On the Combination of IMU and Optical Flow for Action Recognition

1st Taha Alhersh
Data and Web Science Group
University of Mannheim
Mannheim, Germany
taha@informatik.uni-mannheim.de

2nd Heiner Stuckenschmidt
Data and Web Science Group
University of Mannheim
Mannheim, Germany
heiner@informatik.uni-mannheim.de

***Abstract*—Different Action recognition methods use Inertial Measurement Unit (IMU) and optical flow independently. This research aims to explore the usefulness of combining IMU and Optical flow for action recognition. We are investigating the effectiveness of using statistical features to build an expandable feature vector space.**

***Index Terms*—Action Recognition, IMU, Optical Flow**

I. Introduction

The main objective of human behavior analysis is to understand the subjects behavior over time using motion information Figure 1. From egocentric perspective, this will instantiate a relationship over time between objects and hands to achieve a task; such as object recognition, hand detection, foreground segmentation and gaze estimation. Higher level of semantic is action level, which needs longer time to recognize simple events, such as open a jar or get water from the tap. Activity is a higher level of semantic representing a sequence of actions in time frame, could last from several minutes to hours. Examples of daily living activities: preparing a meal, making a coffee or brushing one's teeth. The difference between action and activity is not only about time lapse, but also about a higher semantic level due to more complex interactions between objects and people.

Visual and inertial sensing are two sensory modalities that can be used for action recognition either together or independently. RGB-D videos have been used in deep learning to recognize human actions [12], [5]. IMU only has been used for action recognition [2], [7]. Optical flow can be derived from visual sensing. It represents the apparent motion of objects in consecutive frame pairs. The displacement vector for each pixel of the first frame which called forward optical flow, or from the second frame back to the first frame and called backward optical flow. This forms a field of vectors in u and v directions. The interaction between Optical flow and action recognition has been discussed in [18]. The success of optical flow in many action recognition applications [21], [1], [14], [24] is not the temporal structure. However, it's the invariance to appearance of the representation [18]. Combining both optical flow and inertial data was used for ego-motion estimation [3] or for rotor-craft stabilization [16]. The advantages of combining optical flow and IMU data is the complementary characteristics of optical flow and inertial sensors. For instance, IMU data have large measurement uncertainty at slow motion and lower relative uncertainty at high velocities. Inertial sensors can measure very high velocities and accelerations. On the other hand, optical flow can track features very accurately invariant to appearance of the representation at low velocities. For high velocity, tracking is less accurate since the resolution must be reduced to obtain a larger tracking window with the same pixel size and, hence, a higher tracking velocity [16].

This paper presents an ongoing exploratory research to study the interaction and feasibility of combining optical flow and IMU data for action recognition. Our work is similar to Stein and McKenna work [19], however, we suggest using more IMU sensors rather than *Accelerometer* to increase the robustness. On the other hand, they have used Histograms of Relative Tracklets (RETLETS) and compare it to Histogram of Oriented Gradient (HOG) as baseline.

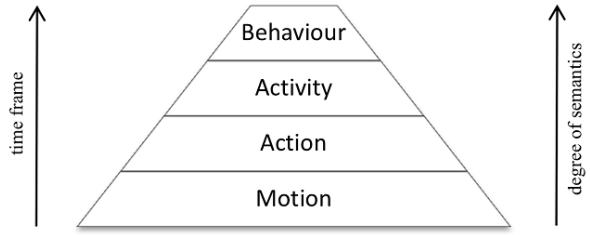

Fig. 1. Human behaviour analysis pyramid (reprinted from [15]).

II. PRELIMINARIES

Literally, optical flow refers to the displacement of intensity patterns [10]. Theoretically, it is the motion of visual features such as points, objects, shapes etc. through a continuous view of the environment. It represents the motion of the environment relative to an observer [1]. Optical flow can be considered as a variational optimization problem to find pixel correspondences between any two consecutive frames [11]. Research paradigms in this field have evolved from considering optical flow estimation as a classical problem [4] , to more high level approaches using machine learning, for

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

example, convolutional neural networks (CNN) as state-of-art method [9], [13], [23], [20]. Optical flow generated can be processed in many methods for different applications. This section is discussing the approach aligned with this research.

A. Histogram of Oriented Gradient (HOG)

Motion-based feature of optical flow can depends on oriented histograms of various kinds of local differences or differentials. For example, Histogram of Oriented Gradient (HOG), Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) [6], [22]. HOG method tiles a detector window with a dense grid of cells, with each cell containing a local histogram over orientation bins. At each pixel, the image gradient vector is calculated and converted to an angle, voting into the corresponding orientation bin with a vote weighted by the gradient magnitude. Votes are accumulated over the pixels of each cell. The cells are grouped into blocks and a robust normalization process is run on each block to provide strong illumination invariance. The normalized histograms of all of the blocks are concatenated to give the window-level visual descriptor vector for learning [6].

III. Data

The Carnegie Mellon University database (CMU) [8] contains measurements of the human activity involved in cooking and food preparation. Forty subjects have been recorded cooking five different recipes: brownies, pizza, sandwich, salad and scrambled eggs. In this research we have used the following modalities for 6 annotated subjects preparing brownies:

- Head mounted high spatial resolution (800 x 600) camera at low temporal resolution (30 Hertz).
- Two Wired IMUs (3DMGX) on right and left hands each with a triaxial accelerometer, gyro and magnetometer sensor sampling at 125 Hz.

We have extracted the corresponding IMU data and frames for 4 actions ("take-oil", "put-baking", "open-fridge", "stir-egg"). For each extracted action, the data for both modalities has been synchronized using time stamps provided. However, length of data for each action vary between different subjects.

IV. Method

Current objective is to construct a feature space for both IMU data and HOG of optical flows that can be used to measure the similarities between different feature vectors belonging to the same action. Data variation between subjects for the same action added challenges to data processing. Our proposed method consists of three main parts as follow:

1) Extract IMU features: we have used a dynamic sliding window based on the length of the IMU data to overcome the variation in IMU data lengths. From each window a 8 bins histogram is calculated with mean, variance and median of each window, to construct a feature vector with the same length for all subjects and actions.
2) Extract averaged HOGs from generated optical flows using [13] for each consecutive frames in the video. For each action an averaged HOGs will be generated with two dimensions for all subjects.
3) Conduct feature analysis for each feature vectors generated by IMU and averaged HOGs.

V. Preliminary Results

Some initial set of experiments were conducted using the proposed method explained above. The first results indicate the feasibility of using our approach and highlighted different issues will be included in the conclusions section.

Figure 5 shows the feature vector differences between various sensors for various actions. It is obvious that the produced features can distinguish between "take-oil", "put-baking" and ("open-fridge" or "stir-egg") while there is overlapping between ("put-baking" and "stir-egg"). So, we have conducted a T-Test to confirm this observation as in Table II. The results of T-Test confirms our observation that the feature vectors between "put-baking" and "stir-egg" are not significant and thus hard to be distinguished. More features are needed to classify actions.

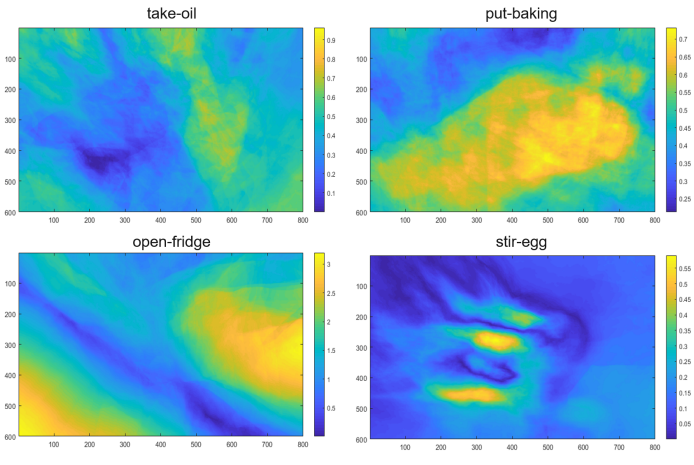

Fig. 2. Visualization of averaged HOGs for the four action used in this experiment.

Results of actions averaged HOGs are illustrated in Figure 2. The visualizations of averaged HOG for each action shows that it is easy to distinguish actions from each other. This information can be used as a complementary feature for IMU features to produce more robust feature vectors.

Distances between HOGs for different actions can be used as a quantitative measurement to evaluate the similarities between actions. Figure 4 shows the distances in log scale between all combinations of actions used in this research. Different metrics were used to calculate the distances between actions HOGs:

(i) **Chi Square:** is a metric that can be used to compare histograms and can be defined as:

$$d(x,y) = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{(x_i + y_i)} \quad (1)$$

(ii) **L1:** can be defined as:

$$\|s\|_1 = \sum_{i=1}^{n} |y_i - y_i)| \tag{2}$$

(iii) **Earth Mover's Distance (EMD):** is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, which we call the ground distance is given [17]. The EMD between histograms x and y is given by:

$$emd(x, y) = \sum_{i=1}^{n} |cd_x(i) - cd_y(i)| \tag{3}$$

where,

$$cd_x(i) = \sum_{j=1}^{i} x_j \tag{4}$$

and,

$$cd_y(i) = \sum_{j=1}^{i} y_j \tag{5}$$

(iv) **Euclidean:**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

(v) **Squared Euclidean (SQ Euclidean):**

$$d(x, y)^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{7}$$

Real value distances for averaged HOGs for pairwise combination of actions using previously mentioned metrics are shown in Table I. The actual differences provide more information about measurement differences inside the same metric, in which *Chi Square* metric provides the maximum difference among all pairwise actions as shown in Figure 3.

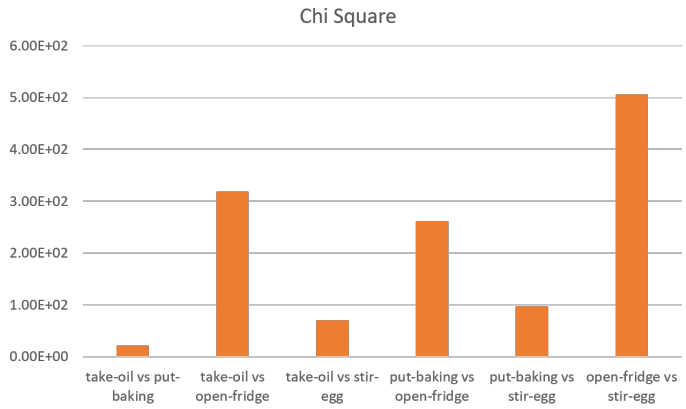

Fig. 3. Visualization of averaged HOG for the four action used in this experiment.

VI. Conclusions and Future Work

The preliminary investigations provide insights that combining optical flow and IMU data can be complementary to each other and indeed a promising direction. Nevertheless, substantial work needs to be done in order to formulate a robust combined approach that shows a convincing performance across different actions. In particular, the main open tasks are the following:

(a) **IMU statistics:** Even that the used statistical feature are considered to be naive , however, the idea of building an expandable feature space to accommodate different actions is promising. So, constructing significant statistical features for IMU data will be crucial for strengthen the pipeline.

(b) **Optical flow analysis:** The other open problem in this research is finding a better representation and analysis for optical flow that can be fused with IMU data. One idea could be using the correlated angles between both of them. Optical flow considered to be good feature for action recognition because it is invariant to appearance.

(c) **Sensor fusion:** Fusion between optical flow and IMU data still an opened problem in action recognition context. One approach can be statistical feature for both sensors. But, more investigations are needed to find alternatives for combining features.

(d) **Classification:** Using good classification approaches could enhance the recognition tasks for derived features from both modalities.

(e) **Pipeline development:** This open task refers to the overall assembly of the different parts to produce multi-modality action recognition system that can be used in various real-life scenarios. Then, the research can be upgraded to a higher level of semantics such as activity recognition or human behavior analysis.

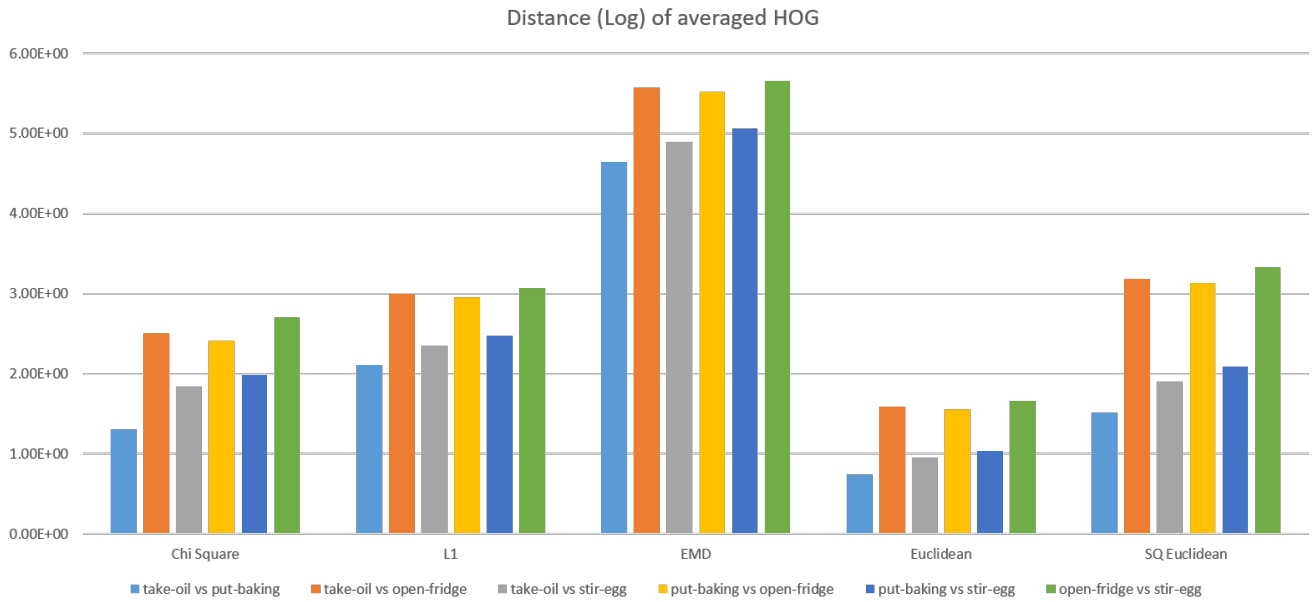

Fig. 4. Different distance metrics (Chi Square, L1, Earth Mover's Distance (EMD), Euclidean and Squared Euclidean (SQ Euclidean)) in Log scale between different combination of actions feature vectors for HOG.

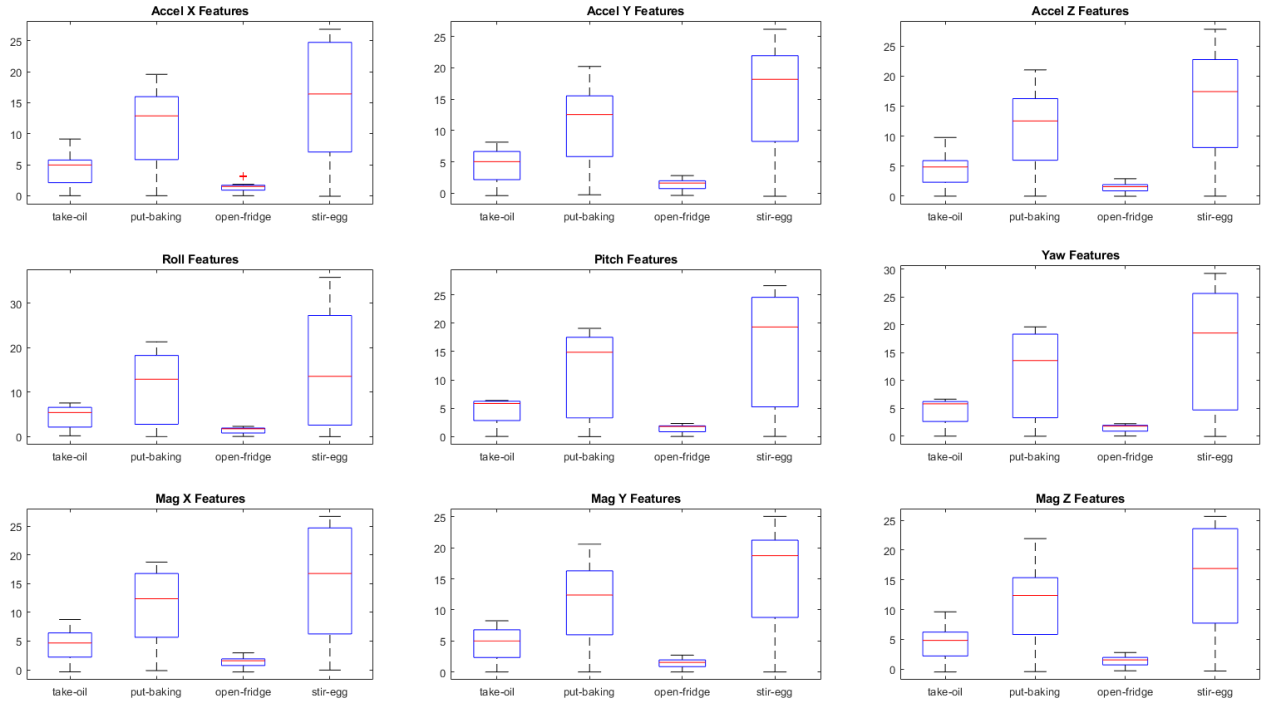

Fig. 5. IMU features for all IMU sensors used in various actions for the experiment.

TABLE I
PAIRWISE DISTANCE BETWEEN AVERAGED HOGS FOR DIFFERENT ACTIONS USING CHI SQUARE, L1, EMD, EUCLIDEAN AND SQUARED EUCLIDEAN METRICS

Action1	Action2	Chi Square	L1	EMD	Euclidean	Squared Euclidean
take-oil	put-baking	2.03E+01	1.29E+02	4.46E+04	5.52E+00	3.31E+01
take-oil	open-fridge	3.18E+02	9.80E+02	3.72E+05	3.88E+01	1.54E+03
take-oil	stir-egg	6.95E+01	2.24E+02	7.77E+04	8.88E+00	8.01E+01
put-baking	open-fridge	2.60E+02	8.96E+02	3.34E+05	3.64E+01	1.36E+03
put-baking	stir-egg	9.58E+01	2.96E+02	1.16E+05	1.09E+01	1.24E+02
open-fridge	stir-egg	5.05E+02	1.18E+03	4.50E+05	4.57E+01	2.12E+03

TABLE II
P-VALUE FOR T-TEST USING PAIRWISE ACTIONS FORM IMU FEATURE VECTORS FOR DIFFERENT SENSORS

P-Value										
Action1	**Action2**	**Accel-X**	**Accel-Y**	**Accel-Z**	**Roll**	**Pitch**	**Yaw**	**Mag-X**	**Mag-Y**	**Mag-Z**
take-oil	put-baking	0.0072	0.0077	0.0066	0.0181	0.0112	0.0112	0.0078	0.0069	0.0096
take-oil	open-fridge	0.0034	0.0036	0.0026	0.0013	0.001	0.0011	0.0044	0.002	0.0052
take-oil	stir-egg	0.002	0.0014	0.0013	0.0124	0.0024	0.0036	0.002	0.0009	0.0016
put-baking	open-fridge	0.1172*	0.1357*	0.099*	0.6825*	0.3141*	0.3219*	0.1321*	0.1207*	0.1947*
put-baking	stir-egg	0.2486	0.2425	0.2352	0.3618	0.2702	0.2907	0.2507	0.2234	0.2454
open-fridge	stir-egg	0.0001	0.0001	0.0001	0.0017	0.0002	0.0003	0.0001	0	0.0001
* $\times 1.0e^{-03}$										

REFERENCES

[1] Samet Akpinar and Ferda Nur Alpaslan. Video action recognition using an optical flow based representation. In *IPCV*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.

[2] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015.

[3] Michael Bloesch, Sammy Omari, Péter Fankhauser, Hannes Sommer, Christian Gehring, Jemin Hwangbo, Mark A Hoepflinger, Marco Hutter, and Roland Siegwart. Fusion of optical flow and inertial measurements for robust egomotion estimation. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 3102–3107. IEEE, 2014.

[4] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2011.

[5] Huseyin Coskun, David Joseph Tan, Sailesh Conjeti, Nassir Navab, and Federico Tombari. Human motion analysis with deep metric learning. *arXiv preprint arXiv:1807.11176*, 2018.

[6] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006.

[7] Juan Carlos Davila, Ana-Maria Cretu, and Marek Zaremba. Wearable sensor data classification for human activity recognition based on an iterative learning framework. *Sensors*, 17(6):1287, 2017.

[8] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *Robotics Institute*, page 135, 2008.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[10] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[11] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[12] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, 2017.

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[14] S Santhosh Kumar and Mala John. Human activity recognition using optical flow based feature set. In *Security Technology (ICCST), 2016 IEEE International Carnahan Conference on*, pages 1–5. IEEE, 2016.

[15] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al. Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1):72, 2016.

[16] Hugo Romero, Sergio Salazar, Rogelio Lozano, and Ryad Benosman. Fusion of optical flow and inertial sensors for four-rotor rotorcraft stabilization. *IFAC Proceedings Volumes*, 40(15):209–214, 2007.

[17] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998.

[18] Laura Sevilla-Lara, Yiyi Liao, Fatma Guney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. *arXiv preprint arXiv:1712.08416*, 2017.

[19] Sebastian Stein and Stephen J McKenna. Recognising complex activities with histograms of relative tracklets. *Computer Vision and Image Understanding*, 154:82–93, 2017.

[20] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[21] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *CVPR*, 2018.

[22] Jasper Uijlings, Ionut Cosmin Duta, Enver Sangineto, and Nicu Sebe. Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, 4(1):33–44, 2015.

[23] Anne S Wannenwetsch, Margret Keuper, and Stefan Roth. Probflow: Joint optical flow and uncertainty estimation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1182–1191. IEEE, 2017.

[24] Marco Wrzalik and Dirk Krechel. Human action recognition using optical flow and convolutional neural networks. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pages 801–805. IEEE, 2017.