Visual features for ego-centric activity recognition: A survey

Girmaw Abebe Tadesse
Institute of Biomedical Engineering
University of Oxford
Oxford, United Kingdom
girmaw.abebe@eng.ox.ac.uk

Andrea Cavallaro
Centre for Intelligent Sensing
Queen Mary University of London
London, United Kingdom
a.cavallaro@qmul.ac.uk

ABSTRACT

Wearable cameras, which are becoming common mobile sensing platforms to capture the environment surrounding a person, can also be used to infer activities of the wearer. In this paper we critically discuss features for ego-centric activity recognition using videos. These features can be learned from data or designed to effectively encode motion magnitude, direction and other dynamics. Features can be derived from optical flow, from the displacement of key-points or the intensity centroid. We also discuss how features are effectively filtered and fused for specific tasks. Features presented in this paper can also be applied to other wearable systems that use accelerometer and gyroscope data.

CCS CONCEPTS

•General and reference → Surveys and overviews; •Computing methodologies → Computer vision; Activity recognition and understanding; •Human-centered computing → *Ubiquitous and mobile computing;*

KEYWORDS

First-person vision, wearable camera, activity recognition, survey

ACM Reference format:
Girmaw Abebe Tadesse and Andrea Cavallaro. 2018. Visual features for ego-centric activity recognition: A survey. In *Proceedings of 4th ACM Workshop on wearable systems and applications , Munich, Germany, June 10, 2018 (WearSys'18),* 6 pages.
DOI: 10.1145/3211960.3211978

1 INTRODUCTION

Sensing the surrounding environment with a small, high-quality and efficient wearable camera, also known as egocentric or first-person vision (FPV) [26], has been made possible by fast progress in embedded technology [16]. FPV can support lifelogging [15, 21], augmented reality [6, 33], and activity recording [19, 51]. Moreover, egocentric activity recognition has multiple applications that include health monitoring [53, 54] and sport activities analysis and segmentation [4, 27].

Surveys on wearable camera systems [8, 10, 24, 26, 35, 60] are generally wide in scope and include multiple application domains, i.e. object, action and activity recognition [8, 10, 35] as well as lifelogging and video summarization [24, 60]. Kanade and Hebert [26] addressed the challenges in developing environment-aware wearable camera systems for the localization and recognition of object, people and 3D scene structure. However, no low-level features to encode activities were covered. Bambach [8] reviewed algorithms for lifelogging video summarisation and the recognition of object-driven activities such as cooking. Betancourt et al. [10] covered tasks such as physical scene reconstruction and interaction detection without discussing in details key low-level features. Nguyen et al. [35] hierarchically structured activities of daily living as motion-level events (e.g. hand detection), basic actions (e.g. closing a jar) and complex activities (e.g. making a cup of coffee) but the specific features employed for the classification were not described. Zhou and Gurrin [60] and Jacquemard et al. [24] focused, respectively, on the evaluation of devices and ethical challenges in lifelogging. The use of different wearable sensors for activity detection and classification was also recently reviewed [11, 13].

However, no survey has yet focused on details of features for the classification of user's activities using wearable camera systems. In this paper, we review four main categories of motion features (see Table 1) in FPV for recognition of the wearer's activity. The majority of features are still handcrafted and designed to encode salient characteristics of the motion data (Fig. 1). Apparent motion can be estimated using optical flow, virtual-inertial data or the displacement of keypoints. We also describe the extraction of features automatically learned from data using deep neural networks. Deep architectures learn high-level representations of the input data that help generalize across different classification challenges.

This paper is organized as follows. Section 2 describes features extracted form optical flow data. Section 3 presents keypoint-based features where motion is estimated from the temporal displacement of keypoints. Recent trends to extract virtual-inertial features from video are described in Section 4. Section 5 reviews automatic learning of motion features from data. Finally, Section 6 concludes the review and outlines future directions.

2 OPTICAL FLOW-BASED FEATURES

Optical flow is the main source of motion features for video-based activity recognition [4, 37, 38, 54]. Optical flow can be derived using direct motion estimation technique [23] to achieve sub-pixel accuracy. A grid representation of the optical flow is often preferred to a dense representation in order to avoid redundancy in the assumption of global motion dominance in FPV [4, 37, 54, 58].

We can group optical flow-based features into three categories: raw grid, direction and/or magnitude histogram and frequency-domain features. *Raw grid features* include grid representation and

WearSys'18, Munich, Germany
978-1-4503-5842-2/18/06…$15.00
DOI: 10.1145/3211960.3211978

Table 1: Common feature extraction and motion filtering techniques used in the state-of-the-art. Optical-flow is most frequently used due to its sub-pixel accuracy and direct motion estimation technique. RANSAC: random sample consensus.

			[3]	[1]	[4]	[43]	[38]	[37]	[54]	[53]	[56]	[27]	[58]	[34]	[55]	[59]
Features	Optical flow-based	Raw grid feature						✓	✓	✓	✓			✓	✓	
		Grid direction histogram		✓	✓	✓						✓				
		Grid magnitude histogram		✓	✓							✓				
		Grid gradient histogram				✓										
		Grid frequency feature		✓	✓							✓				
	Keypoint-based	Direction histogram											✓			✓
	Virtual-inertial-based	Intensity centroid		✓	✓											
		Average grid		✓												
	Learned	Pooled deep-appearance feature	✓			✓										
		Deep motion feature	✓				✓									
	Filtering	Thresholding						✓					✓			✓
		RANSAC-based filtering										✓	✓		✓	✓
		Gaussian smoothing	✓	✓	✓			✓								
		Average pooling							✓	✓	✓				✓	
	Early fusion		✓	✓	✓	✓		✓				✓			✓	

the concatenation of horizontal and vertical grid components [53–56]. Poleg et al. [37] used the *radial projection response* of grid optical flow vectors to discriminate moving from stationary camera wearers. Similarly, hard-coded rules on grid vector direction (θ) are employed in [34] to classify activities, e.g. *Left-turn* satisfies $0^\circ < \theta < 90^\circ$ or $270^\circ < \theta < 360^\circ$. *Raw grid features* have limited discriminative capabilities as specific motion characteristics (e.g. magnitude) are not exploited to the required level to achieve a robust motion feature with a compact representation.

Direction (see Fig. 1 (b)) and *magnitude* (see Fig. 1 (c)) histograms of the flow vector can provide more discriminant features [4, 27]. Motion magnitude and direction components are generally exploited separately to increase the discrimination. For example, *Sit-down* and *Stand-up* can be distinguished by exploiting their direction patterns, whereas magnitude information helps differentiate *Walk* and *Sprint*. A histogram is a compact representation of the direction and/or magnitude components of the grid flow data [4, 27, 43]. The histogram might be applied using independent direction and magnitude bins [4], joint spatial and direction bins [43], or joint magnitude, direction and magnitude variance bins [27]. The inclusion of spatial bins [43] is comparatively less effective since multiple motion-driven activities can be performed in similar environment settings. In addition, Ryoo et al. [43] employed motion boundary histogram (MBH) as one of the multiple motion features from optical flow data that compensates the camera motion. MBH is obtained by applying a spatial derivative on the horizontal and vertical optical flow components separately, followed by a magnitude-weighted histogram of motion direction [52].

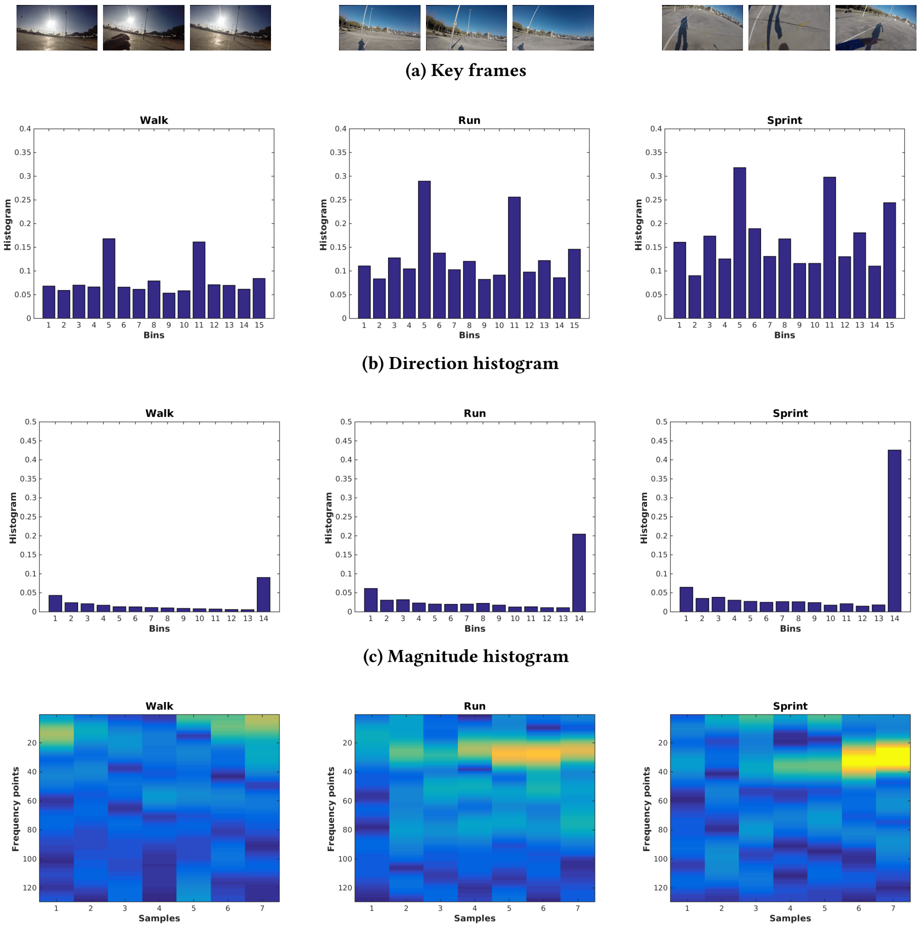
Frequency-domain features exploits low-level motion dynamics, which helps distinguish activities with similar direction patterns, e.g. *Sprint* and *Run* (see Fig. 1 (c)) as the latter involves less frequent changes in motion dynamics [4]. Kitani et al. [27] extracted frequency-domain features from the horizontal and vertical grid components independently, whereas the features were extracted on the direction and magnitude equivalents of the grid components in [1, 4]. The frequency-domain features can be represented by grouping the Fourier responses into equally spaced bands [4] or by selecting the low-frequency coefficients [1, 27]. Though the low frequency coefficients are robust to noise, the representation does not include the full spectrum characteristics. Similarly to the number of magnitude and direction bins for the histogram representations, the number of frequency bands needs to be carefully selected in order to avoid under or over-quantization.

Filtering is commonly applied to discard falsely matched descriptors [4, 54, 56, 58, 59]. Common filtering techniques include *thresholding* [37, 58, 59], *random sample consensus (RANSAC)* [17, 27, 55, 58, 59], *Gaussian smoothing* [4, 37] and *temporal averaging [53–56]*. In thresholding, motion vectors whose magnitude is smaller than a threshold are removed [37, 58, 59]. RANSAC can be employed to discard outliers of optical flow vectors [27, 55] or matched descriptors [58, 59]. Gaussian smoothing can also be applied on the whole motion data when there is a high variance [4, 37]. Average pooling of temporally adjacent grid vectors can also improve recognition performance [53–56].

Though optical flow might pose relatively higher computational cost, it provides sub-pixel accuracy and plausible approaches, e.g. Horn-Schunck [22] could be applied in-cases of less/no texture regions. Besides, sparse optical flow computation can ease computation as global motion often dominates local ones due to full- or upper-boy motion of the wearer. Though different discriminative characteristics could be extracted from optical flow, e.g. from its direction, magnitude, and dynamics; multiple existing methods tend to focus on encoding only a specific subset of these characteristics. Generally, optical flow is a commonly used motion information but all available discriminative details need to be effectively encoded for activity recognition in FPV.

3 KEYPOINT-BASED FEATURES

Spatial change of keypoints across frames can also be used to deduce apparent motion. According to the type of spatial structure of the interest points, keypoint detection can be *blob-based* or *corner-based*. Examples of blob-based detectors include scale-invariant feature transform [30], speeded-up robust features [9] and centre surround extremas [5]. Examples of corner-based detectors include features from accelerated segment test [41], adaptive and



(a) Key frames

(b) Direction histogram

(c) Magnitude histogram

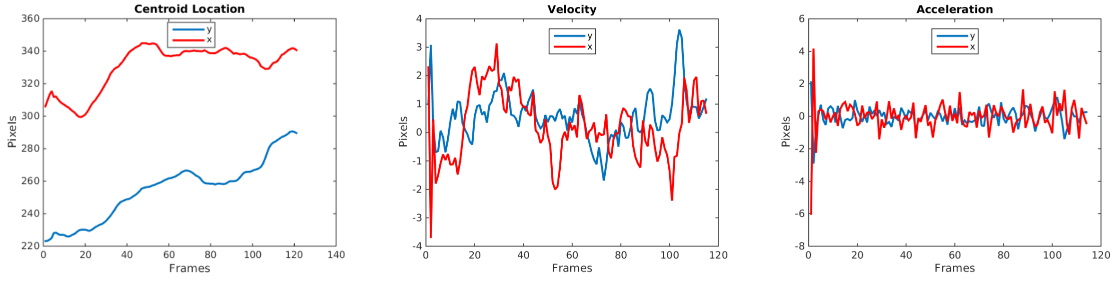
(d) Direction spectrogram

Figure 1: Examples of features derived from the optical flow to encode different motion characteristics of *Walk*, *Run* and *Sprint*. (a) Key frames of the activity in the corresponding column; (b) direction histogram representations; (c), (d) the activities can be easily discriminated using the magnitude histogram and the frequency-domain analysis of the direction information, respectively.

generic corner detection based on the accelerated segment test [32] and binary robust invariant scalable keypoints [29]. After a keypoint is detected, its neighbourhood is described using a *binary* or *non-binary* descriptor with characteristics such as invariant to rotation [18, 44, 49]. Using a binary descriptor makes matching computationally easier since Euclidean distance can be replaced by a Hamming distance that can be calculated using a bitwise XOR operation [7, 29, 42].

Zhang et al. [59] proposed a keypoint-based feature, inspired by the earlier work of Shi and Tomasi [46]. The matched keypoints were further refined by uniqueness (one-to-one correspondence) and epipolar constraints [20]. The frame motion was estimated as a set of displacement vectors between matched descriptor pairs. The direction of each displacement vector that satisfied a magnitude threshold was quantized using a histogram representation. The work was later upgraded to achieve multi-resolution detection of interest points in [58]. Average standard deviation [59] and combined standard deviation [58] of the histogram representation were used in order to include temporal characteristics in the feature space, which resulted in enhanced classification accuracy. Since Zhang et al. [58, 59] did not exploit the magnitude information and encode the dynamics in-detail, their recognition performances are often inferior to more advanced features that exploit those characteristics [4, 27].

Generally, keypoint-based methods are computationally efficient and can handle large displacements. Particularly, binary descriptors offer faster matching of keypoints, which is useful in resource-limited platforms such as wearable systems. However, their usefulness is reduced in poorly textured first-person videos, which are often blurred due to high egomotion induced by the camera wearer.



(a) (b) (c)

Figure 2: A step-by-step visualization of virtual-inertial data extraction from an exemplar *Walk* video segment ($\approx$ $4s$) with 640×480 resolution at $30fps$. (a) The intensity centroid tracked across frames; (b), (c) the velocity and acceleration vectors extracted using consecutive temporal derivatives, respectively.

4 VIRTUAL-INERTIAL FEATURES

In addition to optical flow and keypoint displacement, features can also be derived from the apparent motion encoded as virtual-inertial data, which contain velocity and acceleration vectors generated from the video (see Fig. 2). Virtual-inertial features provide inertial characteristics without using the actual inertial sensors, and thus avoid synchronization issues. The inertial data can be encoded from the movement of *intensity centroid* (see Fig. 2 (a)). The intensity centroid of an image, which is equivalent to the mass-centre of a rigid object in physics, is derived from the first-order image moments, which are computed as weighted averages of the whole intensity values [12, 40, 42]. The first-order derivative on the intensity centroids of successive frames gives instantaneous velocity abstraction (see Fig. 2 (b)). Corresponding acceleration vectors are generated by applying another temporal derivative on the velocity vectors (see Fig. 2 (c)). The velocity and acceleration vectors along with their magnitude components provide the complete set of the virtual-inertial data. Virtual-inertial features are extracted from the these velocity and acceleration components similarly to the extraction of the state-of-the-art inertial features from accelerometer data [4].

The inertial features can be extracted in time and frequency domains. The common time-domain inertial features include *zero-crossing*, *minimum*, *maximum*, *median*, *energy*, *kurtosis*, *mean* and *standard deviation* [4, 28, 36, 39, 54]. Zero-crossing measures the oscillatory behaviour of a vector in reference to zero magnitude value. Note that zero-crossing is not applied on magnitude vectors; however, the same intuition can be extracted in a reference to a non-zero threshold value. Kurtosis quantifies whether the distribution of an inertial vector is heavy-tailed or light-tailed with respect to a Gaussian distribution. A high kurtosis represents a heavy tail in the distribution, which signals a high probability of outliers [4]. Due to its high order definition, kurtosis is sensitive to noise. However, its ensemble along with other features improves the discriminating potential [4, 54]. Virtual inertial features can also be extracted from the mean optical flow vectors such that the horizontal and vertical components across frames constitute additional time-series vectors [1].

5 LEARNED FEATURES

Motion features can be learned from optical flow [31, 38, 48, 57] or pooled from deep appearance descriptors [1, 3, 31, 43] exploiting deep neural networks that automate feature engineering using successive layers of the neural networks.

Convolutional neural networks (CNNs) have been successful in learning high-level appearance features [25, 45]. The temporal pooling of frame-level appearance features reflect the variation of appearance information. Ryoo et al. [43] proposed different pooling operations, which treat each descriptor across frames as time-series data.

Summation and maximum pooling on the raw appearance features are less effective to encode the temporal variations as they do not show how a feature element changes over time. For this purpose, a *time-series gradient* pooling was proposed to encode the short and long temporal variations by applying first-order temporal derivative on each descriptor element [43].

The summation and histograms of positive and negative gradients can be applied to encode the variation. Comparatively, the gradient histogram representation describes the short-term variation more effectively since its score depends only on the sign of the gradients. The discriminative capacity of the feature space can be improved by encoding a more detailed temporal characteristics using frequency-domain analysis. This technique could also be applied to simple appearance descriptors such as the histogram of oriented gradients [43]. In addition to the temporal variation of the appearance, static appearance information can be useful when activities are correlated with certain environmental settings, e.g. *Go-upstairs/downstairs* involves staircases [43, 52].

Poleg et al. [38] proposed a compact CNN taking a sparse grid volume as input and learned motion features that are demonstrated to outperform the handcrafted features in their previous work [37]. The network was derived from the temporal component of an existing network [47], and it was designed with a 3D convolutional layer followed by a 3D pooling to handle the 3D input data [50]. A 2D convolution layer applied afterwards eliminates the temporal dependency early in the network. Finally, two fully connected convolutional layers are stacked followed by the softmax layer. Though the deep motion features are shown to be transferable

Table 2: Summary of advantages and disadvantages of main feature groups employed in the state of the art.

Feature groups	Advantages	Disadvantages
Optical flow-based	Direct motion estimation Sub-pixel accuracy	Computationally expensive w.r.t keypoint-based methods
Keypoint-based	Easier than optical flow-based Can encode both appearance and motion	Challenging when there is no texture
Virtual-inertial	Provides inertial features without using the actual sensor	Less discriminative when a user is stationary
Learned	Avoids feature engineering Improved performance with enough data Transferable knowledge	Requires more data for training Computational expensive than all other methods Features are less interpretable

across datasets of similar nature [38], the interpretation of the knowledge learned at different layers requires further study.

Abebe et al. [3] employed spectrogram-based representation of short-term motion feature in FPV that is later normalized into RGB-like images. This approach helps to exploit existing CNN frameworks that are pretrained on natural images, e.g. ImageNet [14], for motion features encoding, i.e. transfer learning. The spectrogram representation allows simpler networks to learn motion features using 2D convolutions compared to 3D convolution-based approaches [50]. This approach has been validated further in other mobile and wearable sensory data such as inertial time-series from accelerometer and gyroscope [2]. The spectrogram of multiple axial motion components were stacked as an image to achieve cross-domain knowledge transfer using vision-based deep neural networks.

Generally, though deep learning is posed to become the standard in feature encoding, deep neural networks are often treated as black-boxes, and the interpretation of learned motion features is still limited. Deep learning offers transferability of knowledge across different tasks [2] but it still requires further research to significantly outperform handcrafting approaches in FPV.

6 CONCLUSION

We reviewed features for activity recognition in videos captured by wearable cameras. The features are designed to exploit available motion peculiarities, such as magnitude, direction and dynamics. Optical flow-based techniques are in general more robust as they can estimate motion in the presence of weak texture and motion blur. Table 2 summarises the FPV features discussed in this paper.

Activity recognition often requires the use of multiple discriminative features [4, 27, 37, 58, 59] or an additional modality, such as inertial sensors, to complement visual features [34, 53, 54]. In these cases, early feature fusion can be performed prior to classification [4, 27, 37, 43, 55]. Feature-level fusion has to be carefully applied on multiple feature groups with equivalent scales and dimensions, otherwise the discriminative potential of a lower-dimensional feature group or a feature group with smaller scale could be undermined as the result of the feature-level fusion which is often implemented using concatenations.

With the growing size of publicly available datasets and the success of deep networks across different application domains, high-level learned features are expected to outperform hand-crafted features robustly. While learned features pose to overtake handcrafted features, common strategies on the architecture of the network and the integration of appearance and motion information still requires further study. Knowledge transfer offers a promising potential to the future development of unsupervised or weakly supervised tasks as labelling the growing size of FPV data becomes difficult.

7 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.

REFERENCES

[1] Girmaw Abebe and Andrea Cavallaro. 2017. Hierarchical modeling for first-person vision activity recognition. *Neurocomputing* 267 (December 2017), 362–377.

[2] Girmaw Abebe and Andrea Cavallaro. 2017. Inertial-Vision: cross-domain knowledge transfer for wearable sensors. In *Proc. of International Conference on Computer Vision (ICCV)*. Venice, Italy, 1392–1400.

[3] Girmaw Abebe and Andrea Cavallaro. 2017. A long short-term memory convolutional neural network for first-person vision activity recognition. In *Proc. of International Conference on Computer Vision (ICCV)*. Venice, Italy, 1339–1346.

[4] Girmaw Abebe, Andrea Cavallaro, and Xavier Parra. 2016. Robust multi-dimensional motion features for first-person vision activity recognition. *Computer Vision and Image Understanding (CVIU)* 149 (2016), 229 – 248.

[5] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. 2008. CenSurE: Center surround extremas for realtime feature detection and matching. In *Proc. of European Conference on Computer Vision (ECCV)*.

[6] Chris Aimone, James Fung, and Steve Mann. 2003. An EyeTap video-based featureless projective motion estimation assisted by gyroscopic tracking for wearable computer mediated reality. *Personal and Ubiquitous Computing* 7 (2003), 236–248.

[7] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. 2012. FREAK: Fast retina keypoint. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*.

[8] Sven Bambach. 2015. A Survey on Recent Advances of Computer Vision Algorithms for Egocentric Video. (2015).

[9] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)* 110, 3 (December 2008), 346–359.

[10] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. 2015. The Evolution of First Person Vision Methods: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 5 (2015), 744–760.

[11] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2017. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications* 76, 3 (01 Feb 2017), 4405–4425.

[12] Brian P Clarkson, Kenji Mase, and Alex Pentland. 2000. Recognizing user context via wearable sensors. In *Proc. of International Symposium on Wearable Computers (ISWC)*. Atlanta, USA, 69.

[13] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. 2017. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal* 17, 2 (2017), 386–403.

[14] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami, USA, 248–255.

[15] Aiden Doherty, Paul Kelly, and Charlie Foster. 2013. Wearable Cameras: Identifying Healthy Transportation Choices. *Pervasive Computing* 12 (2013), 44–47.

[16] Alireza Fathi. 2013. *Learning Descriptive Models of Objects and Activities from Egocentric Video.* Ph.D. Dissertation. Georgia Institute of Technology.

[17] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (June 1981), 381–395.

[18] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. 2014. SVO: Fast semi-direct monocular visual odometry. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*. Hong Kong, China, 15–22.

[19] Wendy Glauser. 2013. Doctors among early adopters of Google Glass. *Canadian Medical Association Journal* (2013), 109.

[20] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

[21] Steve Hodges, Emma Berry, and Ken Wood. 2011. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory* 19, 7 (October 2011), 685–696.

[22] Berthold K. P. Horn and Brian G. Schunck. 1981. Determining optical flow. *Artificial Intelligence* 17, 1-3 (1981), 185 – 203.

[23] Michal Irani and P Anandan. 1999. About direct methods. In *Proc. of International Workshop on Vision Algorithms: Theory and Practice*. Corfu, Greece, 267–277.

[24] Tim Jacquemard, Peter Novitzky, Fiachra OfiBrolcháin, Alan F Smeaton, and Bert Gordijn. 2014. Challenges and opportunities of lifelog technologies: a literature review and critical analysis. *Science and engineering ethics* 20, 2 (2014), 379–409.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of ACM International Conference on Multimedia*. Florida, USA, 675–678.

[26] T. Kanade and M. Hebert. 2012. First-Person Vision. *Pro. of IEEE* 100, 8 (2012), 2442–2453.

[27] Kris Makoto Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*. Colorado, USA, 3241–3248.

[28] Oscar D. Lara and Miguel A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials* 15, 3 (November 2013), 1192–1209.

[29] Stefan Leutenegger, Margarita Chli, and Roland Yves Siegwart. 2011. BRISK: Binary robust invariant scalable keypoints. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. Barcelona, Spain, 2548–2555.

[30] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 12 (2004), 91–110.

[31] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. 2016. Going Deeper into First-Person Activity Recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA, 1894–1903.

[32] Elmar Mair, Gregory D Hager, Darius Burschka, Michael Suppa, and Gerhard Hirzinger. 2010. Adaptive and generic corner detection based on the accelerated segment test. In *Proc. of European Conference on Computer Vision (ECCV)*. Crete, Greece, 183–196.

[33] Steve Mann, James Fung, and Eric Moncrieff. 1999. Eyetap technology for wireless electronic news gathering. *ACM SIGMOBILE Mobile Computing and Communications Review* 3 (1999), 19–26.

[34] Yunyoung Nam, Seungmin Rho, and Chulung Lee. 2013. Physical activity recognition using multiple sensors embedded in a wearable device. *ACM Transactions on Embedded Computing Systems* 12, 2 (February 2013), 26:1–26:14.

[35] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, and others. 2016. Recognition of Activities of Daily Living with Egocentric Vision: A Review. *Sensors* 16, 1 (2016), 72.

[36] Jorge Luis Reyes Ortiz. 2015. *Smartphone-Based Human Activity Recognition*. Springer.

[37] Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal segmentation of egocentric videos. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*. Ohio, USA, 2537–2544.

[38] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact CNN for indexing egocentric videos. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*. New York, USA, 1–9.

[39] Daniel Rodriguez-Martin, Albert Sama, Carlos Perez-Lopez, Andreu Catala, Joan Cabestany, and Alejandro Rodriguez-Molinero. 2013. SVM-based posture identification with a single waist-located triaxial accelerometer. *Expert Systems with Applications* 40, 18 (December 2013), 7203–7211.

[40] Paul L Rosin. 1999. Measuring corner properties. *Computer Vision and Image Understanding (CVIU)* 73, 2 (Februrary 1999), 291–307.

[41] Edward Rosten and Tom Drummond. 2006. Machine learning for high-speed corner detection. In *Proc. of European Conference on Computer Vision (ECCV)*.

[42] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: an efficient alternative to SIFT or SURF. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. Barcelona, Spain, 2564 – 2571.

[43] Michael S Ryoo, Brandon Rothrock, and Larry Matthies. 2015. Pooled motion features for first-person videos. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*. Boston, USA, 896–904.

[44] Davide Scaramuzza and Friedrich Fraundorfer. 2011. Visual odometry. *IEEE Robotics & Automation Magazine* 18, 4 (December 2011), 80–92.

[45] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc. of International Conference on Learning Representations (ICLR)*. Banff, Canada.

[46] Jianbo Shi and Carlo Tomasi. 1994. Good features to track. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA, 593 – 600.

[47] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. Montreal, Canada, 568–576.

[48] Suriya Singh, Chetan Arora, and C. V. Jawahar. 2016. First Person Action Recognition Using Deep Learned Descriptors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA, 2620–2628.

[49] Philip HS Torr and Andrew Zisserman. 1999. Feature based methods for structure and motion estimation. In *Proc. of International Workshop on Vision Algorithms: Theory and Practice*. Corfu, Greece, 278–294.

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 4489–4497.

[51] S Vallurupalli, H Paydak, SK Agarwal, M Agrawal, and C Assad-Kottner. 2013. Wearable technology to improve education and patient outcomes in a cardiology fellowship program-a feasibility study. *Health and Technology* 3 (2013), 267–270.

[52] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)* 103, 1 (2013), 60–79.

[53] Kai Zhan, Steven Faux, and Fabio Ramos. 2014. Multi-scale Conditional Random Fields for first-person activity recognition. In *Proc. of IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Budapest, Hungary, 51–59.

[54] Kai Zhan, Steven Faux, and Fabio Ramos. 2015. Multi-scale Conditional Random Fields for first-person activity recognition on elders and disabled patients. *Pervasive and Mobile Computing* 16, Part B (January 2015), 251–267.

[55] Kai Zhan, Vitor Guizilini, and Fabio Ramos. 2014. Dense motion segmentation for first-person activity recognition. In *Proc. of IEEE International Conference on Control Automation Robotics & Vision (ICARCV)*. Marina Bay Sands, Singapore, 123–128.

[56] Kai Zhan, Fabio Ramos, and Steven Faux. 2012. Activity recognition from a wearable camera. In *Proc. of IEEE International Conference on Control Automation Robotics & Vision (ICARCV)*. Guangzhou, China, 365 – 370.

[57] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. 2016. Real-time Action Recognition with Enhanced Motion Vector CNNs. *arXiv preprint arXiv:1604.07669* (2016).

[58] Hong Zhang, Lu Li, Wenyan Jia, John D Fernstrom, Robert J Sclabassi, Zhi-Hong Mao, and Mingui Sun. 2011. Physical Activity Recognition Based on Motion in Images Acquired by a Wearable Camera. *Neurocomputing* 74, 12 (June 2011), 2184–2192.

[59] Hong Zhang, Lu Li, Wenyan Jia, John D Fernstrom, Robert J Sclabassi, and Mingui Sun. 2010. Recognizing physical activity from ego-motion of a camera. In *Proc. of IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*. Buenos Aires, Argentina, 5569–5572.

[60] Lijuan Marissa Zhou and Cathal Gurrin. 2012. A survey on life logging data capturing. *SenseCam 2012* (2012).