Supplementary Material - MS-EVS: Multispectral event-based vision for deep learning based face detection

Saad Himmi* Vincent Parret† Ajad Chhatkuli* Luc Van Gool*

1. Dataset Collection

1.1. Consent

Capturing and publishing human data should be carried out with care. Following the Computer Vision Foundation ethics guidelines [4], all the shared data has been consented to, explicitly, in writing. Some individuals agreed to be captured but only for internal use, therefore some sequences used to compute our results are not shared publicly with N-SpectralFace.

1.2. Hardware setup

For both N-SpectralFace and Real-SpectralFace, the data has been captured with two cameras and a beamsplitter. We share these optical setups in fig. 1. In order to capture both datasets, we connect the cameras to a single laptop that runs the recording scripts. The beamsplitter is essential to artificially increase the number of multispectral bands we capture at once. For N-Spectral face, we use:

- A multispectral CMS-C camera from Silios Technologies [13]. It captures data over 9 bands: 8 bands in the visible (430nm to 700nm) and 1 grayscale (panchromatic) band over all the visible spectrum (c.f. fig. 2). In good lighting conditions, it records at a framerate of 60fps.

- A Basler dart grayscale camera (reference: daA1920-160um) with a long-pass infrared (IR) filter mounted in front. In good lighting conditions, it can record up to 164fps.

For Real-SpectralFace, we use two identical DAVIS328 [2] event cameras, but vary the light filter in front. In fig 1b, the left event camera has an IR cut, a filter that only lets visible light through, and the right event camera has the same IR long-pass as used in N-SpectralFace (cut-off $850nm$). Without a beamsplitter, it would not have been possible to capture multispectral events for Real-SpectralFace that capture the exact same scene. In a stereo camera setup, the spectral bands would have a slightly offset viewpoint. The processing of the data is detailed in section 2.

1.3. N-SpectralFace dataset composition

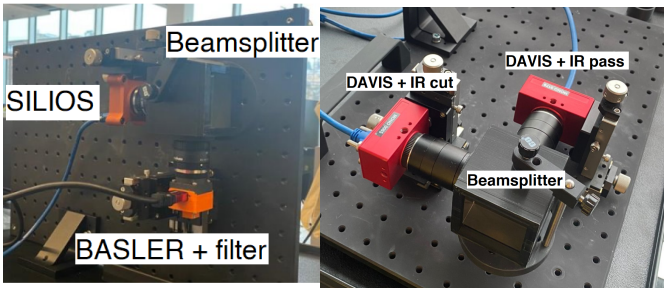
For the dataset, our goal is to feature multiple people, in different places with different lighting conditions. The lighting is an important parameter as it directly relates to the amount of infrared light one has in the scene. We distinguish between three possibilities:

- Outdoor sequences. Thanks to sunlight, there is a large amount of infrared light (c.f. fig. 12).

- Indoor sequences in a bright room. Thanks to the glass walls, enough sunlight enters the room and the scene has enough infrared light.

- Other indoor sequences. All the indoor lights being LED, no infrared light is initially present in the scene (c.f. fig. 12). To alleviate this issue and have meaningful infrared data for these sequences, we use an incandescent bulb pointed towards the scene we capture.

N-SpectralFace is diverse but not strictly balanced. The people and place proportions are respectively reported in figures 3a and 3b.



(a) N-SpectralFace (b) Real-SpectralFace

Figure 1. Beamsplitter and camera setups. Each subfigure represents the setup of a single dataset. In each pair of cameras, the camera lenses are identical. Best viewed in color.

*Computer Vision Laboratory, ETH Zürich, Switzerland
†Stuttgart Laboratory 1, Sony Semiconductor Solutions Europe, Sony Europe B.V.

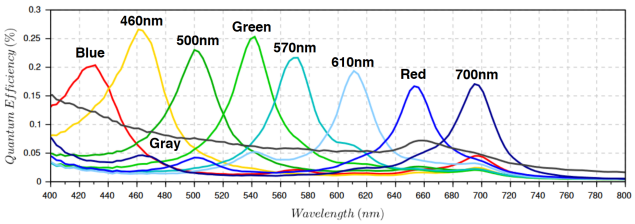
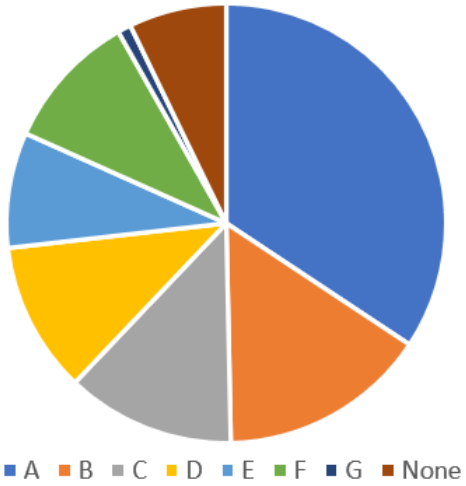
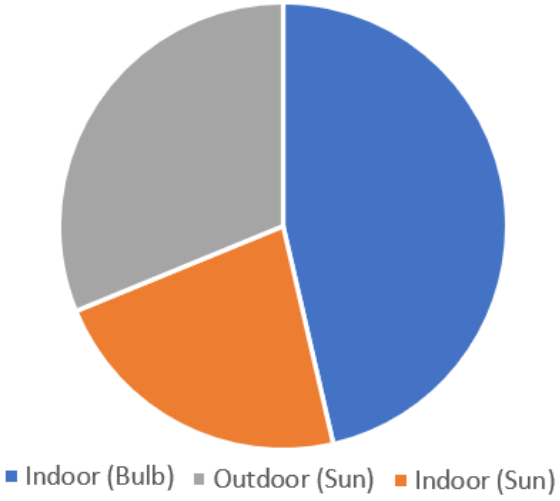


Figure 2. Quantum Efficiency of the SILIOS multispectral camera, adapted from [13]. Warning: the colors of the bands do not correspond to their wavelength. Notice how the grayscale band is almost equally sensitive for all visible bands. Best viewed in color.



(a) People. Each letter is a different person.



(b) Recording places (and associated lighting).

Figure 3. N-SpectralFace full dataset composition (58 sequences). Focus on the people and places distribution. Best viewed in color.

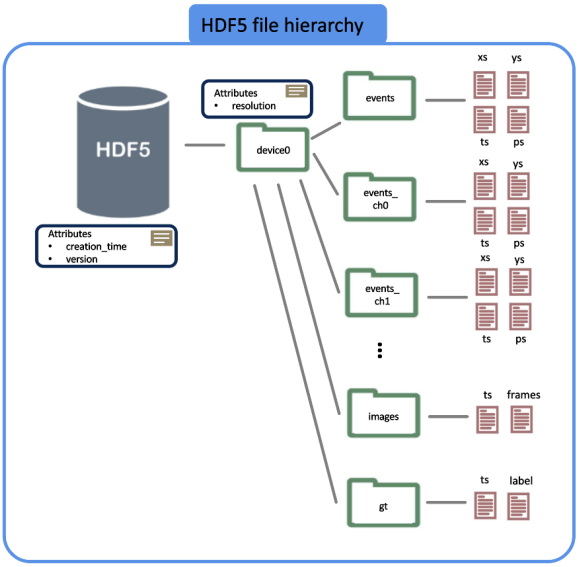


Figure 4. HDF5 hierarchical file structure for each sequence. Multispectral events, conventional images and ground truth are stored in the same dictionary. Notice how each branch has a "ts" leaf for timestamps, all the data in an H5 file is aligned in time (unique time reference).

2. Preparing the datasets

2.1. Data format

On the one hand, representing an image in the memory is relatively easy. All the pixel values are stored in a contiguous array. If the image is multispectral, it is then saved as a contiguous 3-dimensional tensor, with N channels. On the other hand, storing event-based data is less trivial. Indeed, event cameras capture an asynchronous and continuous stream of events, and each event is a (x, y, t, p) tuple. Moreover, N-MobiFace, N-YoutubeFaces, and N-SpectralFace are bimodal datasets, meaning it consists of conventional frames and event-based data aligned on a single timescale. Contiguous arrays are obviously not a good data structure to store our multispectral events and conventional frames. We choose to use the hierarchical HDF5 file structure, represented in fig. 4. An H5 file can be described as a dictionary where the terminal nodes (leaves) are our data. For example, the *"device0/events"* branch has 4 leaves corresponding to location, polarity and timestamp of each event. The leaves are stored as lists and all the leaves of a single branch have the same length (e.g. the number of frames or the number of events). Multispectral events are simply represented as N branches, each corresponding to a single spectral band.

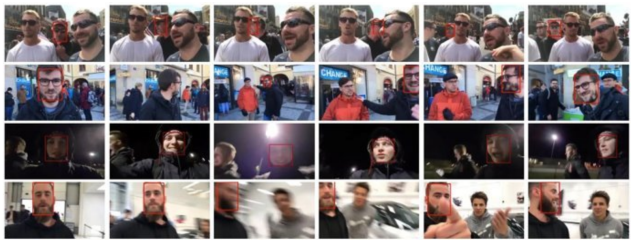


Figure 5. Example samples from Mobiface, adapted from [7]. Notice the occlusions, out of plane rotation, illumination variations and motion blur. Best viewed in color.

Figure 6. Example samples from YoutubeFaces, adapted from [14]. Notice that the faces to be detected are less challenging. Best viewed in color.

2.2. Data cleaning

Some sequences from MobiFace [7] and YoutubeFaces [14] were not relevant for our work and kept out of their respective Neuromorphic version. In particular, multiple sequences in YoutubeFaces consist of a still face, e.g. when the extract is only a voiceover and a picture of a celebrity. However, in these situations we would end up with a labeled face but no events at all to detect it, as static scenes captured by a static camera do not generate events. Therefore, we did not include these sequences in N-YoutubeFaces. Also, we do not use the portrait sequences in N-MobiFace in our training and validation because it consists of only five sequences out of approx. 3500 total landscape sequences (both N-datasets combined). Refer to fig. 5 and 6 for some samples of N-MobiFace and N-YoutubeFaces.

2.3. Data labelling

YoutubeFaces [14] and MobiFace [7] are both labeled for single face tracking and therefore many face labels are missing. We present an approach to fix the labels but it can also be used to label a face dataset from scratch. As we have access to ground truth labels for some faces, we opt for reusing this information. For each image in a video sequence, we run inference for an ensemble of N face detection methods. We save all the outputs and their confidence if it is above a confidence threshold (preferably high). Naturally, ground truth labels have a confidence of 1.0. Then, we run Non-Maximum Suppression (NMS) with a low Intersection-over-Union threshold to only keep the most confident labels in the image and discard duplicates. After this operation, a new set of pseudo ground truth labels for the image is created. If there was only a single face in the sample, the original ground truth is automatically kept. If not, it is very likely that at least one of the N labeling detectors will detect it and it will be labeled. For N-MobiFace and N-YoutubeFaces, we used two face detection neural networks (N=2): MTCNN[1] [16] and Yolov5[2] trained on CrowdHuman [11]. For N-SpectralFace and Real-SpectralFace, only Yolov5 is used (N=1) to enforce label consistency. The auto-labeling process steps are shown in fig. 7. Figure 8 shows that the computed groundtruth labels transfer well to the simulated EVS data.

3. Training details

3.1. Training from scratch?

When training our baseline model on N-MobiFace and N-YoutubeFaces, we experimented with using the ImageNet [3] pretrained weights for ResNet (included in PyTorch) as well as training everything from scratch. The experiments we carried showed us that the ImageNet weights greatly improve the APS performance but not the EVS performance, compared to training from scratch:

- On the one hand, the APS GS performance (AP@.5) improved by 9% and APS RGB performance improved by 20%.
- On the other hand, the EVS GS performance (AP@.5) improved by 2% only and the EVS RGB performance improved by 8%

From the numbers, it is clear that the ImageNet weights greatly favor APS models over EVS. As there exists no comparable large dataset for EVS images, we train all models from scratch for a more fair comparison. Reproducing this experiment on N-YoutubeFaces showed similar results.

These results are not very surprising as ImageNet is a conventional image-based dataset and nothing guarantees

[1]Model and weights: https://github.com/timesler/facenet-pytorch

[2]Model: https://github.com/ultralytics/yolov5, Weights: https://github.com/deepakcrk/yolov5-crowdhuman

(a) N-Mobiface. Left: Intermediate labels (N detectors), Right: Final labels

(b) N-YoutubeFaces. All samples with final labels.

Figure 7. Illustration of the pseudo-labeling process, before and after the Non-Maximum Suppression (NMS). Red boxes: MTCNN, Blue boxes: Yolov5, Yellow boxes: Initial ground truth, Green boxes: Final label after NMS. Best viewed in color.

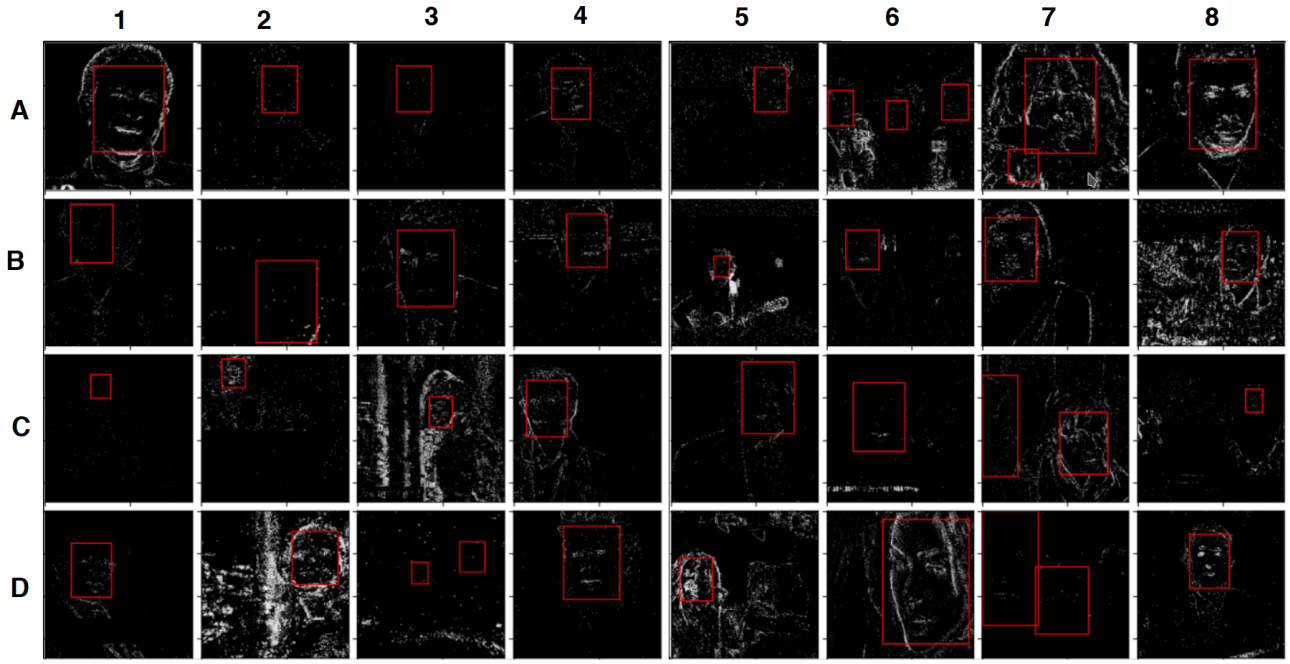


Figure 8. Example of final labels in N-YoutubeFaces, in the EVS domain. Notice the bounding boxes around no events: these are not False Positives but static faces.

that the learned features will be useful for EVS data. In [5], *Geirhos et al. (2018)* show that "[...]ImageNet-trained CNNs are strongly biased towards recognising textures rather than shapes[...]". However, conventional image-based data being dense and high resolution, it provides a lot of information about object textures. On the other hand, event-based data is sparse and based on intensity changes, highlighting object edges and therefore object shapes rather than texture. This could be an explanation to why ImageNet weights initialization do not benefit EVS models as much but it would need further experimentation and analysis. For all the reasons, we chose to train both APS and EVS models from scratch, for a fair comparison.

3.2. Hyperparameters

To reproduce our results, here are the hyperparameters used during training and fine-tuning:

- **Input image size:** [346,346] during training, [256, 393] during fine-tuning.
- **Event frame representation:** Binary image, with $50ms$ time window. Events are aggregated along their time dimension, polarity is ignored.
- **Optimizer:** SGD with 90% momentum or SGD without momentum. For each model, we trained both and kept the best weights
- **Scheduler:** Cosine Annealing with Warm Restarts, a periodic scheduler with increasing period. First restart after 11 epochs, multiply the period by 2 at every restart.
- **Initial learning rate:** 0.01 when training from scratch, 0.001 for fine-tuning (slower).
- **Batch size:** 32 for training, 64 for validation.
- **Number of epochs:** 300, each epoch is composed of 2000 random samples (conventional frame or $50ms$ event-interval).
- **Augmentations:** Random flipping (left-right), in-plane rotations and random cropping.
- **Pretrained weights:** None, not even for the Resnet blocks, c.f section 3.1. During fine-tuning, we used the weights trained on grayscale N-MobiFace and N-YoutubeFaces.

Training is performed on a single GPU with the same random seed.

4. Physics: Other relevant Spectra

In this section, we report some useful spectra:

- The grayscale DAVIS camera sensitivity in figure 9.
- The color DAVIS camera sensitivity in figure 10.
- The SILIOS camera sensitivity in figure 2.
- The Basler camera sensitivity in figure 11.
- Different light emission spectrums in figure 12.
- The skin reflectance spectrum in figure 13.

Note that Camera Sensitivity, Quantum Efficiency (QE), and Spectral Response Curve are similar concepts, they all give an idea of the ability a camera has to capture light from different wavelengths. The difference lies in how this ability is measured and reported, QE gives the proportion of photons that is captured for each wavelength, while the Spectral Response Curve normalizes the contribution of each wavelength with respect to the peak.

5. Dataset Licenses

Mobiface [7] is licensed for research and non-commercial use only. All videos in the dataset are protected by Youtube terms of service [15]. As of today, the dataset is available upon completion of a form[3].

YoutubeFaces [14] also consists of videos downloaded from YouTube and follow the platform's terms of service [15]. The dataset is publicly shared on their website[4]

[3]Link: https://mobiface.github.io/
[4]Link: https://www.cs.tau.ac.il/~wolf/ytfaces/

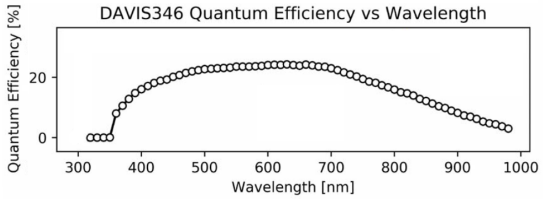


Figure 9. Quantum Efficiency of the DAVIS346 MONO event camera, adapted from [12]. Notice the sensitivity to wavelengths in the near-infrared ($> 800nm$): infrared events can exist.

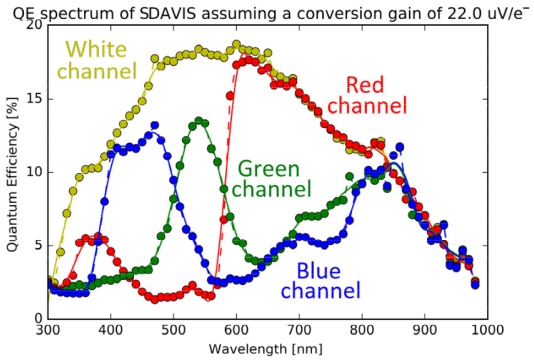


Figure 10. Quantum Efficiency of the Color-DAVIS event camera, adapted from [9]. Notice how all color bands are equally sensitive to near-infrared ($> 800nm$). Best viewed in color.

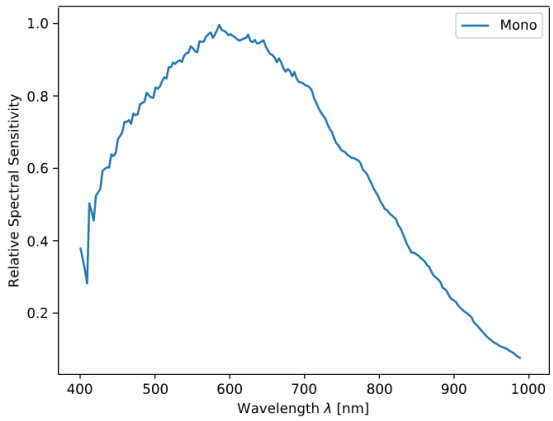


Figure 11. Spectral Response Curve of the Basler grayscale camera, adapted from [1].

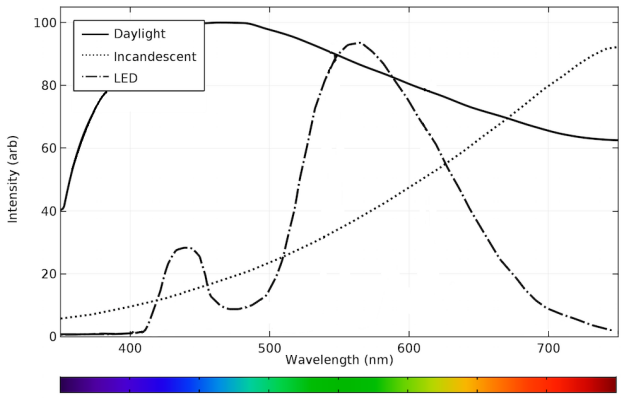


Figure 12. Emission spectrum for outdoor sunlight, LED and incandescent bulb, normalized.

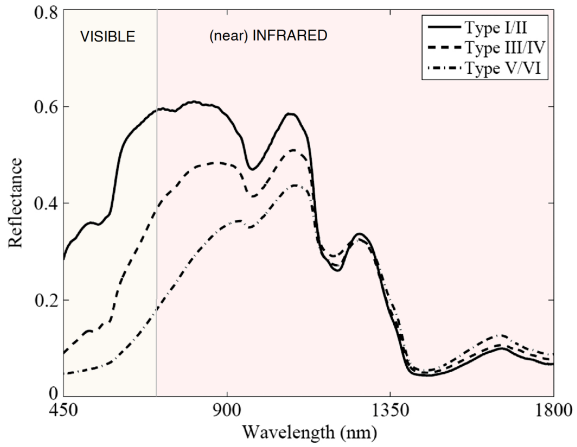


Figure 13. Skin reflectance spectrum for different types of skin, adapted from [8]. Notice how all curves end up converging in the infrared.

6. Qualitative Results

To get a better understanding of face detection on event-based data, and particularly on multispectral events, we show some additional examples in this section. Note that multispectral samples must be "false-colored" in RGB for visualization. In general, we represent the "most blue" band in Blue (lower wavelength), and the two bands with highest wavelength in Green and Red. Therefore, if an input contains infrared, it is guaranteed that it will be represented in red.

Ground truth label is represented in red, and face detection prediction is in green. Please look at figures 14, 15, and 16 for N-SpectralFace examples of the Single channel experiment (Grayscale against Infrared). For multi-channel examples, figures 17, and 18 show a few examples of APS and EVS inferences. Increasing the amount of multispectral channels or introducing infrared both seem to reduce the number of False Positives in the scene, c.f. fig. 17 and 18. Finally, examples on real multispectral events (Real-SpectralFace) are shown in figures 19 and 20.

7. About Multispectral EVS HW feasibility

Note that, when it comes to multispectral sensing, the narrower the band is, the less light intensity the sensor will capture, as we simply truncate part of the Quantum Efficiency curve and therefore the final integral (equation 2) is mathematically less or equal to the initial integral (equation 1). While low light is an issue for regular cameras, the high dynamic range and high sensitivity in low light properties of event cameras make them especially suitable for multispectral imaging. Also, compared to hyperspectral imaging, the bands are relatively wide and therefore irradiance is reasonably high.

Furthermore, we believe that hardware feasibility in practice would not be a problem. As EVS pixels are larger than regular APS pixels, it should be possible to apply the same filters as used in regular cameras pixel-wise. The existing application of polarization filters on EVS pixels [6] or RGB Bayer pattern on EVS pixels [9, 10] also suggest that it would be feasible.

Figure 14. Face detection on three samples (Indoor, with an incandescent bulb). Left is GS data, right is IR. The first row is APS, others are EVS. The first two rows show a static face, while the third row shows a moving face. Ground truths are in red, predictions in green. Notice the different event density around the face when there is a movement, vs. when there is none.

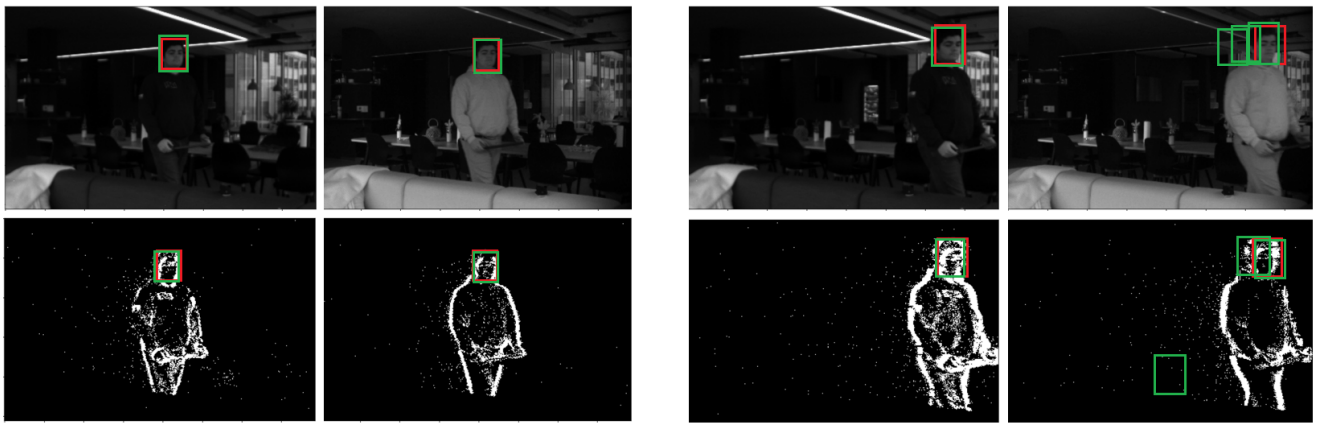
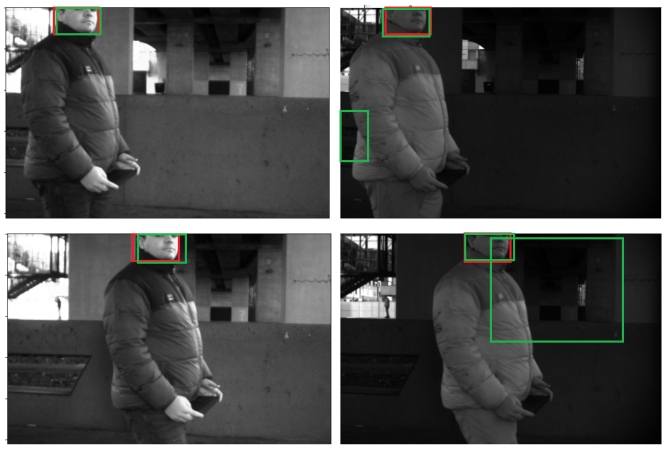
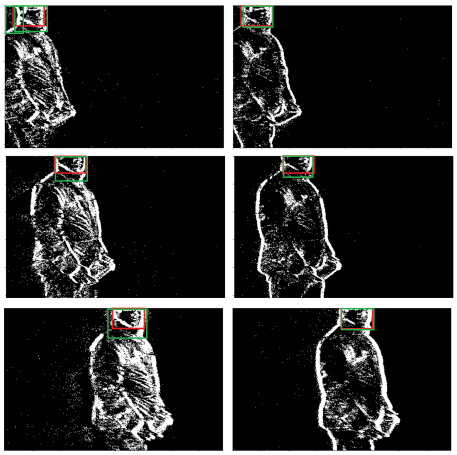


Figure 15. Face detection on two samples from N-SpectralFace (Indoor with sunlight). For each group of pictures, the first row is APS GS and APS IR, and the second row is EVS GS and EVS IR. Ground truths are in red, predictions in green. Notice how the single channel IR models show more false positives.

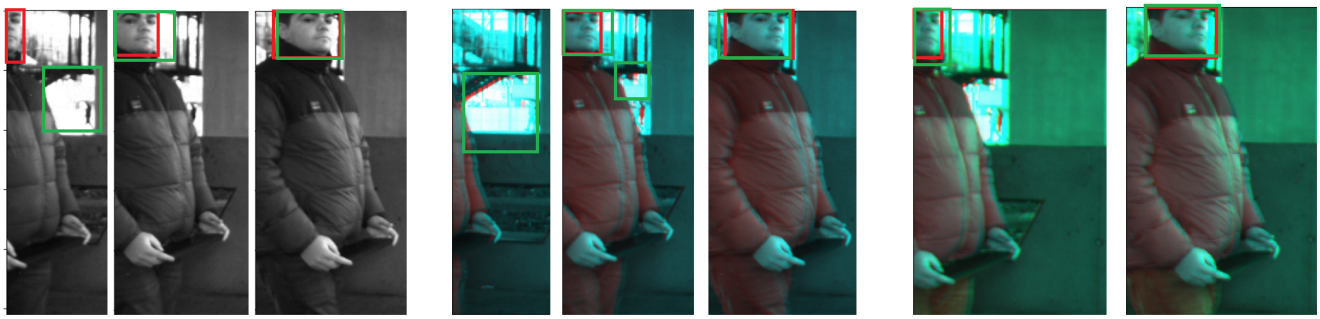


(a) APS samples. Left: GS (good), Right: IR, notice the False Positives.

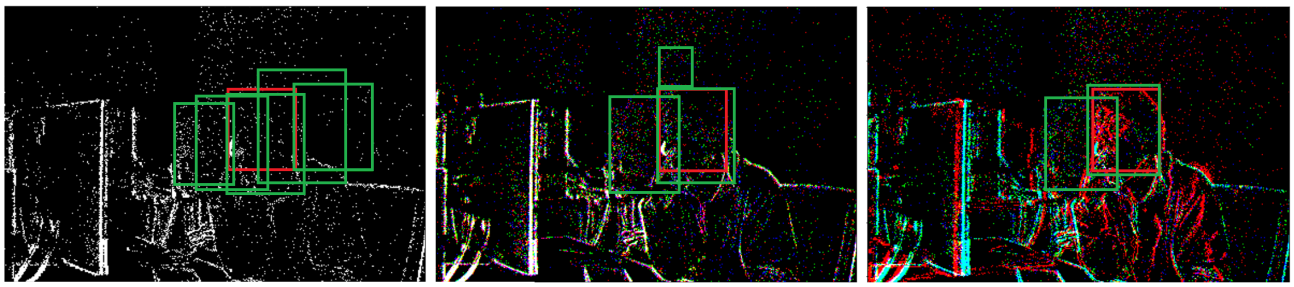


(b) EVS samples. Left: GS, Right: IR

Figure 16. Face detection on outdoor samples (with sunlight), N-SpectralFace. Ground truths are in red, predictions in green.



(a) APS data, outdoor. From left to right: GS, GS+IR, BGR+IR. Notice how the predictions are getting better when more multispectral bands are added (from left to right).



(b) EVS data, indoor with bulb. From left to right: GS, BGR, BGR+IR.

Figure 17. Face detection on multispectral samples from N-SpectralFace. Ground truths are in red, predictions in green. Notice how the number of False Positives decreases for EVS models from left to right: multispectral EVS models are clearly more robust, which could explain the better mAP. Best viewed in color.

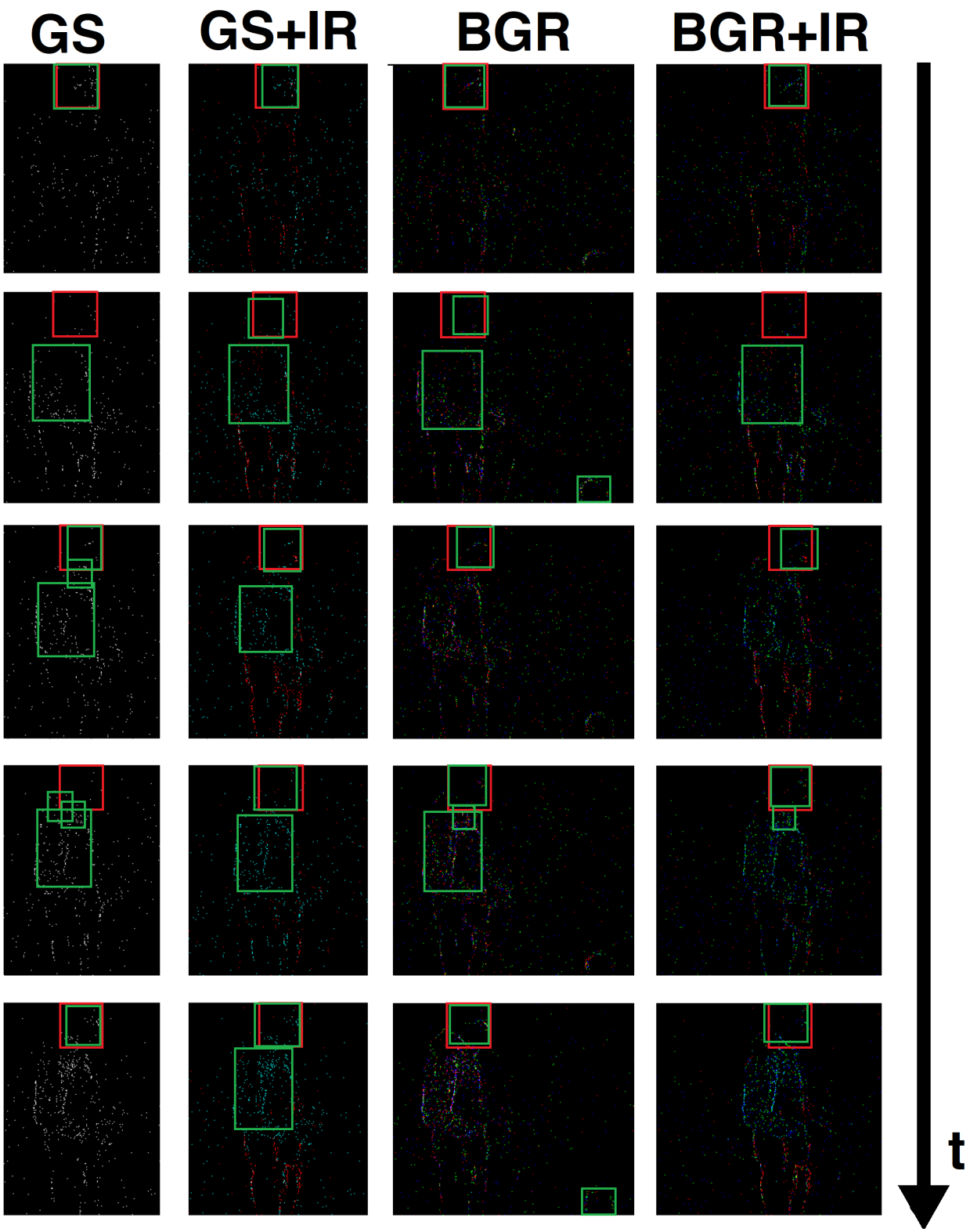


Figure 18. Sequence of face detection examples on multispectral events from N-SpectralFace, through time. Each column is a different set of input channels. From top to bottom, the frames are consecutive. Some samples are cropped for space constraints. Ground truths are in red, predictions in green. Notice the different False positives and False negatives in the scene: the best model is BGR+IR. Best viewed in color.

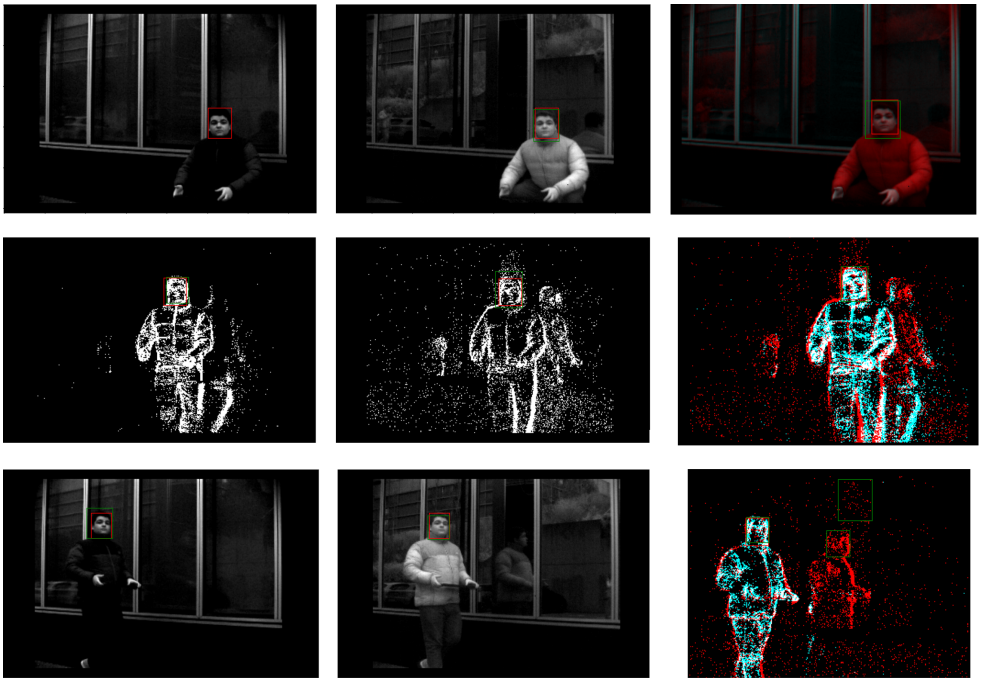
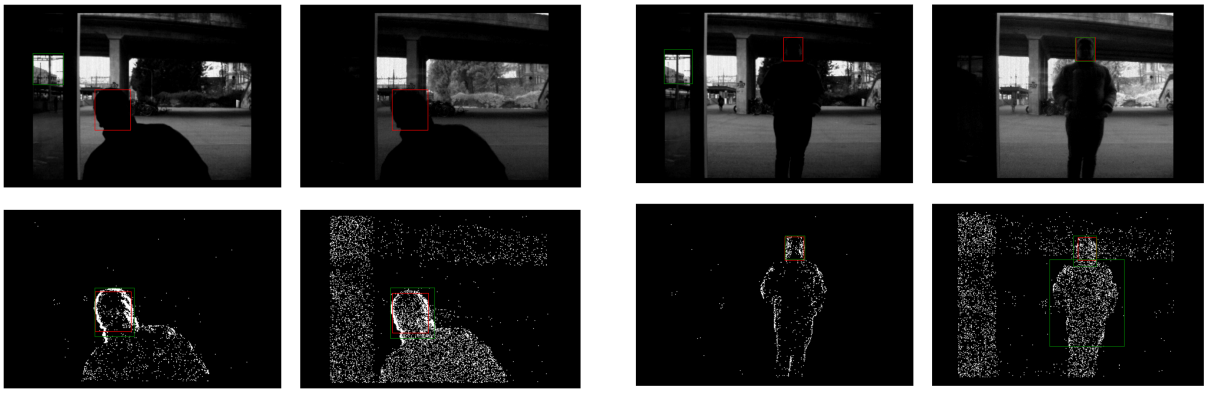


Figure 19. Mix of APS and EVS samples from the Real-SpectralFace dataset, with real multispectral events. Each column is a different set of spectral bands, in order: GS, IR and GS+IR. Notice how APS IR have a better contrast than APS GS for face detection. Best viewed in color.



(a) First sequence

(b) Second sequence

Figure 20. Face detection examples from Real-SpectralFace in a scene that needs High Dynamic Range (HDR). Top: APS GS and APS IR. Bottom: EVS GS and EVS IR. Notice how hard it is to distinguish the face in the APS samples, that is why it is never detected, while with EVS the samples are clear, regardless of spectral the band. Best viewed in color.

References

[1] Basler AG. Basler product documentation — daa1920-160um, 2023. Available at https://docs.baslerweb.com/daa1920-160um#. Accessed on 30.06.2023. 6

[2] Christian Brändli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. *Solid-State Circuits, IEEE Journal of*, 49:2333–2341, 10 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[4] The Computer Vision Foundation. Ethics guidelines for authors, 2023. available at https://cvpr2023.thecvf.com/Conferences/2023/EthicsGuidelines. Accessed on 29.06.2023. 1

[5] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *CoRR*, abs/1811.12231, 2018. 5

[6] Germain Haessig, Damien Joubert, Justin Haque, Moritz B. Milde, Tobi Delbruck, and Viktor Gruev. Pdavis: Bio-inspired polarization event camera. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3963–3972, 2023. 6

[7] Yiming Lin, Shiyang Cheng, Jie Shen, and Maja Pantic. Mobiface: A novel dataset for mobile face tracking in the wild, 2019. 3, 5

[8] Michael J. Mendenhall, Abel S. Nunez, and Richard K. Martin. Human skin detection in the visible and near infrared. *Appl. Opt.*, 54(35):10559–10570, Dec 2015. 6

[9] Diederik Paul Moeys, Chenghan Li, Julien N.P. Martel, Simeon Bamford, Luca Longinotti, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Color temporal contrast sensitivity in dynamic vision sensors. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2017. 5, 6

[10] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2021. 6

[11] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 3

[12] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 5

[13] Silios Technologies. Cms series: Multispectral cameras. Available at https://www.silios.com/cms-series. Accessed on 28.06.2023. 1, 2

[14] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011. 3, 5

[15] YouTube. Terms of service, 2023. Available at https://www.youtube.com/static?template=terms. Accessed on 29.06.2023. 5

[16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. 3