Deep Saliency Map Generators for Multispectral Video Classification

Companion paper for the RRPR 2022 Workshop

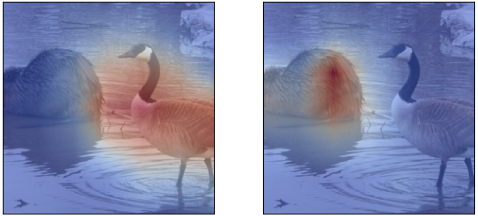
Jens Bayer, David Münch, Michael Arens
Department Object Recognition, Fraunhofer IOSB
21.08.2022

Fraunhofer IOSB

Motivation

[0]

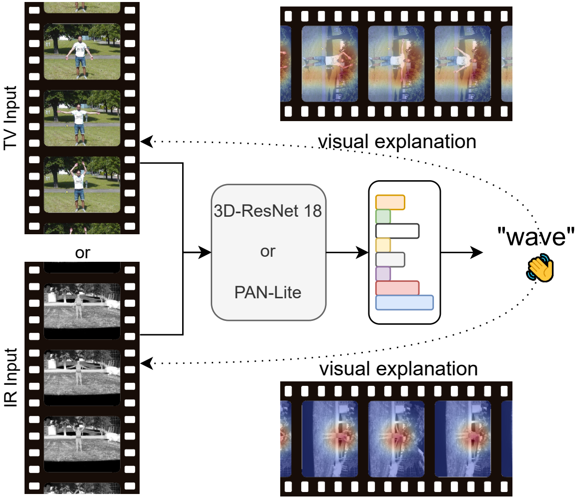
- Saliency Maps give a hint on what is important for a networks decision.
- Usually, the methods are applied to ordinary images.



Goose and Beaver “explained” with Grad-CAM.

Motivation

- Saliency Maps give a hint on what is important for a networks decision.
- Usually, the methods are applied to ordinary images.
- *How do these methods behave when applied to three-dimensional multispectral input data?*



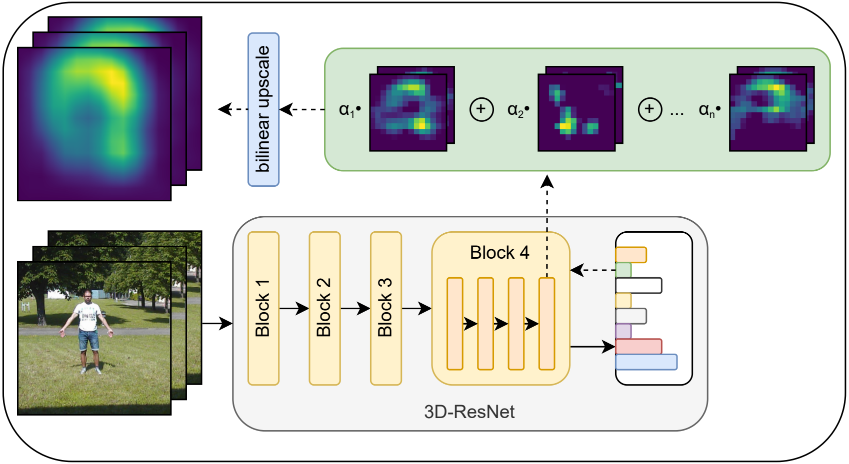
Fraunhofer IOSB

Contributions

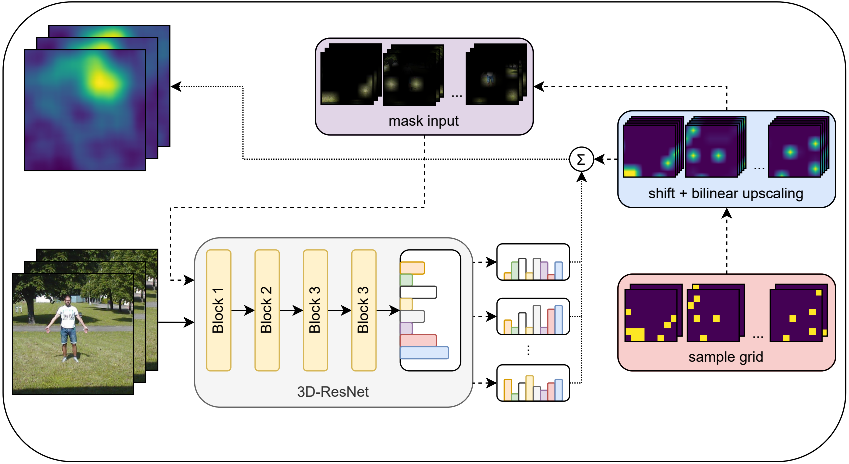
- Adapt and evaluate three saliency map generators for multispectral 3D input data.
- Exemplarily shown on video input data for human action recognition with IR and TV images.
- Investigated two different network architectures.

Fraunhofer IOSB

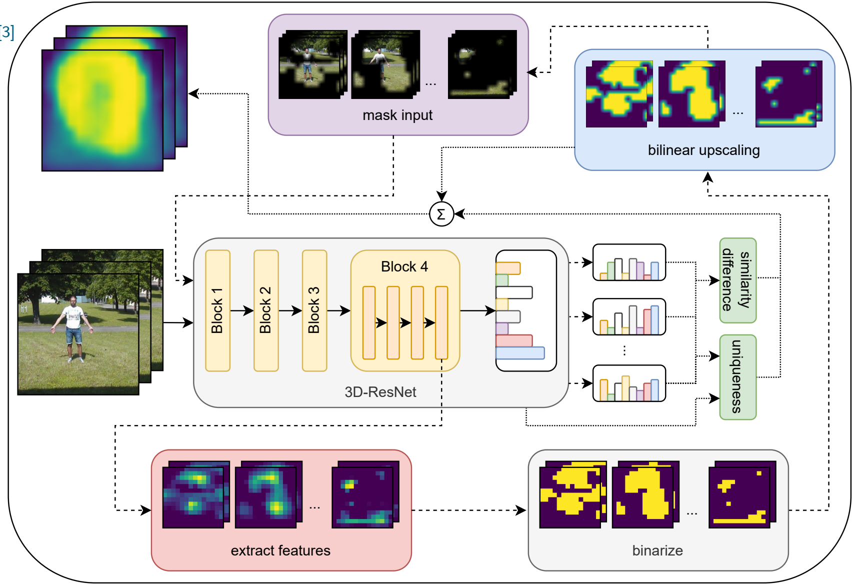
Grad-CAM [1]


bilinear upscale
α1•
α2•
αn•
...
Block 1
Block 2
Block 3
Block 4
3D-ResNet

RISE [2]


mask input
...
shift + bilinear upscaling
...
Σ
Block 1
Block 2
Block 3
Block 3
3D-ResNet
sample grid
...

SIDU [3]



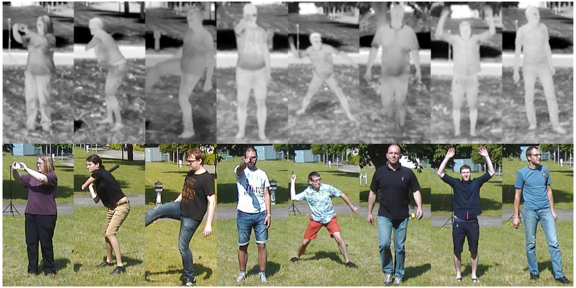
Fraunhofer IOSB

Networks

- 3D-ResNet 18
  - ResNet-based network for video input data.
  - 3D filter kernels.
- Persistent Appearance Networks (PAN)
  - Follows the two-stream paradigma.
  - Persistence of Appearance motion cue instead of optical flow.
  - Pre-trained ResNet50 backbone.

Multispectral Action Dataset [4]

- Eight different Actions.
- Ten different Actors.
- Two different Camera perspectives.
- Recorded in the thermal infrared and the visual domain.
- Resolution:
  - IR: 640×480px
  - TV: 960×540px
- Train and test split with ratio 8:2.



Fraunhofer IOSB

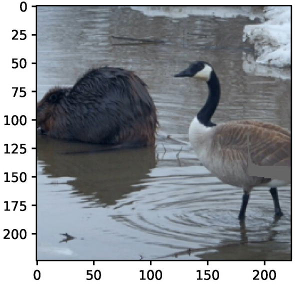
Experimental Setup

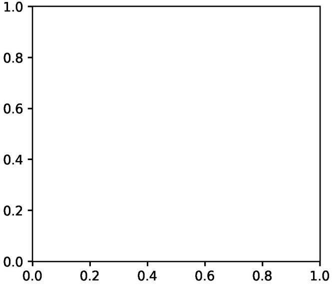
- All networks achieved a 0.9 test accuracy.
- All experiments use the same train and test split for IR and TV.
- A sequence length of 32 frames is ensured.
- Sequences are rescaled to a fixed 256px height and center cropped to 224×224px.

Metrics: Deletion Metric [2]

- Remove pixels successively according to the saliency map.
- The more pixels are removed, the more the probability of the predicted class decreases.
- If the most crucial pixels are removed first, there is a sharp drop in the curve.

⇒ Area under the curve should be close to zero.

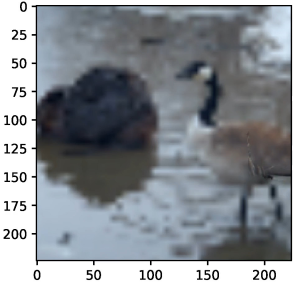


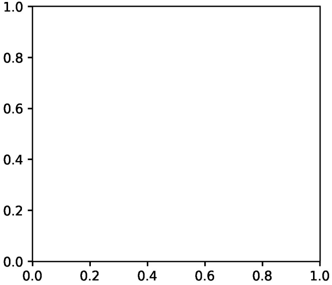


Metrics: Insertion Metric [2]

- Blur the input.
- Replace the blurred pixels with the unblurred one successively, according to the saliency map.
- The more pixels are recovered, the more the probability of the predicted class increses.
- If the most crucial pixels are recovered first, there is a sharp rise in the curve.

$\Rightarrow$ Area under the curve should be close to one.





Fraunhofer IOSB

Results: IR input data

Method and Trained Spectrum	3D-ResNet 18		PAN	
	Deletion $\downarrow$	Insertion $\uparrow$	Deletion $\downarrow$	Insertion $\uparrow$
Grad-CAM IR	0.16 ± 0.22	0.71 ± 0.31	0.29 ± 0.28	0.73 ± 0.29
RISE IR	0.19 ± 0.25	0.54 ± 0.34	0.26 ± 0.26	0.58 ± 0.36
SIDU IR	0.15 ± 0.20	0.69 ± 0.32	0.25 ± 0.25	0.73 ± 0.30
Grad-CAM IRTV	$\mathbf{0.15} \pm \mathbf{0.20}$	$\mathbf{0.83} \pm \mathbf{0.25}$	0.18 ± 0.21	$\mathbf{0.77} \pm \mathbf{0.28}$
RISE IRTV	0.17 ± 0.22	0.62 ± 0.32	$\mathbf{0.17} \pm \mathbf{0.23}$	0.58 ± 0.34
SIDU IRTV	0.16 ± 0.22	0.78 ± 0.28	0.18 ± 0.21	0.75 ± 0.29

Results: TV input data

Method and Trained Spectrum	3D-ResNet 18		PAN	
	Deletion $\downarrow$	Insertion $\uparrow$	Deletion $\downarrow$	Insertion $\uparrow$
Grad-CAM TV	0.16 ± 0.22	0.64 ± 0.33	0.28 ± 0.35	0.65 ± 0.31
RISE TV	0.18 ± 0.23	0.44 ± 0.32	0.31 ± 0.32	0.47 ± 0.34
SIDU TV	$\mathbf{0.15} \pm \mathbf{0.20}$	0.63 ± 0.33	0.30 ± 0.34	0.61 ± 0.33
Grad-CAM IRTV	0.15 ± 0.23	$\mathbf{0.79} \pm \mathbf{0.28}$	$\mathbf{0.22} \pm \mathbf{0.25}$	0.76 ± 0.28
RISE IRTV	0.16 ± 0.22	0.69 ± 0.30	0.24 ± 0.29	0.62 ± 0.34
SIDU IRTV	0.17 ± 0.23	0.75 ± 0.29	0.23 ± 0.25	$\mathbf{0.79} \pm \mathbf{0.26}$

Modified parameters

- Grad-CAM:
  - None
- RISE:
  - Number of masks
  - Grid size
  - Flip probability
- SIDU:
  - Binarization threshold

Parameters: RISE - Number of masks

Image Spectrum	Parameters	3D-ResNet 18		PAN	
		Deletion $\downarrow$	Insertion $\uparrow$	Deletion $\downarrow$	Insertion $\uparrow$
TV	$n = 10^2$	0.18 ± 0.13	0.55 ± 0.23	0.27 ± 0.18	0.51 ± 0.22
	$n = 10^3$	$\mathbf{0.17 \pm 0.09}$	0.60 ± 0.24	0.22 ± 0.18	0.59 ± 0.22
	$n = 10^4$	0.18 ± 0.08	$\mathbf{0.62 \pm 0.24}$	$\mathbf{0.22 \pm 0.17}$	$\mathbf{0.61 \pm 0.24}$
IR	$n = 10^2$	0.18 ± 0.13	0.48 ± 0.27	0.17 ± 0.14	0.51 ± 0.23
	$n = 10^3$	0.16 ± 0.10	0.54 ± 0.26	0.16 ± 0.14	0.54 ± 0.23
	$n = 10^4$	$\mathbf{0.16 \pm 0.08}$	$\mathbf{0.57 \pm 0.26}$	$\mathbf{0.16 \pm 0.13}$	$\mathbf{0.55 \pm 0.23}$

Parameters: RISE - Grid size

Image Spectrum	Parameters	3D-ResNet 18		PAN	
		Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑
TV	$s = 2 \times 8 \times 8$	0.18 ± 0.10	$\mathbf{0.64 \pm 0.20}$	$\mathbf{0.23 \pm 0.17}$	$\mathbf{0.58 \pm 0.24}$
	$s = 4 \times 8 \times 8$	0.18 ± 0.19	0.56 ± 0.22	0.24 ± 0.17	0.50 ± 0.23
	$s = 8 \times 8 \times 8$	$\mathbf{0.16 \pm 0.21}$	0.43 ± 0.30	0.26 ± 0.14	0.45 ± 0.22
IR	$s = 2 \times 8 \times 8$	0.18 ± 0.10	$\mathbf{0.61 \pm 0.23}$	$\mathbf{0.17 \pm 0.14}$	$\mathbf{0.54 \pm 0.22}$
	$s = 4 \times 8 \times 8$	$\mathbf{0.18 \pm 0.12}$	0.61 ± 0.24	0.18 ± 0.12	0.54 ± 0.24
	$s = 8 \times 8 \times 8$	0.18 ± 0.16	0.58 ± 0.23	0.22 ± 0.13	0.50 ± 0.23

Parameters: RISE - Flip probability

Image Spectrum	Parameters	3D-ResNet 18		PAN	
		Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑
TV	$p = 0.10$	0.18 ± 0.10	$\mathbf{0.61 \pm 0.23}$	0.23 ± 0.18	0.60 ± 0.23
	$p = 0.25$	$\mathbf{0.18 \pm 0.12}$	0.61 ± 0.24	$\mathbf{0.23 \pm 0.19}$	$\mathbf{0.66 \pm 0.24}$
	$p = 0.50$	0.18 ± 0.16	0.58 ± 0.23	0.27 ± 0.21	0.65 ± 0.26
IR	$p = 0.10$	0.18 ± 0.10	0.58 ± 0.23	$\mathbf{0.17 \pm 0.15}$	0.53 ± 0.23
	$p = 0.25$	$\mathbf{0.17} \pm 0.09$	$\mathbf{0.59 \pm 0.23}$	0.17 ± 0.16	$\mathbf{0.62 \pm 0.23}$
	$p = 0.50$	0.18 ± 0.12	0.56 ± 0.24	0.19 ± 0.17	0.60 ± 0.24

Parameters: SIDU - Binarize threshold

Image Spectrum	τ	3D-ResNet 18		PAN	
		Deletion $\downarrow$	Insertion $\uparrow$	Deletion $\downarrow$	Insertion $\uparrow$
TV	-1	0.35 ± 0.17	0.46 ± 0.19	0.40 ± 0.21	0.49 ± 0.22
	-0.5	0.35 ± 0.17	0.46 ± 0.19	0.40 ± 0.21	0.49 ± 0.22
	0	$\mathbf{0.18 \pm 0.08}$	$\mathbf{0.67 \pm 0.18}$	0.26 ± 0.14	0.62 ± 0.25
	0.5	0.18 ± 0.09	0.67 ± 0.18	0.23 ± 0.16	$\mathbf{0.68 \pm 0.22}$
	1	0.18 ± 0.09	0.67 ± 0.19	$\mathbf{0.23 \pm 0.15}$	0.68 ± 0.22
IR	-1	0.29 ± 0.15	0.42 ± 0.19	0.36 ± 0.18	0.48 ± 0.23
	-0.5	0.29 ± 0.15	0.42 ± 0.19	0.36 ± 0.18	0.48 ± 0.23
	0	0.17 ± 0.08	0.64 ± 0.20	0.21 ± 0.12	0.59 ± 0.24
	0.5	$\mathbf{0.17 \pm 0.08}$	$\mathbf{0.64 \pm 0.20}$	$\mathbf{0.18 \pm 0.12}$	$\mathbf{0.63 \pm 0.21}$
	1	0.17 ± 0.08	0.63 ± 0.20	0.18 ± 0.12	0.63 ± 0.21

Conclusion

- The investigated methods can be used with three-dimensional input data, yet the performance differs.
- Grad-CAM outperforms RISE and SIDU.
- RISE:
  - More masks lead to better results with RISE but increases the computation time significantly.
  - A smaller temporal grid resolution improves the metric scores.
  - Only a small impact of the flip probability.
- SIDU:
  - The default value for the threshold results in most cases the best or close to the best metric scores.

I am looking forward to your questions and comments!

References I

[0] https://www.theweathernetwork.com/photos/view/animals/beaver-and-goose-canadian-icons/34747590

[1] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. ISSN 15731405. doi: 10.1007/s11263-019-01228-7.

[2] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

[3] Satya M. Muddamsetty, N. S. Jahromi Mohammad, and Thomas B. Moeslund. SIDU: Similarity Difference And Uniqueness Method for Explainable AI. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3269–3273. IEEE, 10 2020. ISBN 978-1-7281-6395-6. doi: 10.1109/ICIP40778.2020.9190952. URL http://arxiv.org/abs/2101.10710https://ieeexplore.ieee.org/document/9190952/.

[4] Barbara Hilsenbeck, David Münch, Ann-Kristin Grosselfinger, Wolfgang Hubner, and Michael Arens. Action Recognition in the Longwave Infrared and the Visible Spectrum Using Hough Forests. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 329–332. IEEE, 12 2016. ISBN 978-1-5090-4571-6. doi: 10.1109/ISM.2016.0072. URL https://ieeexplore.ieee.org/document/7823639/.

Appendix

Saliency maps