMAF	48	47

Table 10 presents how the combination of feature subgroups improves system performance in the proposed method. The concatenation of GOFF subgroups (S.No 1-5), in accordance with their ranking in Fig. 19, and later with VIF (S.No 6) realizes the full implementation of RMF where the highest recognition performance is achieved in each dataset. VIF is highly discriminant as it improves the F_1 -score by 8%, 7% and 10% and the KNN accuracy by 5%, 14% and 9%, respec-

tively, in the BAR, JPL and DogCentric datasets. In order to re-evaluate the significance of each subgroup, we remove, one-by-one, the previously added GOFF subgroups (S.No 7-11), where a gradual performance reduction is experienced. The different ranking of the subgroups in different datasets (Fig. 19), the improvement of recognition performance when we concatenate them in S.No 1-6 and the gradual decline in S.No 7-11 disclose the importance of all feature subgroups

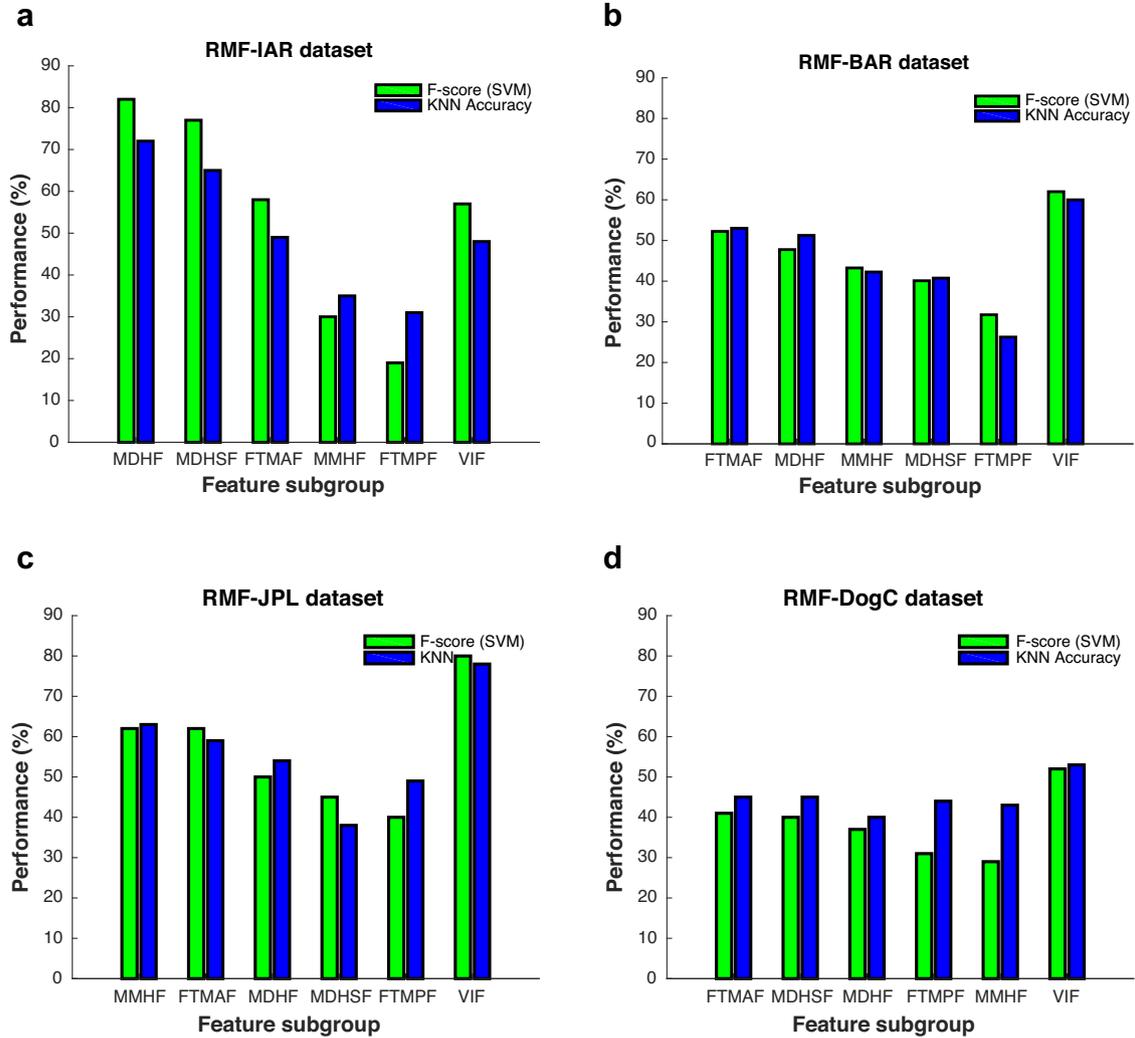


Fig. 19. Independent performance of GOFF subgroups (sorted by the F_1 -score of SVM outputs) and VIF of the proposed RMF in the (a) IAR and (b) BAR (c) JPL and (d) DogCentric datasets. MDHF: motion direction histogram feature; MDHSF: motion direction histogram standard deviation feature; FTMAF: Fourier transform of motion across frames; MMHF: motion magnitude histogram feature; FTMPF: Fourier transform of grid motion per frame; VIF: vision-based inertial feature. According to the variation among activities (Fig. 4), direction-based features top the ranking in IAR dataset whereas magnitude and frequency-based features become more discriminant in the BAR and JPL datasets. The different in the ranking of feature subgroups in different datasets reveals the importance of all subgroups for efficient encoding of motion information.

Table 11

Report on the independent performance (%) of VIF subgroups. **IAR**: indoor activity recognition; **BAR**: basketball activity recognition; **JPL**: JPL-interaction dataset; **DogC**: DogCentric dataset; \mathcal{F} : F_1 -score; **Min.**: minimum; **Max.**: maximum; **Med.**: median; **En.**: energy; **Kur.**: kurtosis; **Z-c.**: zero-crossing; **Std.**: standard deviation; **FF**: frequency-domain feature; **All**: concatenation of all feature subgroups in VIF.

Dataset	Measure	Min.	Max.	Med.	En.	Kur.	Z-c.	Mean	Std.	FF	All
IAR	\mathcal{F}	32	31	40	24	19	16	29	38	53	57
	KNN	34	31	42	30	23	27	32	40	44	48
BAR	\mathcal{F}	26	22	36	20	20	23	34	35	65	62
	KNN	32	30	43	27	20	23	31	40	56	60
JPL	\mathcal{F}	38	34	34	31	29	32	36	36	77	80
	KNN	40	33	39	22	41	41	39	44	66	78
DogC	\mathcal{F}	17	17	22	16	18	23	26	21	45	48
	KNN	27	23	36	18	29	30	32	23	45	47

in order to achieve the highest performance. We leave as a future work the use of a feature selection method to automatically build a shorter RMF vector from the GOFF and VIF subgroups. Independent performance evaluation of VIF subgroups in another experiment (Table 11) shows that the frequency-domain feature (FF) is the single best performing feature.

In order to measure the robustness of the methods, we artificially introduce a white Gaussian noise with different signal-to-noise ratio (SNR) values in the motion data. The motion implies the grid optical flow in both RMF and AP, whereas it refers to the pixel-wise displacement of matched interest points in MGRE. We apply the noise on MBH once the motion-based histograms were computed. Fig. 20 illustrates

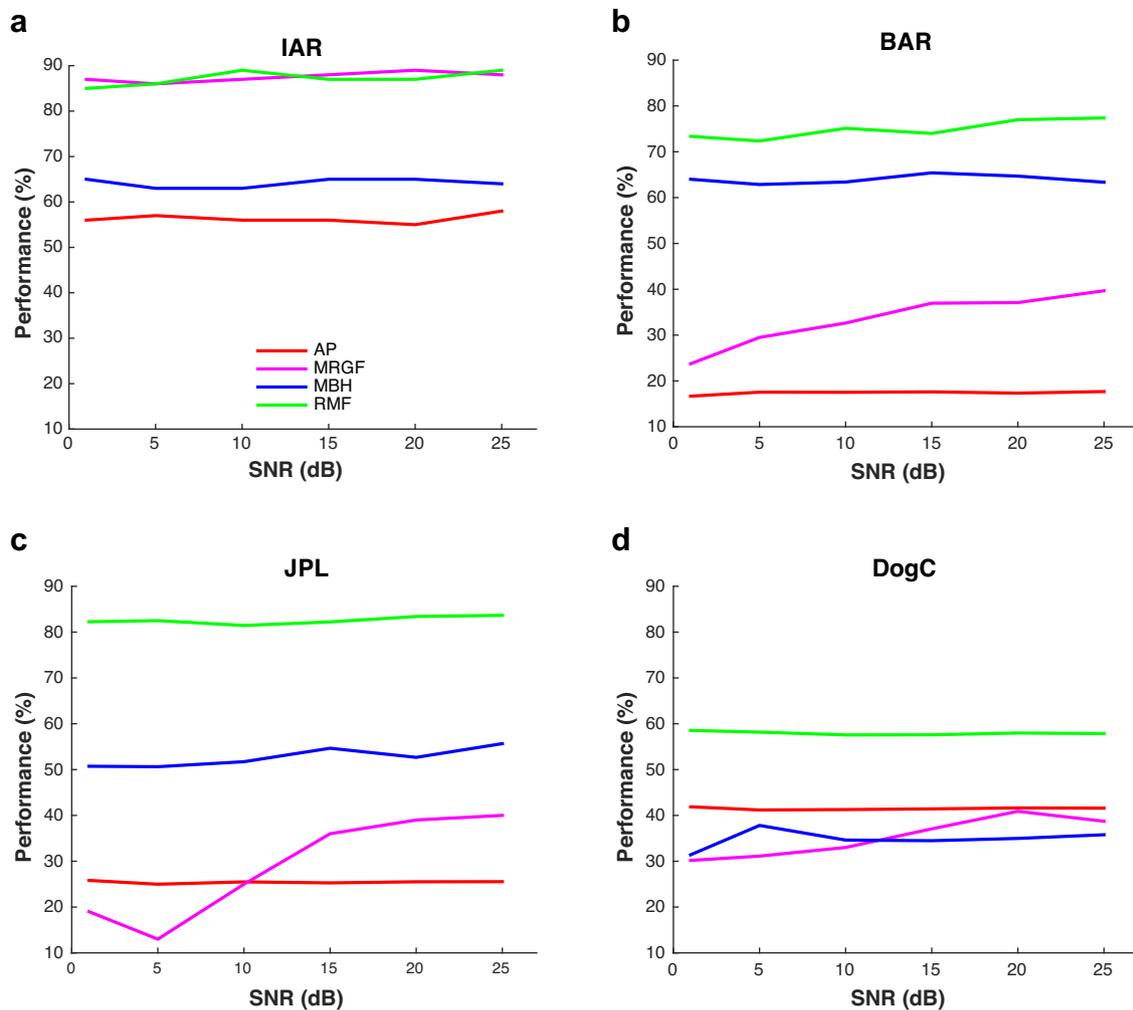


Fig. 20. Robustness analysis of the methods when a Gaussian noise, SNR values ranging from 1 dB to 25 dB, is introduced in the motion data. Significant robustness is observed for our proposed RMF, consistently across the datasets.

that RMF is the only one that achieves consistent stability for a range of SNR values across the four datasets.

We experiment further the robustness of RMF by testing on a new noisy dataset (Sitges) collected in streets with pedestrians (Fig. 21). A subject performs all the BAR activities except for *Dribble* and the replacement of *Shoot* with *Jump*. Some of the challenges introduced in this new dataset include highly dynamic occlusions by pedestrians, which might be both in similar and opposite directions to the direction of the user and a lack of illumination since the recording was performed around sunset opposite to the BAR dataset which was collected in the morning just after a sunrise. We train activity models using the BAR dataset and test them on the Sitges dataset. RMF achieved a performance of $\mathcal{F} = 56\%$ validated on SVM, higher than any of the other methods considered (Table 12). The SVM-based confusion matrix of the proposed method is shown in Fig. 22. Similarly to Fig. 18b, *Run* and *Jog* are hardly classified. However, misclassification of erratic samples, mainly, to *Left-right turn* and *Pivot* happens because of rotation-like motions introduced due to a large field-of-view of the camera and closer appearance of buildings in this dataset. In general, the results show that RMF has a strong potential to discriminate activities even in crowded environments from FPV if the models are trained properly.

In addition, we also test the sensitivity of our proposed method for manual variation of parameter settings described in Section 3.2. In

Table 12

Comparative performance (%) of the methods when they are trained on the BAR dataset and tested on the Sitges dataset, recorded in streets with pedestrians. \mathcal{F} : F_1 -score of SVM output. The proposed method surpasses the state-of-the-art motion features, MBH following closely.

	Methods				
	Baseline	AP	MRGF	MBH	RMF
\mathcal{F}	15	11	34	51	56
KNN	14	22	36	53	51

particular, we vary the parameters to be tuned in GOFF and VIF: direction histogram bins in MDHF and MDHSF (β_p), magnitude histogram bins in MMHF (β_m), frequency bands in FTMAF (N_f) and number of low frequency coefficients in FTMPF (N_s) and in VIF (N_c). We also measured mean and standard deviation for the variation of parameter settings and recognition performances. Table 13 depicts the stability of RMF for the manual variations of the parameter settings across the four datasets. SVM classifier results more stable outputs with F_1 -score variation ranging from $\sigma = 0.9$ in DogCentric to $\sigma = 2.2$ in JPL, in comparison with KNN that varies from $\sigma = 0.8$ in DogCentric to $\sigma = 3.9$ in the BAR dataset.

Table 13

Sensitivity analysis of the proposed method for variations of parameter settings in the IAR, BAR, JPL and DogCentric datasets; β_p : direction histogram bins in MDHF and MDHSF; β_m : magnitude histogram bins in MMHF; N_f : frequency bands in FTMAF; N_s in FTMPF and N_c in VIF: low frequency coefficients; \mathcal{F} : F_1 -score of SVM (%); μ : mean; σ : standard deviation.

Parameters						IAR		BAR		JPL		DogC	
	β_p	β_m	N_f	N_s	N_c	\mathcal{F}	KNN	\mathcal{F}	KNN	\mathcal{F}	KNN	\mathcal{F}	KNN
	36	15	25	25	10	88	78	80	78	86	82	61	58
	24	10	20	20	15	84	77	76	75	85	81	61	58
	16	5	15	15	20	83	75	75	70	86	83	62	57
	10	20	10	20	25	83	72	75	69	85	81	61	56
	18	25	30	10	5	85	76	77	78	82	79	60	58
	48	30	5	30	18	86	78	78	70	80	78	61	58
	30	5	35	5	8	85	76	77	76	84	80	59	58
μ	26.0	15.7	20.0	17.8	14.4	84.8	76.0	76.8	73.7	84.0	80.5	60.7	57.5
σ	13.1	9.7	10.8	8.5	7.1	1.7	2.1	1.7	3.9	2.2	1.7	0.9	0.8

a**b**

Fig. 21. Key-frames from the newly collected data to validate the flexibility of our method; (a) activities viewed from an external camera, (b) frames from first-person videos acquired by a chest-mounted wearable camera while a user performs the corresponding activity in (a). The activities from left to right are *Walk*, *Left-right turn*, *Jog* and *Jump*.

	RMF-Sitges									
	Bow	Def.	Jog	Lef.	Piv.	Run	Sho.	Sit.	Spr.	Walk
Bow	92									
Defend	31	31	12	12	15					
Jog		6	15	70	6					3
Left-right	3			85						3
Pivot					100					
Run			20	64	12	4				
Shoot							100			
Sit-stand			6		3			72		19
Sprint				30	10				60	
Walk		4		24	6				29	37

Fig. 22. Confusion matrix of the SVM-validated proposed method for the new dataset recorded in streets with pedestrians. Erratic samples are classified as *Left-right turn* activity dominantly and *Pivot* secondarily, because the high field-of-view of the wearable camera and closer appearance of the buildings introduce a sense of rotational motion in the FPV.

6. Conclusions

We designed a set of robust multi-dimensional motion features for first-person vision (FPV) based on optical flow and change of intensity centroid. Discriminant features were extracted and combined in the proposed feature set that incorporates motion magnitude, direction and periodic characteristics. To improve recog-

nition performance, we combined optical flow-based features with virtual inertial features extracted from the video. The proposed features were validated on the classification of ambulatory activities using our two datasets (we make them available to the research community) and further two public available interaction-based datasets. Results demonstrate that the proposed method outperforms state-of-the-art features, especially in classifying activities that contain complex ego-motions. The robustness to noise and stability under different parameter settings were also demonstrated by the proposed RMF representation. Finally, RMF outscored existing methods in more challenging environments unseen during training. As future work, we plan to apply efficient feature selection methods to reduce the feature dimension while keeping its discriminative, robustness and stability characteristics.

Acknowledgment

G. Abebe was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA Agency of the European Commission under EMJD ICE FPA no 2010–2012. A. Cavallaro acknowledges the support of the Artemis JU and the UK Technology Strategy Board through the COP-CAMS Project, under Grant 332913.

References

- [1] T. Shiratori, H.S. Park, L. Sigal, Y. Sheikh, J.K. Hodgins, Motion capture from body-mounted cameras, *ACM Trans. Gr. (TOG)* 30 (4) (2011) 31.
- [2] S. Bambach, A survey on recent advances of computer vision algorithms for ego-centric video (2015) arXiv:1501.02825.

- [3] Y. Bai, C. Li, Y. Yue, W. Jia, J. Li, Z.-H. Mao, M. Sun, Designing a wearable computer for lifestyle evaluation, in: Proceedings of the Northeast Bioengineering Conference (NEBEC), Philadelphia, USA, 2012, pp. 93–94.
- [4] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, K. Wood, Sensecam: A retrospective memory aid, in: Proceedings of the International Conference on Ubiquitous Computing (UbiComp), California, USA, 2006, pp. 177–193.
- [5] S. Hodges, E. Berry, K. Wood, Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory, *Memory* 19 (7) (2011) 685–696.
- [6] A.R. Doherty, A.F. Smeaton, Automatically segmenting lifelog data into events, in: Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Klagenfurt, Austria, 2008, pp. 20–23.
- [7] N. Caprani, N.E. O'Connor, C. Gurrin, Investigating older and younger peoples' motivations for lifelogging with wearable cameras, in: Proceedings of the International Symposium on Technology and Society (ISTAS), Toronto, Canada, 2013, pp. 32–41.
- [8] K. Zhan, S. Faux, F. Ramos, Multi-scale conditional random fields for first-person activity recognition on elders and disabled patients, *Pervasive Mobile Comput.* 16, Part B (2015) 251–267.
- [9] K. Zhan, S. Faux, F. Ramos, Multi-scale conditional random fields for first-person activity recognition, in: Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, Hungary, 2014, pp. 51–59.
- [10] Y. Nam, S. Rho, C. Lee, Physical activity recognition using multiple sensors embedded in a wearable device, *ACM Trans. Embed. Comput. Syst.* 12 (2) (2013) 26:1–26:14.
- [11] K. Zhan, F. Ramos, S. Faux, Activity recognition from a wearable camera, in: Proceedings of the IEEE International Conference on Control Automation Robotics & Vision (ICARCV), Guangzhou, China, 2012, pp. 365–370.
- [12] H. Zhang, L. Li, W. Jia, J.D. Fernstrom, R.J. Scabassi, Z.-H. Mao, M. Sun, Physical activity recognition based on motion in images acquired by a wearable camera, *Neurocomputing* 74 (12) (2011) 2184–2192.
- [13] H. Zhang, L. Li, W. Jia, J.D. Fernstrom, R.J. Scabassi, M. Sun, Recognizing physical activity from ego-motion of a camera, in: Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society (EMBC), Buenos Aires, Argentina, 2010, pp. 5569–5572.
- [14] K.M. Kitani, T. Okabe, Y. Sato, A. Sugimoto, Fast unsupervised ego-action learning for first-person sports videos, in: Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Colorado, USA, 2011, pp. 3241–3248.
- [15] Y. Cho, Y. Nam, Y.-J. Choi, W.-D. Cho, Smartbuckle: human activity recognition using a 3-axis accelerometer and a wearable camera, in: Proceedings of the International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments, Colorado, USA, 2008, pp. 1–3.
- [16] A. Fathi, Learning Descriptive Models of Objects and Activities from Egocentric Video, Georgia Institute of Technology, 2013 (Ph.D. thesis).
- [17] K. Matsuo, K. Yamada, S. Ueno, S. Naito, An attention-based activity recognition for egocentric video, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, USA, 2014, pp. 565–570.
- [18] S. Sundaram, W.W.M. Cuevas, High level activity recognition using low resolution wearable vision, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Miami, USA, 2009, pp. 25–32.
- [19] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Providence, USA, 2012, pp. 2847–2854.
- [20] M.S. Ryoo, L. Matthies, First-person activity recognition: What are they doing to me? in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), Portland, USA, 2013, pp. 2730–2737.
- [21] S. Narayan, M.S. Kankanhalli, K.R. Ramakrishnan, Action and interaction recognition in first-person videos, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, USA, 2014, pp. 526–532.
- [22] M. Muir, C. Conati, Understanding student attention to adaptive hints with eye-tracking, in: Advances in User Modeling, Girona, Spain, 2012, pp. 148–160.
- [23] M. Kumar, T. Garfinkel, D. Boneh, T. Winograd, Reducing shoulder-surfing by using gaze-based password entry, in: Proceedings of the Symposium on Usable Privacy and Security (SOUPS), Pittsburgh, USA, 2007, pp. 13–19.
- [24] L. Piccardi, B. Noris, O. Barbey, A. Billard, G. Schiavone, F. Keller, C. von Hofsten, Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children, in: Proceedings of the IEEE International Symposium on Robot and Human interactive Communication, Jeju, South Korea, 2007, pp. 594–598.
- [25] T. Toyama, T. Kieninger, F. Shafait, A. Dengel, Gaze guided object recognition using a head-mounted eye tracker, in: Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Santa Barbara, USA, 2012, pp. 91–98.
- [26] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, *IEEE Commun. Surv. Tutor.* 15 (3) (2013) 1192–1209.
- [27] P.L. Rosin, Measuring corner properties, *Comput. Vis. Image Underst. (CVIU)* 73 (2) (1999) 291–307.
- [28] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: an efficient alternative to SIFT or SURF, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 2564–2571.
- [29] B.P. Clarkson, K. Mase, A. Pentland, Recognizing user context via wearable sensors, in: Proceedings of the International Symposium on Wearable Computers (ISWC), Atlanta, USA, 2000, p. 69.
- [30] Y. Iwashita, A. Takamine, R. Kurazume, M. Ryoo, First-person animal activity recognition from egocentric videos, in: Proceedings of the International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 2014, pp. 4310–4315.
- [31] P.H. Torr, A. Zisserman, Feature based methods for structure and motion estimation, in: Proceedings of the International Workshop on Vision Algorithms: Theory and Practice, Corfu, Greece, 1999, pp. 278–294.
- [32] D. Scaramuzza, F. Fraundorfer, Visual odometry [tutorial], *IEEE Robot. Autom. Mag.* 18 (4) (2011) 80–92.
- [33] C. Forster, M. Pizzoli, D. Scaramuzza, Svo: Fast semi-direct monocular visual odometry, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 2014, pp. 15–22.
- [34] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), *Comput. Vis. Image Underst. (CVIU)* 110 (3) (2008) 346–359.
- [35] J. Shi, C. Tomasi, Good features to track, in: Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 1994, pp. 593–600.
- [36] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* 24 (6) (1981) 381–395.
- [37] M. Irani, P. Anandan, About direct methods, in: Proceedings of the International Workshop on Vision Algorithms: Theory and Practice, Corfu, Greece, 1999, pp. 267–277.
- [38] R.F. Guerreiro, P.M. Aguiar, Global motion estimation: feature-based, featureless, or both? in: Proceedings of the International Conference on Image Analysis and Recognition (ICIAR), Varzim, Portugal, 2006, pp. 721–730.
- [39] N. Ravi, N. Dandekar, P. Mysore, M.L. Littman, Activity recognition from accelerometer data, in: Proceedings of the International Conference on Innovative Applications of Artificial Intelligence (IAAI), Pittsburgh, Pennsylvania, 2005, pp. 1541–1546.
- [40] L. Bao, S.S. Intille, Activity recognition from user-annotated acceleration data, in: Proceedings of the International Conference on Pervasive Computing, 2004, pp. 1–17.
- [41] S.-W. Lee, K. Mase, Activity and location recognition using wearable sensors, *IEEE Pervasive Comput.* 1 (3) (2002) 24–32.
- [42] F.R. Allen, E. Ambikairajah, N.H. Lovell, B.G. Celler, Classification of a known sequence of motions and postures from accelerometry data using adapted gaussian mixture models, *Physiol. Meas.* 27 (10) (2006) 935.
- [43] U. Maurer, A. Smaligic, D. Siewiorek, M. Deisher, Activity recognition and monitoring using multiple sensors on different body positions, in: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, Cambridge, USA, 2006, pp. 4pp.–116.
- [44] D. Karantonis, M. Narayanan, M. Mathie, N. Lovell, B. Celler, Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring, *IEEE Trans. Inf. Technol. Biomed.* 10 (1) (2006) 156–167.
- [45] O.D. Lara, A.J. Pérez, M.A. Labrador, J.D. Posada, Centinela: A human activity recognition system based on acceleration and vital sign data, *Pervasive Mobile Comput.* 8 (5) (2012) 717–729.
- [46] P. Lukowicz, J.A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, T. Starner, Recognizing workshop activity using body worn microphones and accelerometers, in: Pervasive Computing, Vienna, Austria, 2004, pp. 18–32.
- [47] M. Ermes, J. Parkka, L. Cluitmans, Advancing from offline to online activity recognition with wearable sensors, in: Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society (EMBS), Vancouver, Canada, 2008, pp. 4451–4454.
- [48] J.L.R. Ortiz, Smartphone-Based Human Activity Recognition, Springer, 2015.
- [49] D. Rodriguez-Martin, A. Sama, C. Perez-Lopez, A. Catala, J. Cabestany, A. Rodriguez-Moliner, SVM-based posture identification with a single waist-located triaxial accelerometer, *Expert Syst. Appl.* 40 (18) (2013) 7203–7211.
- [50] Z. He, L. Jin, Activity recognition from acceleration data based on discrete cosine transform and SVM, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), San Antonio, USA, 2009, pp. 5041–5044.
- [51] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, *ACM Comput. Surv. (CSUR)* 46 (3) (2014) 1–33.
- [52] B.K.P. Horn, B.G. Schunck, Determining optical flow, *Artif. Intell.* 17 (13) (1981) 185–203.
- [53] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, Canada, 1981, pp. 674–679.
- [54] A. Mannini, A.M. Sabatini, Machine learning methods for classifying human physical activity from on-body accelerometers, *Sensors* 10 (2) (2010) 1154–1175.
- [55] A. Nagasaka, T. Miyatake, Real-time video mosaics using luminance-projection correlation, *Trans. IEICE* (1999) 1572–1580.
- [56] K. Uehara, M. Amano, Y. Arikai, M. Kumano, Video shooting navigation system by real-time useful shot discrimination based on video grammar, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), vol.1, Taipei, Taiwan, 2004, pp. 583–586.