MOVING FORWARD WITH BWC: THE FALEB DATASET FOR MULTIMODAL IMAGE ANALYSIS

Sameer Hans, Jean-Luc Dugelay

EURECOM
450 Route des Chappes, 06410 Biot, France

Mohd Rizal Mohd Isa, Mohammad Adib Khairuddin

UPNM
Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia

ABSTRACT

Body worn cameras (BWCs) have grown in popularity over the last decade. They are becoming one of the most essential tools used by law enforcement for surveillance. Limited academic research has been conducted on image and video processing using BWCs. The number of datasets based on BWCs is incredibly few. For this objective, we introduce FALEB (Face, Action, License, Egocentric look using Body worn cameras): a multimodal dataset for image processing using BWCs. This work includes two distinct insights: (1) introduction of a dataset specific to body cameras with the applications of facial recognition, action recognition, license plate recognition, and egocentric look, and (2) baseline experiments on the dataset. We investigate the methodologies employed in extracting meaningful patterns from BWC footage, the effectiveness of deep learning models in recognizing faces and categorizing actions, and the potential applications of these advancements. By focusing on events uniquely 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 complete dataset is available for research purposes and can be accessed by contacting the authors[1].

***Index Terms*—** Body Worn Camera, Multimodal Dataset, Face recognition, Action recognition, Egocentric motion

1. INTRODUCTION

BWCs are being implemented in several industries. They are used in various parts across the globe, where they are an essential tool for improving law enforcement's accountability, transparency, and evidence gathering [1].

Face authentication systems have advanced significantly and are being employed in many different applications, including social networks, security systems, and surveillance. Deep learning-based feature extraction techniques yield very high performance for such systems. In recent years, the Convolutional Neutral Network (CNN) has become a very popular method for facial recognition. Their achievements have been fueled by the huge amount of data available and the enormous efforts made by the research community to produce vast labeled datasets like CASIAWebFace and VGGFace2.

Action recognition, which focuses on automatically identifying and categorizing human behaviors within video footage, is a promising use-case of BWC data. Recent developments in deep learning 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 3D CNNs.

BWCs have opened up new avenues for the study of egocentric motion. Unlike third-person observational data, egocentric motion records a person's actions straight from their perspective, including posture, gait, and dynamic movement patterns. Since these features are unique to each person, they provide a rich data source for user recognition applications.

This work introduces FALEB, a multimodal dataset for image processing using BWCs. This study is the first that provides a publicly available annotated dataset with events specific to law. It is divided into 4 sections: face recognition, action recognition, license plate recognition, and egocentric user recognition. The contribution of this study (introduction of the dataset along with initial experiments) is as follows:

- The first section contains 485 videos from 97 subjects for each environment for facial recognition: indoor, outdoor, and dark. The videos are classified according to the discussion context, considering expressions of happy, sad, angry, and neutral emotion per subject. We evaluate different model architectures to analyze their recognition performance, along with a comparative study on different environments.
- The second section of the dataset contains annotated videos of 99 subjects with actions specific to law (divided into two different scenarios), along with metadata such as GPS position and heart rate of the user. These law-specific 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 the performance of action recognition models, with the approach of sequential fine-tuning [2].

[1]The dataset can be obtained by visiting https://faleb.eurecom.fr/

- 18 videos are present from different parking lots for research in the area of license plate recognition (LPR) for identifying the minimum distance, view angle and illumination that the officer should focus on recognizing a license plate from a busy road or a parking space.

- In the final section of the dataset, we provide videos with an egocentric view of 23 subjects, where we explore the feasibility of identifying users based on their egocentric motion using BWCs. We evaluate deep learning models to analyze the motion patterns to achieve robust user recognition.

The paper is organized as follows. In section 2, we survey previous studies related to BWCs. In section 3, we introduce the steps followed in collecting data for the different activities. We report our experiments and results in sections 4, 5, 6, and 7. The conclusions and future work follow in section 8.

2. RELATED WORK

BWC footage presents unique challenges like first-person viewpoint, low resolution, unbalanced data distribution across activities, and limited annotated training data. The motion patterns are often influenced by factors [3] such as camera placement, environmental conditions, and walking speed.

Prior studies have explored face recognition [4, 5], action recognition [6, 7], and egocentric vision [8, 9] using various datasets and techniques. Several datasets [10, 11, 12] exist for action recognition, but lack the specific context of BWC usage. While crowd scene datasets like NWPU-Crowd might be relevant for scenarios with bystanders, they do not capture the specific interactions between officers and suspects. Although these works demonstrate the potential of wearables, they often lack training data and law-specific scenarios. Additionally, most existing datasets are limited to controlled environments, leaving a gap in real-world applicability. Table 1 shows the summary of related works. FALEB addresses these limitations by introducing a versatile dataset designed for diverse image processing tasks specific to law enforcement. To the best of our knowledge, no other study findings utilizing police BWCs in realistic scenarios consisting of relevant law-specific events have been published in the literature.

3. DATA COLLECTION

For the data collection, we gathered volunteers from UPNM university. They were recorded using Cammpro[2] I826 Body camera GPS. The recording took place over different sessions spread across a week. The camera was fixed on the middle of the chest of the user as recommended in [13]. The recordings have a video resolution of 2304×1296 pixels at 30 fps.

For facial recognition, we had 97 volunteers. We record 5 videos per subject, each showing them talking for 10-15 seconds, specifying the expressions of neutral, happy, angry, and sad. To ensure consistent and accurate expression of emotions, each participant was provided a script before the session, which included example sentences designed to elicit the target emotions. This activity was done in three different environments: indoor, outdoor, and dark. The indoor environment was well-lit with a consistent background and lighting conditions, and the dark environment was in the same place with the lights off. The outdoor environment had natural sunlight with varying intensities. The distance between the user and the subject was kept around 5-6 feet according to the reactionary gap [14]. The videos for indoor and outdoor environments lie in the visible[3] spectrum. The recording in the dark environment was done using infrared[4] feature of the camera. This data is useful for experiments on matching Near-Infrared (NIR) face images to Visible spectrum (VIS) face image, which is a challenging task.

The action recognition activity was shot in an outdoor setting (divided into 2 scenarios). The first scenario includes the actions of walking, talking, showing hands, sitting, going forward and backward, standing, pushing, and running away, and 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 include 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. We had 47 subjects (all male) for the first scenario, and 52 subjects for the second scenario (32 male and 20 female). Garmin[5] vivoactive 5 is used as an additional sensor to record GPS data and heart rate of the user. There will be sudden changes in GPS and fluctuations in heart rate when the user chases the subject, which are useful parameters for the other officers to know when the user sprints. Finally, for every subject's video, we have 8 actions for scene 1 and 11 actions for scene 2 (13 distinct actions in total).

For license plate recognition, the recording was done in 18 different parking lots. The user takes a normal walk in the parking lot and records the plates from different angles.

The egocentric activity was recorded in outdoor setting and divided into 2 scenarios. Two endpoints (A and B) were designated at opposite ends of the campus, approximately 8-minute walk apart. In the first scenario, the user walks from A to B at normal pace. For the second scenario, they follow the same path back (B to A) at slow jogging pace. Before starting, all subjects received clear instructions on how to perform the tasks. 23 subjects participated in this activity, and each sub-

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

[3]The visible spectrum (VIS) is the region perceivable by the human eye, which includes wavelengths from 400 nm to 700 nm.

[4]The camera produces near-infrared (NIR) images. The NIR region spans wavelengths ranging from 780 nm to 2500 nm.

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

Study	Tasks	Key Techniques	Limitations
[9]	Egocentric Hand Identification	ResNet18+3D convolutions	Limited generalizability
[5]	Face Recognition	CNN with 5 different loss functions	Limited data environment-wise
[6]	Action Recognition	Graph-based semi-supervised learning	Lacking law-specific actions
[8]	Egocentric user identification	LPC+Kernel SVM	Focused only on controlled data
[4]	Face Recognition	Eigenfaces, Fisherfaces and Wavelet Transforms	Limited data size
[7]	Action Recognition	Transformer+sequential fine-tuning	Limited number of actions

Table 1: Summary of Related Works based on Wearables

ject contributed two egocentric videos: one around 8 minutes long, capturing their walk from A to B, and another around 5 minutes long, documenting their slow jog from B to A.

Figure 1 shows the structure of the full dataset.

4. FACE RECOGNITION WITH BWC

4.1. Preprocessing

After getting the frames from the videos, faces are detected and cropped using the Dlib library. The frames are resized into 224×224 to ensure uniformity with various model architectures. The frames for the experiments are selected according to sharpness metrics using Laplace variance (if their sharpness is higher than a threshold fixed as 0.002 for the frames). Figure 2 shows some samples obtained after preprocessing. The frames are selected such that we have samples from all the expressions for diversity. We divide the training, validation, and test set in the ratio of 70:15:15.

4.2. Experiments

We used the VGG16 [15] architecture, Inception Resent V1 [16], and Bidirectional Encoder representation from Image Transformers (BEiT) [17]. These models were chosen for their respective strengths: VGG16 as a baseline model, Inception ResNet V1 for its popularity and proven performance, and BEiT for its novelty and recent advancements in the field. During training, the models are trained over 50 epochs, with learning rate fixed as 0.001, and by using Adam optimizer and Cross-entropy loss.

4.2.1. Self-Environment analysis

For the analysis, we consider 20 subjects randomly chosen with a size of 140 frames per subject. The selected frames are normal (as captured by the camera) to see the performance of different models on the videos. BEiT model performs the best among all the models giving accuracy of 98.1% and 99.3% for indoor and outdoor environments respectively. We also receive high accuracies in the range of [96%, 99%] for the indoor and outdoor environments when considering the models of VGG16 and Inception ResNet. We record accuracy value of 97.95% for dark environment with BEiT model.

4.2.2. Cross-Environment analysis

For this experiment, the training is done in one particular environment and tested on different validation and test environment. Rank-1 and Rank-5 accuracy values are recorded for VGG-Face model. When the training set is indoor and we test on outdoor set, we get a remarkable Rank-1 accuracy in the outdoor test set (87.5%). When the training set is outdoor, there is around 12-point drop in the Rank-1 accuracy (75.75%) as compared with the previous test. At Rank-5, we achieve high accuracy value of 95.83%.

4.2.3. Cross-Spectrum analysis

In this experiment, we selected 20 subjects, and the training set has images from both VIS and NIR spectra for fine-tuning the model. This approach aims to train the model on diverse conditions, potentially improving its robustness to variations. We experiment with the models described earlier. The training set consists of 98 images per subject (49 VIS and 49 NIR spectrum images). For testing, we create validation and test sets from both spectra (VIS and NIR) and see the performance of the models on both these spectra separately. The sizes of the validation and test sets are based on the ratio 70:15:15. We get comparable results from existing experiments [18] done on datasets with traditional cameras. On the VIS test set, we obtain accuracy of 96.56% and on the NIR test set, we obtain accuracy of 93.36% with Inception ResNet model.

5. ACTION RECOGNITION WITH BWC

5.1. 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. 3 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. 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 spatial and temporal jittering. We also horizontally flip them with 50% probability.

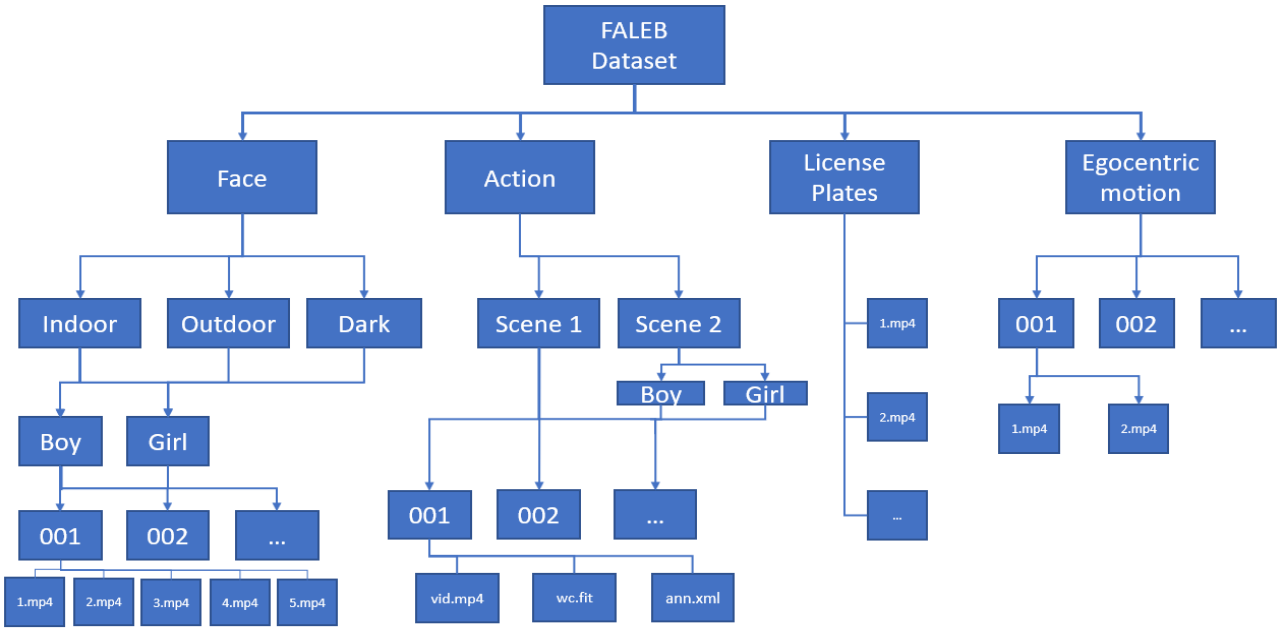

Fig. 1: Structure of the dataset. (More details on website)

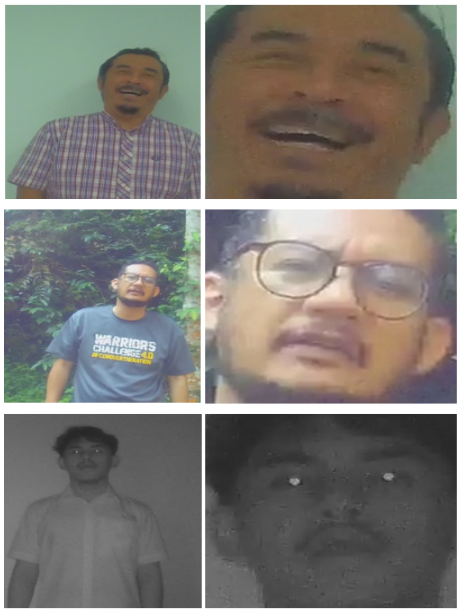

Fig. 2: The rows represent normal and preprocessed in indoor, outdoor, and dark environments respectively.

5.2. Experiments

We used C3D [19], I3D [11], SlowFast network [20], and TimeSformer [21] models for our experiments. These models were chosen as follows: 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.

5.2.1. Scene 1 Analysis

We fine-tune the model in two sequential phases, first on just the similar "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. We obtain accuracies of 95.7%, 88.75%, and 86.25% for TimeSformer, C3D and I3D models respectively.

5.2.2. Scene 2 Analysis

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 model on the entire 11 actions for this scene. We see comparable performance to previous scene, especially when comparing actions that are similar. We receive accuracies of 88.5%, 88.33%, 68.44%, and 63.64 for TimeSformer, C3D, I3D, and SlowFast respectively.

Fig. 3: The first row represents actions of show hands, push, and run from scene 1; the second row represents hands behind head, sit inside car, and door close from scene 2.

6. LPR WITH BWC

We employed a pre-trained YOLOv5 model (on a car_plates dataset), to detect and recognize license plates within the BWC footage. The model demonstrated promising results under favorable conditions, such as clear visibility, close proximity, and stable camera positioning. However, performance was significantly affected (1 in 10 plates) by challenging conditions commonly encountered in real-world scenarios. Detection accuracy decreased with increasing distance between the camera and the vehicle. At close range (within 5 meters), the model achieved reliable detection. However, at distances beyond 10 meters, detection rates dropped significantly due to reduced plate resolution. Oblique viewing angles posed a significant challenge for the model. Rapid camera movement, often caused by the officer's movement, introduced motion blur and further reduced accuracy. While the model performed reasonably well in well-lit environments, both underexposure (low light) and overexposure (glare) negatively impacted performance. In low-light conditions, plates became difficult to distinguish from the background, while glare caused saturation and loss of detail.

7. USER IDENTIFICATION BY EGOCENTRIC VIEW

7.1. Preprocessing

The videos are divided into sequences of 4 seconds, which are adequate to capture a few steps of the user's motion. These videos are converted into frames and organized based on the two scenarios (walking and slow jogging). At 30 fps, each 4-second sequence results in 120 frames. The frames are then resized to a standard resolution of 224×224 pixels to ensure uniformity with various model architectures. During training, input clips are randomly cropped into $16 \times 112 \times 112$ patches, enabling both spatial and temporal jittering to improve generalization. These augmentation strategies help to enhance the model's ability to identify users under varied conditions.

To further prepare the data for the task, normalization is applied to the pixel values of the frames using the mean and standard deviation of the ImageNet dataset, for standardizing pixel intensity values across all channels. The training, validation, and test sets are split in ratio of 65:15:20.

7.2. Egocentric user recognition

The models were trained for 20 epochs, and metrics of accuracy, precision, recall, F1-score, and loss were tracked for training, validation, and testing phases.

Among the models, I3D demonstrated the most consistent performance, achieving a test accuracy of 89.9% and a balanced F1-score of 0.90, showcasing its robust ability to capture temporal dynamics in egocentric motion. The TimeSformer model, also achieved competitive results, with a test accuracy of 89.23% and F1-score of 0.89, demonstrating its capability to model long-range temporal dependencies. While SlowFast exhibited slightly lower performance with a test accuracy of 88%, it maintained a precision and recall of 88%. C3D demonstrated the lowest test accuracy at 85.36%, highlighting its limitations in capturing the complexity of egocentric motion. Validation accuracy remained consistent across the models, further emphasizing the generalization ability of I3D, SlowFast, and TimeSformer.

8. CONCLUSION

This work introduces a multimodal dataset for image processing using BWCs. 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), which makes the applications of the dataset specific to law enforcement scenarios. This dataset addresses a critical gap in the current research landscape, providing data for law-specific activities with BWCs, along with detailed annotations for face expressions and actions such as making an arrest, attacks on officers, and suspect running away, which are integral to the officers' daily duties. For the preliminary experiments, a comparative analysis between different models is done on different sections of the dataset. The models demonstrated satisfactory performance. For future work, we aim to extend the dataset by incorporating images captured with standard cameras and comparing them with those obtained from BWCs, along with addtional population and sensors. New transformer-based models like Video Swin need to be tested for action recognition. We will also explore advanced fusion techniques like multi-stream architectures for recognition by egocentric view.

9. REFERENCES

[1] Natalie Todak, Janne E. Gaub, and Michael D. White, "Testing the evidentiary value of police body-worn cameras in misdemeanor court," *Crime & Delinquency*, vol. 70, no. 4, pp. 1249–1273, 2024.

[2] Saikat Chakraborty, Riktim Mondal, Pawan Singh, Ram Sarkar, and Debotosh Bhattacharjee, "Transfer learning with fine tuning for human action recognition from still images," *Multimedia Tools and Applications*, vol. 80, 05 2021.

[3] Jason J. Corso, Alexandre Alahi, Kristen Grauman, Gregory D. Hager, Louis-Philippe Morency, Harpreet Sawhney, and Yaser Sheikh, "Video analysis for body-worn cameras in law enforcement," *arXiv preprint arXiv:1604.03130*, 2018.

[4] Wasseem Al-Obaydy and Harin Sellahewa, "On using high-definition body worn cameras for face recognition from a distance," in *Biometrics and ID Management*, Berlin, Heidelberg, 2011, pp. 193–204, Springer Berlin Heidelberg.

[5] Ali Almadan, Anoop Krishnan, and Ajita Rattani, "Bwcface: Open-set face recognition using body-worn camera," *arXiv preprint arXiv:2009.11458*, 2020.

[6] Honglin Chen, Hao Li, Alexander Song, Matt Haberland, Osman Akar, Adam Dhillon, Tiankuang Zhou, Andrea L. Bertozzi, and P. Jeffrey Brantingham, "Semi-supervised first-person activity recognition in body-worn video," *arXiv preprint arXiv:1904.09062*, 2019.

[7] Sameer Hans, Jean-Luc Dugelay, Mohd Rizal Mohd Isa, and Mohammad Adib Khairuddin, "Action recognition in law enforcement: A novel dataset from body worn cameras," in *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods - ICPRAM*. INSTICC, 2025, pp. 605–612, SciTePress.

[8] Yedid Hoshen and Shmuel Peleg, "An egocentric look at video photographer identity," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4284–4292, 2016.

[9] Satoshi Tsutsui, Yanwei Fu, and David Crandall, " Whose hand is this? Person Identification from Egocentric Hand Gestures ," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Los Alamitos, CA, USA, Jan. 2021, pp. 3398–3407, IEEE Computer Society.

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

[11] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. 43, no. 06, pp. 2141–2149, jun 2021.

[13] Julia Bryan, "Effects of Movement on Biometric Facial Recognition in Body-Worn Cameras," *PhD thesis, Purdue University Graduate School*, 5 2020.

[14] "Safe distance," Last accessed: 7 July 2024, https://www.officer.com/home/article/10248804/safely-handling-suspicious-person-stops.

[15] Masaki Nakada, Han Wang, and Demetri Terzopoulos, "Acfr: Active face recognition using convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 07 2017, pp. 35–40.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[17] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei, "BEit: BERT pre-training of image transformers," in *International Conference on Learning Representations*, 2022.

[18] Naheed Jahan Siddiqui, "Novel approach for face recognition using cross-spectral environment," *International Journal of Research in Engineering and Science (IJRES)*, vol. 09 Issue 10, pp. 70–76, 2021.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[21] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.