Action Recognition in Law Enforcement: A Novel Dataset from Body Worn Cameras

Sameer Hans[1], Jean-Luc Dugelay[1], Mohd Rizal Mohd Isa[2] and Mohammad Adib Khairuddin[2]
[1] *EURECOM, 450 Route des Chappes, 06410 Biot, France*
[2] *Universiti Pertahanan Nasional Malaysia (UPNM), Kem Perdana Sg. Besi, 57000 Kuala Lumpur, Malaysia*
{*hans, dugelay*}*@eurecom.fr*, {*rizal, adib*}*@upnm.edu.my*

Keywords: Body Worn Camera, Multimodal Dataset, Action Recognition, Surveillance.

Abstract: Over the past decade, there has been a notable increase in the integration of body worn cameras (BWCs) in many professional settings, particularly in law enforcement. BWCs serve as valuable tools for enhancing transparency, accountability, and security by providing real-time, first-person perspective recordings of interactions and events. These devices capture vast amounts of video data, which can offer critical insights into the behaviors and actions of individuals in diverse scenarios. This paper aims to explore the intersection of BWCs and action recognition methodologies. We introduce FALEBaction: a multimodal dataset for action recognition using body worn cameras, with actions relevant to BWCs and law enforcement usage. We investigate the methodologies employed in extracting meaningful patterns from BWC footage, the effectiveness of deep learning models in recognizing similar actions, and the potential applications and implications of these advancements. By focusing on actions relevant to law enforcement scenarios, we ensure that our dataset meets the practical needs of the authorities and researchers aiming to enhance public safety through advanced video analysis technologies. The entire dataset can be obtained upon request from the authors to facilitate further research in this domain.

1 INTRODUCTION

BWCs are becoming common in a variety of industries. They have been put into practice around the globe, and they are an essential tool for law enforcement to improve accountability (Suat Cubukcu and Topalli, 2023), transparency (Choi et al., 2023), and evidence gathering (Todak et al., 2024).

One of the most promising applications of BWC data is in the field of action recognition, which focuses on the automated identification and classification of human actions within video footage. Action recognition in BWC footage can play a crucial role in promoting transparency and accountability, ultimately strengthening the trust between law enforcement and the community. Advances in deep learning in recent times have greatly improved the performance of action recognition systems. The ability to extract spatial and temporal information from video data has been significantly enhanced by models like Convolutional Neural Networks (CNNs) (Tran et al., 2015) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) (Majd and Safabakhsh, 2020) networks. These models can learn complex patterns and dependencies, accurately classifying a wide range of human actions.

Despite these advancements, action recognition by BWCs remains a challenging task due to various factors (Corso et al., 2018). The dynamic and often unpredictable nature of body worn footage, characterized by varied perspectives, motion blur, and occlusions, requires advanced algorithms capable of processing and interpreting complex visual input. Additionally, distinguishing between similar actions requires models to be highly sensitive to subtle differences in motion and context.

This work introduces FALEBaction, a multimodal[1] dataset specific to the actions that are useful for the applications of BWCs. This study is the first to provide a publicly[2] available dataset with actions specific to the usage of BWCs. It contains annotated

[1] In the context of machine learning and data analysis, a modality refers to a distinct type of data or information, such as text, image, audio, video, sensor reading. A multimodal dataset includes data from two or more of these modalities.

[2] To obtain the dataset, please visit https://faleb.eurecom.fr/

Hans, S., Dugelay, J.-L., Isa, M. R. M. and Khairuddin, M. A.
Action Recognition in Law Enforcement: A Novel Dataset from Body Worn Cameras.
DOI: 10.5220/0013151900003905
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 605-612
ISBN: 978-989-758-730-6; ISSN: 2184-4313
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

videos of 99 subjects with actions specific to law (divided into 2 different scenarios), along with the metadata such as GPS position, and heart rate of the user. We focus on the actions represented in 3.1, which are useful for a law officer in real time. These actions help in identifying if an officer is in a critical situation (when the subject attacks and runs away) or when an officer has made a significant step in their daily routine like making an arrest. We evaluate action recognition models of C3D (Tran et al., 2015), I3D (Carreira and Zisserman, 2018), SlowFast network (Feichtenhofer et al., 2019), and TimeSformer (Bertasius et al., 2021), with an approach to improve the accuracy in case of similar actions by following the method of sequential fine-tuning (Chakraborty et al., 2021) (a specialized form of transfer learning), along with a comparative study on the two different scenarios to evaluate the recognition performance in two different kinds of settings.

The paper is organized as follows. In section 2 we survey related work on action recognition and BWCs. In section 3, we introduce the steps followed in the data collection for the activity. We report our experimental setup and implementation in section 4. The experiments and the results are presented in section 5. Finally, the conclusions and future work follow in section 6.

2 RELATED WORKS

Existing action recognition datasets like HMDB51 (Kuehne et al., 2011), UCF101 (Soomro et al., 2012), Something-something (Goyal et al., 2017), Kinetics (Carreira and Zisserman, 2018), or EPIC-Kitchens (Damen et al., 2020) offer diverse actions but lack the specific context of BWC footage. While crowd scene datasets like NWPU-Crowd (Wang et al., 2021) might be relevant for scenarios with bystanders, they do not capture the specific interactions between officers[3] and suspects[4].

The advancement of deep learning revolutionized the field of action recognition by enabling end-to-end learning of spatiotemporal features directly from raw video data. CNNs were extended to process video data, leading to the development of 3D CNNs such as C3D. These models can capture both spatial and temporal information by applying 3D convolutions over the video frames (Duan et al., 2022). Table 1 shows performance of some of the models on existing datasets.

Very limited work exists on action recognition by BWCs. BWC footage presents unique challenges like first-person viewpoint, low resolution due to camera limitations, unbalanced data distribution across activities, privacy concerns over identifiable information, and limited annotated training data. Moreover, the existing work focuses on egocentric vision (Núñez-Marcos et al., 2022; Chen et al., 2019; Meng et al., 2018) and not on plausible activities for action recognition specific to BWCs like running, pushing, sitting inside the car, and making an arrest.

The study (Chen et al., 2019) is based on ego-activity recognition in first-person video in which they propose a system for classifying ego-activities in body worn video footage using handcrafted features and a graph-based semi-supervised learning method. They achieve comparable performance to supervised methods on public datasets, however the challenges include a lack of sufficient training data, and actions specific to the usage of body cameras.

(Chao et al., 2022) presents a multimodal dataset for human action detection that makes use of wearable sensors and a depth camera to identify actions more precisely. It has 880 sequences of 22 human acts carried out by 5 participants. Based on depth images of various actions, the recognition rate ranges from 58% to 97%. Although the number of participants in this dataset is extremely small, it is nonetheless helpful as a starting point for the task of action recognition by wearables.

In this paper, we introduce enough data for each of the actions that are specific to the usage of BWCs along with a sufficient number of subjects and a variety of environmental conditions; which would be helpful for law enforcement to assess the outcomes of body camera footage. To the best of our knowledge, no other study findings utilizing police BWCs in real life scenarios consisting of relevant actions have been published in the literature.

3 DATA COLLECTION

For the data collection, students from UPNM volunteered. The subjects were recorded using Cammpro[5] I826 Body camera. The recording took place over different sessions spread across a week. The camera was fixed on the middle of the chest of the user (Bryan, 2020). All the recordings were done with a video resolution of 2304 × 1296 pixels at 30 fps.

The activity was recorded in an outdoor setting. It was divided into 2 scenarios. In the first scenario,

[3] Officer is the user of the camera.

[4] The subject in question is the suspect.

[5] https://www.cammpro.com/

Table 1: Existing Dataset Evaluations as discussed in Related Works.

Dataset	Number of Classes	Model	Accuracy [%]
UCF101	101	C3D	85.2
UCF101	101	VideoSwin (Liu et al., 2022)	98
HMDB51	51	C3D	65.4
KINETICS-400	400	VideoSwin	82.7
Something-something	174	VideoSwin	69.6

we include the actions of walking, talking, showing hands, sitting, going forward and backward, standing, pushing, and running away. The second scenario has the same actions as the first one with some additional actions (an arrest is made instead of the subject running away). So, the additional actions comprise hands behind the head, turning around, sitting inside the car, and opening and closing the car doors. The subjects were provided structured scripts on how to act for the scenes. In total, we have 47 subjects (all male) for the first scenario, and 52 subjects for the second scenario (32 male and 20 female). Garmin[6] vivoactive 5 was used as an additional sensor to record GPS data and heart rate of the user. There will be fluctuations in the heart rate and sudden changes in GPS when the user chases the subject, which are useful parameters for the other officers to know when the user runs suddenly. These additional attributes are also useful in identifying an action more accurately.

After collecting all the videos, they are annotated using the CVAT[7] tool according to their actions for each video of a subject. We get XML files for each video, which consist of information like the frame number and the associated action label for that frame. Therefore, for every subject's video, we have 8 actions for the scene 1 and 11 actions for the scene 2.

3.1 Actions

The scenes depict the suspect's actions. However, we also mention about the officer in the scenes as the camera is inherently dynamic (due to the officer's movements).

- Backward: The suspect moves backward.
- Forward: The suspect moves forward and the officer is moving backward.
- Show Hands: The suspect is raising hands in surrender.
- Sit: The suspect sits on the ground.
- Stand: The suspect is standing and talking.
- Walk: The suspect is walking toward the officer and the officer also walks toward the suspect. For scene 2 only, there is an additional walk where the officer and the suspect walk together towards the car.
- Push: (Scene 1 Only) The suspect aggressively pushes the officer away and the officer moves backward.
- Run: (Scene 1 Only) The suspect is running away and the officer is chasing the suspect.
- Door open: (Scene 2 Only) We see the car door opening.
- Door close: (Scene 2 Only) We see the car door closing.
- Hands behind head: (Scene 2 Only) The suspect complies by placing both hands behind their head.
- Turn around: (Scene 2 Only) The officer makes the suspect turn around (to make him walk towards the car).
- Sit inside car: (Scene 2 Only) The officer makes the suspect sit inside the car.

4 SETUP

In this section, we discuss the steps followed for preprocessing of the videos, the networks used, and the implementation details of the networks.

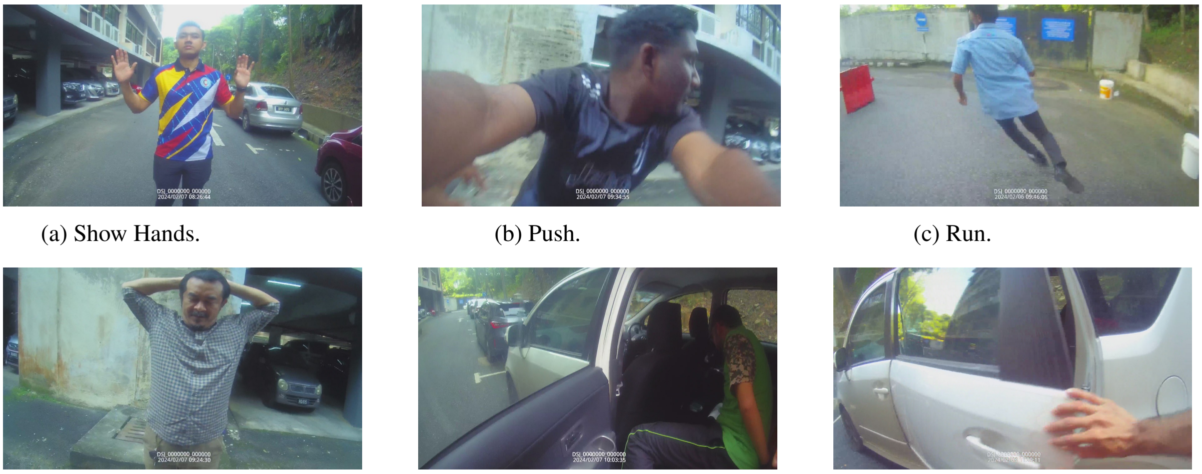
4.1 Preprocessing

Each annotated video is split according to its action labels. The dataset contains 13 unique action categories, where for the first scene, we have 8 action categories and for the second scene, we have 11 action categories. Fig. 1 shows some samples of the actions present in the dataset. In total, we obtain 954 individual video clips showcasing an action.

We divide the training, validation, and test set in the ratio of 65:15:20. As we split the videos according to the actions (and not the subjects), a subject can appear in both train and test set. Clips are resized to have a frame size of 128×171. On training, we randomly crop input clips into $16 \times 112 \times 112$ crops for

[6] https://www.garmin.com/

[7] https://www.cvat.ai/



(a) Show Hands. (b) Push. (c) Run.

(d) Hands behind head. (e) Sit inside car. (f) Door close.

Figure 1: Samples of action frames from the dataset. The first row represents specific actions of scene 1, and the second row represents some actions of scene 2.

spatial and temporal jittering. We also horizontally flip them with 50% probability.

Table 2 shows the total duration of each action and the proportion of that action in their respective scenarios.

4.2 Implementation Details

We used C3D, I3D, SlowFast network, and TimeSformer models for our experiments. These models were chosen for their respective strengths: C3D as a baseline model, I3D for its popularity and proven performance, SlowFast Network for its advancements in the field, and TimeSformer for its novelty.

- **C3D:** We implement a pretrained C3D model, trained on the Sports-1M dataset (Karpathy et al., 2014), which consists of 1.1 million sports videos belonging to one of the 487 sports categories. This pretrained model is fine-tuned on both the scene datasets to evaluate the performance of the model and gain some insights on the videos by BWCs. Training is done by SGD optimizer. The learning rate is fixed as 0.001 after various experiments. The initial layers of the model are frozen, and we add fc8 layer to match the number of classes in the new dataset. This layer starts with random weights and is trained from scratch. The optimization is stopped after 50 epochs.
- **I3D:** We experiment with I3D architecture pretrained on Kinetics-400 dataset. This model is used to initialize our network, where we replace the final projection layer to match the number of output classes corresponding to the actions in the dataset. The model was trained using the CrossEntropy loss function, and optimized using the Adam optimizer with a learning rate of 0.001. During training, both training and validation metrics, including loss, accuracy, precision, recall, and F1-score, were tracked to evaluate performance. We also ensured that each input video was processed as a stack of frames, allowing the I3D model to leverage its 3D convolutional layers to capture temporal dynamics, improving its ability to recognize actions across time.
- **SlowFast Network:** The SlowFast network operates by processing video inputs through two distinct pathways: the slow pathway, which samples frames at a lower frame rate to capture long-range temporal patterns, and the fast pathway, which processes higher frame-rate sequences to capture finer motion details. In our implementation, the SlowFast model is pretrained on Kinetics-400. The final fully connected layer of the network is replaced with a new layer corresponding to the number of actions in the dataset. To accommodate both fast and slow temporal dynamics, the video frames are split into two pathways, with the slow pathway subsampling every fourth frame, while the fast pathway uses all the frames. The training process uses cross-entropy loss to minimize classification error, with an Adam optimizer tuned with a learning rate of 0.001.
- **TimeSformer:** The TimeSformer is a deep learning architecture designed specifically for video understanding tasks like action recognition. It applies transformers directly to the video's spatial and temporal dimensions. It processes video frames as a sequence of patches, incorporating at-

Table 2: Class Proportion in FALEBaction.

Action Labels	Scene 1		Scene 2	
	Duration [seconds]	Proportion [%]	Duration [seconds]	Proportion [%]
Backward	101.40	3.98	122.40	2.84
Forward	81.50	3.20	98.67	2.29
Show Hands	165.53	6.50	410.53	9.51
Sit	361.80	14.20	428.77	9.93
Stand	663.67	26.04	656.60	15.21
Walk	696.70	27.34	1422.37	32.96
Push	66.07	2.59	-	-
Run	411.77	16.16	-	-
Door close	-	-	307.70	7.13
Door open	-	-	189.60	4.39
Hands behind head	-	-	131.17	3.04
Sit inside car	-	-	260.10	6.03
Turn around	-	-	288.10	6.68
Total	2548.43	100	4316.00	100

tention mechanisms across both time and space, allowing for more efficient and scalable learning of video features. In our implementation, the model is pretrained on Kinetics-400. We fine-tune the model by modifying the classifier head. During training, the model's performance is evaluated on validation data after each epoch, and its performance is seen on test data every 5 epochs to monitor accuracy and loss, aiming to improve the classification performance.

5 EXPERIMENTS

5.1 Scene 1 Analysis

In the first experiment (E1), we fine-tune the models directly on the actions of scene 1. We have 231, 64, and 80 videos for train, validation, and test sets respectively. We receive high test accuracies of 95.7%, 81.25%, 81% ,and 76.25% for TimeSformer, C3D, I3D, and SlowFast network respectively. However, we observe that most errors occur in similar actions, like moving backward and forward. As the camera is not fixed and moves along with the user, this creates confusion for the model. For example, when the suspect is moving forward, the officer moves slightly backward and when the suspect is moving backward, the officer remains still.

For the second experiment (E2), we group the similar actions and see the performance of the models. So, ”backward” and ”forward” are grouped as a single action (Motion). In total, we now have 7 actions for this scene. In this case, we have 203 videos in the train set, 56 videos in the validation set, and 70 videos in the test set. While fewer actions are considered, there is a substantial improvement in test accuracies as compared to experiment 1.

On evaluating the confusion matrix, the models are seen performing poorly on similar actions (backward and forward), and also because they have a very low proportion in the scene 1 actions as compared to others (3.98% and 3.20% for backward and forward respectively). To potentially improve the performance, for the third experiment (E3), we fine-tune the model in two sequential phases, first on just the confusing ”backward” and ”forward” actions, and then on the full set of 8 actions. This two-phase approach mimics a hierarchical learning process, where the model initially concentrates on differentiating subtle distinctions between closely related actions, and then expands its knowledge to the remaining classes in the second phase. This approach shows improvement in the performance of the model, particularly in the case of similar actions. We obtain accuracies of 88.75%, and 86.25% for C3D and I3D models respectively. There is a significant increase in the accuracy as compared to the first experiment when we follow this approach. For the TimeSformer model, this optimization was not very relevant and had similar results as the first experiment.

Table 3 shows the results of the 3 approaches discussed above.

5.2 Scene 2 Analysis

Similar to scene 1 experiments, we carry out the experiments for the scene 2 actions. For the first experiment (T1), there are 362 videos in the train set, 97

Table 3: Scene 1 Experiments.

Models	Test Accuracy [%]		
	E1 (8 actions)	E2 (7 actions)	E3 (8 actions)
C3D	81.25	92.19	88.75
I3D	81.00	90.00	86.25
SlowFast	76.25	78.75	76.00
TimeSformer	95.70	-	-

videos in the validation set, and 120 videos in the test set. There is a decrease in the accuracy as compared to previous scene tests. We receive accuracies of 88.5%, 80.83%, 66.67%, and 60% for TimeSformer, C3D, I3D, and SlowFast respectively. The confusion matrix for this experiment shows that the models produce the most errors in the actions of backward, and forward (confused as ”walk”), and show hands, and hands behind head. The first three actions (backward, forward, and walk) are very similar to each other and the last two (show hands and hands behind head) are identical and difficult to differentiate.

In the second experiment (T2), we merge the actions of ”backward” and ”forward” into a single action (Motion) and the actions of ”show hands” and ”hands behind head” into a single action (Hands). We now have 9 actions for this scene, where there are 303 videos in train set, 81 videos in validation set, and 99 videos in test set. When we fine-tune the models on the actions, again there is a significant increase in the test accuracy as compared to the first experiment.

In the final test (T3), we follow the approach of sequential fine-tuning again, where we first fine-tune the model on the confusing actions only (backward, forward, walk, show hands, and hands behind head), and then fine-tune this new model on the entire 11 actions for this scene dataset. We see a notable improvement in the performance, especially when comparing actions that are similar. We are able to improve the test accuracy to 88.33% as compared with the first experiment for the C3D model.

Table 4 shows the results of the 3 approaches discussed above.

Table 4: Scene 2 Experiments.

Models	Test Accuracy [%]		
	T1 (11 actions)	T2 (9 actions)	T3 (11 actions)
C3D	80.83	89.90	88.33
I3D	66.67	74.75	68.44
SlowFast	60.00	70.54	63.64
TimeSformer	88.50	-	-

5.3 Cross-Scene Analysis

It is essential to ensure that our model can generalize well across different environments. In both scenes, we have some common actions between them (Backward, Forward, Show Hands, Sit, Stand, Walk). For this experiment, we experimented with C3D and TimeSformer models as they had the best performance in previous experiments. The training and validation sets are made from one scene setting and the test set from different setting to evaluate the model’s ability to generalize to unseen settings for the same actions, which is crucial for real-world deployment. Table 5 shows the performance of the models across different scene environments on the test set. The first test is training on scene 1 environment and testing on scene 2 environment. For C3D model, there is a big drop in the test accuracy which becomes 75.76%. When the training set is scene 2 environment, we see a significant improvement in the accuracy value (85%). The model generalizes better from scene 2 to scene 1 (85%) as compared to scene 1 and scene 2. TimeSformer model performs slightly better than C3D (with similar behavior of generalizing better from scene 2 to scene 1). Although this experiment had less number of actions as compared to both scenes, the models had lower accuracy as compared to both scene analysis. Scene 2 has more variation and longer durations for certain actions like walking and showing hands. This increased variation leads the model to learn more robust features, which helps in better generalizing to new scenes.

Table 5: Cross-Scene Experiments.

Experiment	Test Accuracy [%]	
	C3D	TimeSformer
Train Scene 1 Test Scene 2	75.76	80.00
Train Scene 2 Test Scene 1	85.00	86.36

6 CONCLUSION

This work introduces FALEBaction, a multimodal annotated dataset for action recognition using BWCs, which is the first of its kind that is publicly available and based on actions relevant to the daily usage of law enforcement scenarios. The dataset is created by only using BWCs for images and videos, and an additional sensor to record metadata (GPS and heart rate of the user). This dataset addresses a critical gap in the current research landscape, providing comprehensive and detailed annotations for actions such as making an arrest, attacks on officers, and suspects fleeing, which are integral to an officers' daily duties. Although fine-tuning the model on the 2 scenes produces good results, we see improvement in the recognition performance following the approach of transfer learning. For cross-scene tests, the model generalizes well in the case of training from scene 2 and testing on scene 1 due to more variations and longer durations for the actions. TimeSformer outperformed the traditional models such as C3D, I3D, and SlowFast in recognizing complex law enforcement-related actions. Future efforts should focus on tailoring the models to better handle variations between environments to enhance cross-scene action recognition. Additionally, custom architectures or domain-adaptive layers could be introduced to better capture the contextual details of complex law enforcement scenarios, making the model more robust and capable of generalizing across different environments and action dynamics. Apart from action recognition with BWCs, the application of BWCs for face recognition has also been explored only to a limited extent, and there is tremendous potential for advancements in this area.

ACKNOWLEDGEMENTS

This research work is a collaborative study between UPNM-EURECOM and Naval Group.

REFERENCES

Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bryan, J. (2020). Effects of Movement on Biometric Facial Recognition in Body-Worn Cameras. *PhD thesis, Purdue University Graduate School*.

Carreira, J. and Zisserman, A. (2018). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chakraborty, S., Mondal, R., Singh, P., Sarkar, R., and Bhattacharjee, D. (2021). Transfer learning with fine tuning for human action recognition from still images. *Multimedia Tools and Applications*, 80.

Chao, X., Hou, Z., and Mo, Y. (2022). Czu-mhad: A multimodal dataset for human action recognition utilizing a depth camera and 10 wearable inertial sensors. *IEEE Sensors Journal*, 22:1–1.

Chen, H., Li, H., Song, A., Haberland, M., Akar, O., Dhillon, A., Zhou, T., Bertozzi, A. L., and Brantingham, P. J. (2019). Semi-supervised first-person activity recognition in body-worn video. *arXiv preprint arXiv:1904.09062*.

Choi, S., Michalski, N. D., and Snyder, J. A. (2023). The "civilizing" effect of body-worn cameras on police-civilian interactions: Examining the current evidence, potential moderators, and methodological limitations. *Criminal Justice Review*, 48(1):21–47.

Corso, J. J., Alahi, A., Grauman, K., Hager, G. D., Morency, L.-P., Sawhney, H., and Sheikh, Y. (2018). Video analysis for body-worn cameras in law enforcement. *arXiv preprint arXiv:1604.03130*.

Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2020). The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1.

Duan, H., Zhao, Y., Chen, K., Lin, D., and Dai, B. (2022). Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., and Memisevic, R. (2017). The "something something" video database for learning and evaluating visual common sense. pages 5843–5851.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video swin transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201.

Majd, M. and Safabakhsh, R. (2020). Correlational convolutional lstm for human action recognition. *Neurocomputing*, 396:224–229.

Meng, Z., Sánchez, J., Morel, J.-M., Bertozzi, A. L., and Brantingham, P. J. (2018). Ego-motion classification for body-worn videos. In Tai, X.-C., Bae, E., and Lysaker, M., editors, *Imaging, Vision and Learning Based on Optimization and PDEs*, pages 221–239, Cham. Springer International Publishing.

Núñez-Marcos, A., Azkune, G., and Arganda-Carreras, I. (2022). Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197.

Soomro, K., Zamir, A., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*.

Suat Cubukcu, Nusret Sahin, E. T. and Topalli, V. (2023). The effect of body-worn cameras on the adjudication of citizen complaints of police misconduct. *Justice Quarterly*, 40(7):999–1023.

Todak, N., Gaub, J. E., and White, M. D. (2024). Testing the evidentiary value of police body-worn cameras in misdemeanor court. *Crime & Delinquency*, 70(4):1249–1273.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Wang, Q., Gao, J., Lin, W., and Li, X. (2021). Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(06):2141–2149.