Seeing through Wearables: A Comprehensive Face Recognition Dataset from Body Worn Cameras

Sameer Hans[1], Jean-Luc Dugelay[1] and Mohd Rizal Mohd Isa[2]

[1] EURECOM, Sophia-Antipolis, France
[2] Universiti Pertahanan Nasional Malaysia (UPNM), Kuala Lumpur, Malaysia

Abstract. Body worn cameras (BWCs) have become more and more popular over the last decade. They are becoming one of the essential tools for law enforcement officers to carry with them for surveillance purposes. Generally, videos captured by BWCs are used a posteriori through visual inspection in case of major problems between police officers and citizens. Limited academic research has been conducted on image and video processing using BWCs. There are extremely few datasets available that are based on BWCs. For this objective, we introduce FALEBface: a novel dataset for face detection and recognition using BWCs. We also provide some baseline experiments on the proposed dataset. This work includes two distinct insights: (1) introduction of a dataset specific to body cameras with the applications of facial recognition, and (2) evaluation of models in different environments. The experiments are carried out in three environments: indoor, outdoor, and dark which includes a variation of expressions for the subjects, and a comparative study is also done to check the performance of the models across environments and spectra. To facilitate further research in this domain, the entire dataset can be obtained upon request from the authors[3].

Keywords: Body Worn Camera · Face Recognition · Multimodal Dataset · Near-infrared spectrum image · Visible spectrum image.

1 Introduction

Body worn cameras (BWCs) have become increasingly prevalent in various sectors. They have been implemented in different parts of the world, where they serve as a critical tool in the areas of law enforcement for enhancing transparency [8], accountability [20], and evidence collection [21].

Face biometric authentication systems have made significant progress and are now used in a variety of applications such as identifying criminals, surveillance, security systems, and even social networks. Very high performance is achieved for such systems using deep learning-based approaches for feature extraction. In recent years, the Convolutional Neutral Network (CNN) [17] has become a very popular and successful method for facial recognition. The CNN automatically extracts a variety of features of the image and has good robustness to complex

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

environments. Their achievements have been fueled by the huge amount of data accessible online and the enormous efforts made by the research community to produce vast labeled datasets like CASIAWebFace [24] and VGGFace2 [7].

Law enforcement professionals have recently shown interest in using face recognition with BWCs to protect officers, enable situational awareness, and provide evidence for trial. This function could be useful for the suspect too (if the officer has inappropriate behaviour). In this particular domain, there are currently a limited number of studies. Most previous studies focus on body camera placement or relied on traditional machine learning methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [3], which were widely used before the rise of deep learning techniques.

This work introduces FALEBface, a novel dataset for face detection and recognition using body worn cameras. The dataset 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 deep learning models based on VGG-Face [16], Inception ResNet [18], and Vision Transformer (ViT) [9], along with a comparative study on different environments and spectra[4] to identify the individuals across diverse conditions and variations.

The paper is organized as follows. In section 2 we survey previous work related to BWCs. In section 3, we introduce the steps followed in the data collection for the activity. We report our implementation, experiments and results in section 4. Finally, the conclusions and future work follow in section 5.

2 PREVIOUS WORK

While there is a considerable volume of work in the area of egocentric vision, the use of BWCs for image processing tasks in real-world surveillance settings remains relatively limited and underexplored. 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. Even in this limited work, most of the work in the literature is focused on egocentric vision and actions [12, 14] and not on scenarios specific to law like recognition of a suspect at night or matching the face from a different environment.

There exists some work done on facial recognition using body cameras [4, 3]. The authors [3] worked with 20 subjects, and evaluated with traditional methods for facial image representation. They obtained accuracy in the range [68%, 75%]. This study had very limited data size to perform further evaluations.

In the most prominent study for facial recognition using body cameras [4], the authors tried to implement a dataset that is based on cropped images in indoor and outdoor environments. They also introduced some subjects in the dark environment, but it is very limited (only 2 subjects) to perform any evaluations.

[4] The camera generates images and videos in visible and near-infrared spectra.

Although they introduce facial images of 132 subjects, the dataset is limited by a lack of expression diversity (many of the subjects are present with a neutral expression) and lack of proper environmental conditions. They also include some tests based on ResNet architecture with different loss functions. For future work, they proposed to evaluate some lightweight CNN models.

EgoFace [10] presents a novel lightweight framework for face performance capture and videorealistic reenactment using a single egocentric RGB camera. The method estimates facial expressions from a single oblique view using a deep encoder, and synthesizes a frontal videorealistic face with an adversarially trained network. The system is trained in a supervised manner and handles diverse lighting, movement, and facial expressions. While EgoFace focuses on videorealistic facial reenactment from egocentric views, our dataset targets identity recognition under surveillance-like conditions with head motion, varying environments, and expression diversity, providing a benchmark more aligned with real-world face recognition tasks in mobile contexts.

An in-depth study [12] focuses on action recognition using BWCs. By focusing on actions relevant to law enforcement scenarios, they address the challenges of egocentric motion, dynamic camera perspectives, and the difficulty of recognizing common actions of individuals recorded by law enforcement officers. State-of-the-art models, along with novel approaches are benchmarked, with results highlighting the need for robust fine-tuning strategies and domain adaptation across different scenarios. This work offers a significant step forward for real-world action recognition from wearable devices.

The authors of [11] try to identify users through egocentric motion captured by BWCs. The study benchmarks several deep learning models, along with a two-stream I3D architecture, combining RGB and optical flow inputs, which achieved the best performance, underscoring the complementary strengths of appearance and motion cues in identifying users. The paper also highlights the challenges of domain shift. This work is one of the first to explore user identification from egocentric motion using real-world BWC data, establishing a valuable benchmark for biometric recognition under mobile, first-person conditions.

3 DATA COLLECTION

For the data collection, students from a university volunteered. The subjects were recorded using Cammpro[5] I826 Body camera, as shown in Figure 1. 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 [6]. All the recordings were done with a video resolution of 2304×1296 pixels at 30 fps.

For facial recognition, we had 97 volunteers. We recorded 5 videos per subject, each showing them talking for 10-15 seconds. These videos specified the expressions of neutral, happy, angry, and sad. Each participant was provided a script before the recording session for consistency. The script included example

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

Fig. 1. Cammpro I826 Body Camera.

sentences specifically designed to elicit the target emotions. Participants were instructed to act out these sentences, focusing on both facial expressions and vocal tones to convey the intended emotions. As the camera is positioned on the chest, we record the subject from the head to the torso. This was done in three different environments: indoor, outdoor, and dark. The indoor environment was well-lit with a uniform background and lighting conditions, and the dark environment was in the same place with the lights switched off. The outdoor environment had natural sunlight conditions with varying intensities of light. The distance between the user and subject was kept around 5-6 feet according to the reactionary[6] gap [2]. The recorded videos for indoor and outdoor environments lie in the visible[7] spectrum. The recording in the dark environment was done using the infrared[8] feature of the camera. This data is useful for experiments of matching Near-Infrared (NIR) face images to Visible spectrum (VIS) face images, which is a very challenging task. Figure 2 shows the sample images from the VIS and NIR spectrum. These recordings captured the facial expressions/emotions, speech, hand gestures, and head movements of each subject.

4 Preliminary Assessment of the Dataset

4.1 Preprocessing

The videos are converted into frames and organized according to the expressions. After getting the frames, faces are detected and cropped using the Dlib library

[6] The Reactionary Gap is a concept based on the rule that distance equals time. The gap or distance you stay away from a suspect provides time for you to respond.

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

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

Fig. 2. Examples of VIS images (top row) and NIR images (bottom row).

[13]. The frames are resized into 224×224 to ensure uniformity and compatibility with various model architectures. To improve the contrast and brightness of the frames (and achieve consistent illumination across the frames) [15], histogram equalization is applied to the resized frames. Finally, these frames are selected according to sharpness metrics using Laplace variance. Figure 3 shows some samples obtained after preprocessing. For cross-environment analysis, we perform experiments on two sets of frames: histogram equalized and normal (RGB) frames. The frames for the experiments are selected if their sharpness is higher than a threshold (fixed as 0.01 for equalized and 0.002 for RGB frames). 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 Implementation

We used the VGG16 architecture, Inception Resent V1, and BEiT. These models were chosen for their respective strengths: VGG16 as a baseline model, Inception ResNet V1 for its popularity and proven performance, and Bidirectional Encoder representation from Image Transformers (BEiT) for its novelty and recent advancements in the field.

- **VGG16:** We implement a VGG-Face model [16] architecture, a 16-layer CNN that is trained on over 2 million celebrity images. The face weights are loaded into the implemented architecture and using this model, we extract image features from the output of the fc-6 layer and use them in our subsequent classification stage. Each image is represented by a 4096-dimensional feature vector. For the activation function, we use **softmax**, which is useful in dealing with multi-class classification problems. During the training, we use the Adam optimizer and Sparse Categorical Cross-entropy loss. The learning rate is fixed as 0.001.
- **Inception Resent V1:** We experiment with Inception ResNet (V1) [18] architecture pretrained on CASIAWebFace and VGGFace2. The process typ-

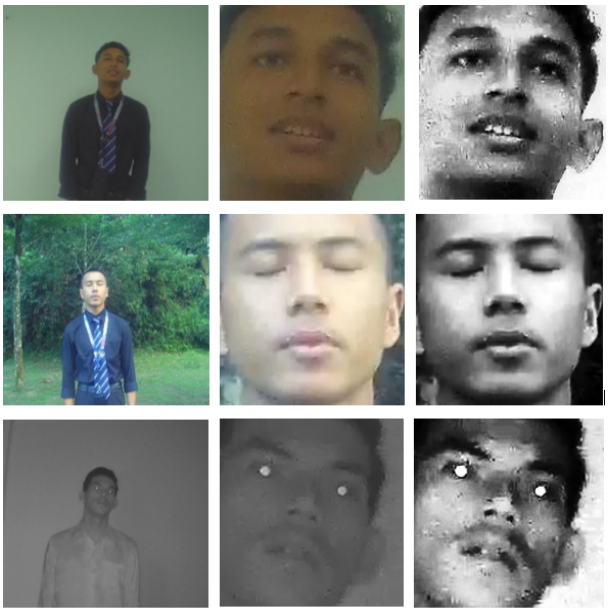

Fig. 3. Samples of acceptable frames in each environment. The first row represents normal, cropped, and equalized frame in the indoor environment. The second row represents normal, cropped, and equalized frame in the outdoor environment. The last row represents normal, cropped, and equalized frame in the dark environment.

ically starts by loading the pretrained weights and setting up a training loop that includes a data loader, Adam optimizer, and learning rate scheduler. The model is trained over 50 epochs, with each epoch consisting of forward passes through the data, loss computation using Cross-entropy loss, and backpropagation to update the model weights.

– **BEiT:** The BEiT [5] model is a Vision Transformer, pretrained in a self-supervised fashion on ImageNet-21k [1], a dataset comprising of 14 million images and 21,841 classes. Images are presented to the model as a sequence of fixed-size patches (resolution 16x16), which are linearly embedded. The model uses relative position embeddings and performs classification of images by mean-pooling the final hidden states of the patches. The fine-tuning takes place by placing a linear layer on top of the pretrained encoder. The learning rate is fixed as 0.001.

4.3 Experiments and Results

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 (not equalized) to see the performance of different models on the videos as captured by the camera. Table 1 shows the performance of different models when fine-tuned on the subjects (environment-wise). 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.

Table 1. Environment-wise analysis.

Model	**Pretrained Dataset**	**Environment**	**Validation Accuracy [%]**	**Rank-1 Accuracy [%]**
VGG16	VGGFace2	Indoor	99.25	96.51
		Outdoor	99.25	98.5
		Dark	97.5	95.25
Inception ResNet	VGGFace2	Indoor	99.76	97.84
		Outdoor	99.7	99.26
		Dark	98.7	97.75
Inception ResNet	CASIA-Webface	Indoor	99.5	97.64
		Outdoor	99.7	99.26
		Dark	99	97.7
BEiT	Imagenet-21K	Indoor	99.7	**98.1**
		Outdoor	99.78	**99.3**
		Dark	98.5	**97.95**

Cross-Environment analysis It is essential to ensure that our model can generalize well across different environments. For this experiment, the training is done in one particular environment and tested on different validation and test environment to evaluate the model's ability to generalize to unseen environments, which is crucial for real-world deployment. We select 20 subject and perform experiments on 3 sizes, where S1 contains 20 frames per subject, S2 contains 40 frames per subject, and S3 contains 140 frames per subject respectively. For each size, the training, validation, and test set are in the ratio 70:15:15. For this experiment, we consider separate sets from both the frames: equalized and not equalized. Table 2 shows the performance of the VGG-Face model (best model in this experiment) when fine-tuned on the subjects. Rank-1 and Rank-5 accuracy values are recorded. The first test is training on indoor and testing on outdoor environment. Although there is a drop in the accuracy from environment-wise testing, we get a remarkable Rank-1 accuracy in the outdoor test set (87.5%) for the non-equalized frames. When the training set is outdoor, there is around 10-point drop again in the Rank-1 accuracy as compared with the previous test. At Rank-5, we achieve high accuracy value of 95.83% when training environment is indoor. The accuracy values are higher if the training sets are more focused on the indoor environment as it has a uniform background and lighting conditions, controlled environment, and reduced variability.

Table 2. Cross-Environment comparison. We experiment with VGG-Face model, where training is done on one environment, and testing for the other environment.

Training Environment	**Test Environment**	**Equalized**	**Size**	**Rank-1 [%]**	**Rank-5 [%]**
Indoor	Outdoor	no	S1	78.33	90
			S2	**87.5**	**95.83**
			S3	77	93.25
		yes	S1	73.33	91.67
			S2	75	**95**
			S3	76	92.75
Outdoor	Indoor	no	S1	73.33	88.33
			S2	72.5	**95.13**
			S3	75.75	91.77
		yes	S1	73.33	91.67
			S2	73.33	85
			S3	75.5	**91.75**

Cross-Spectrum analysis For the dark environment, as we shoot using the infrared feature of the camera, we have samples from the NIR spectrum. In our 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. Table 3 shows the performance of the models in a cross-spectrum environment. We get comparable results from existing experiments [19] 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. Recent advancements in domain adaptation methods [22, 23] offer promising opportunities to bridge the performance gap between VIS and NIR spectra.

Table 3. Accuracy of models in Cross-Spectrum environment.

Model	**Test Spectrum**	**Accuracy [%]**
VGG16	VIS	93.52
	NIR	91.25
Inception ResNet (VGGFace2)	VIS	94.8
	NIR	92.6
Inception ResNet (CASIA-Webface)	VIS	**96.56**
	NIR	**93.36**
BEiT	VIS	94.26
	NIR	80.5

5 Conclusion

This work introduces FALEBface, a novel dataset for image processing using body worn cameras. By focusing exclusively on BWCs, the dataset provides a unique benchmark for law enforcement applications. For the preliminary experiments, a comparative analysis is done on the dataset. Fine-tuning the model on the dataset produces high recognition accuracy of 99.25%, 99.75%, and 96.67% respectively for indoor, outdoor, and dark environments. The high performance observed in these evaluations is indicative of the progress in existing algorithms rather than a lack of complexity in the dataset. However, when transitioning from self-environment analysis to cross-environment evaluations, a notable performance drop was observed. Rank-1 accuracy showed a decline of around 10 points when testing across different environments, indicating that while environment-specific performance is strong, cross-environment generalization remains challenging. Comparable results are also obtained for cross-spectrum environments (VIS and NIR spectra). For future work, we aim to extend the dataset by incorporating images captured with standard cameras and comparing them with those obtained from BWCs. This approach reflects real life scenarios where law

enforcement often has access to images of suspects captured by standard cameras in their databases (gallery), while relying on BWCs during street-level operations. Apart from face recognition, there is limited work in the literature combining action recognition with BWCs. While action recognition has been extensively studied in the community, its application using BWCs remains relatively limited; creating an exciting avenue for new research and advancements in the field. As a part of our ongoing research on BWCs, we aim to experiment with advanced action recognition techniques using BWCs, with actions that are specifically relevant to law enforcement scenarios.

References

1. Imagenet, https://www.image-net.org/. Last accessed: 22 February 2025
2. Safe distance, https://www.officer.com/home/article/10248804/safely-handling-suspicious-person-stops, https://www.officer.com/home/article/10248804/safely-handling-suspicious-person-stops. Last accessed: 22 February 2025
3. Al-Obaydy, W., Sellahewa, H.: On using high-definition body worn cameras for face recognition from a distance. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N.C., Fairhurst, M.C. (eds.) Biometrics and ID Management. pp. 193–204. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
4. Almadan, A., Krishnan, A., Rattani, A.: Bwcface: Open-set face recognition using body-worn camera. arXiv preprint arXiv:2009.11458 (2020)
5. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=p-BhZSz59o4
6. Bryan, J.: Effects of Movement on Biometric Facial Recognition in Body-Worn Cameras. PhD thesis, Purdue University Graduate School (5 2020). https://doi.org/10.25394/PGS.12227372.v1
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face amp; Gesture Recognition (FG 2018). pp. 67–74. IEEE Computer Society, Los Alamitos, CA, USA (may 2018). https://doi.org/10.1109/FG.2018.00020, https://doi.ieeecomputersociety.org/10.1109/FG.2018.00020
8. Choi, S., Michalski, N.D., Snyder, J.A.: 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 (2023). https://doi.org/10.1177/07340168221093549
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
10. Elgharib, M., BR, M., Tewari, A., Kim, H., Liu, W., Seidel, H.P., Theobalt, C.: Egoface: Egocentric face performance capture and videorealistic reenactment (05 2019). https://doi.org/10.48550/arXiv.1905.10822
11. Hans, S., Dugelay, J., Isa, M.R.M.: Identifying individuals through egocentric motion: A study using body worn cameras. In: IEEE (ed.) IWBF 2025, International

Workshop on Biometrics and Forensics, 24-25 April 2025, Munich, Germany. Munich (2025)
12. Hans, S., Dugelay, J., Isa, M.R.M., Khairuddin, M.A.: 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. pp. 605–612. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013151900003905
13. King, D.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (07 2009). https://doi.org/10.1145/1577069.1755843
14. Meng, Z., Sánchez, J., Morel, J.M., Bertozzi, A.L., Brantingham, P.J.: Ego-motion classification for body-worn videos. In: Tai, X.C., Bae, E., Lysaker, M. (eds.) Imaging, Vision and Learning Based on Optimization and PDEs. pp. 221–239. Springer International Publishing, Cham (2018)
15. Mustafa, W., Kader, M.: A review of histogram equalization techniques in image enhancement application. Journal of Physics: Conference Series **1019**, 012026 (06 2018). https://doi.org/10.1088/1742-6596/1019/1/012026
16. Nakada, M., Wang, H., Terzopoulos, D.: Acfr: Active face recognition using convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 35–40 (07 2017). https://doi.org/10.1109/CVPRW.2017.11
17. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015), https://arxiv.org/abs/1511.08458
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
19. Siddiqui, N.J.: Novel approach for face recognition using cross-spectral environment. International Journal of Research in Engineering and Science (IJRES) **09 Issue 10**, 70–76 (2021)
20. Suat Cubukcu, Nusret Sahin, E.T., Topalli, V.: The effect of body-worn cameras on the adjudication of citizen complaints of police misconduct. Justice Quarterly **40**(7), 999–1023 (2023). https://doi.org/10.1080/07418825.2023.2222789
21. Todak, N., Gaub, J.E., White, M.D.: Testing the evidentiary value of police body-worn cameras in misdemeanor court. Crime & Delinquency **70**(4), 1249–1273 (2024). https://doi.org/10.1177/00111287221120185
22. Wen, G., Chen, H., Cai, D., He, X.: Improving face recognition with domain adaptation. Neurocomputing **287** (02 2018). https://doi.org/10.1016/j.neucom.2018.01.079
23. Yang, Y., Hu, W., Lin, H., Hu, H.: Robust cross-domain pseudo-labeling and contrastive learning for unsupervised domain adaptation nir-vis face recognition. IEEE Transactions on Image Processing **PP**, 1–1 (09 2023). https://doi.org/10.1109/TIP.2023.3309110
24. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014), https://arxiv.org/abs/1411.7923