Impact of Fused Visible-Infrared Video Streams on Visual Tracking

Stéphane Vujasinović(✉), Stefan Becker, Norbert Scherer-Negenborn, and Michael Arens

Fraunhofer Institute for Optronics, System Technologies,
and Image Exploitation IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany
stephane.vujasinovic@iosb.fraunhofer.de

Abstract. Currently state-of-the-art trackers rely on fully convolutional neural network (FCNN) for extracting salient features in order to create an appearance representation of the target. Ordinarily, most of them intend to work with input streams from the visible spectrum, yet how does an input stream from the infrared spectrum and a fused visible-infrared stream affect their performances and how does it benefit or detriment them? Towards this end, we compare the performance of various reference trackers utilizing FCNN for feature extraction, on visible, infrared and fused spectrums. By utilizing a carefully processed publicly available data set for the evaluation, containing visible-infrared paired sequences, we ensure to find synchronized and same attributes at the same locations, effectively studying only the impact of a spectral change. Thus, by analyzing quantitative results, we identify visual attributes which benefit or detriment from a fused approach on typical visual tracking scenarios.

Keywords: Visual tracking · Infrared imagery · Image fusion visible-infrared

1 Introduction

Tracking is an elemental task for any practical video application, requiring a level of understanding about the objects of interest. The subject has received increasing attention in recent years, where state-of-the-art trackers mainly focus on using the visible spectrum and deep neural networks. Despite the increase in accuracy and robustness, some limitations still persist. In order to overcome the constraints from a single spectral range during tracking, a multi-spectral approach can be utilized. For example, an additional infrared sensor can provide complementary information to an image obtained in the visible range. On the one hand, visible images offer rich content (i.e. colors, texture) and should be preferred when the thermal properties of an object are close to the surrounding environment, on the other hand, infrared images are better suited in case of a change in lighting conditions or gloomy environments [3,7]. This approach could

© Springer Nature Switzerland AG 2019
A. Morales et al. (Eds.): IbPRIA 2019, LNCS 11867, pp. 101–112, 2019.
https://doi.org/10.1007/978-3-030-31332-6_9

highly benefit applications such as, surveillance [1], traffic monitoring [2] and medical imaging [6].

Among some challenges arising in both domains, the potential added value is accompanied by an overhead like increased amount of raw data that needs to be processed. Therefore, an ideal fusion method should preserve the positive characteristics of the individual channels, but reduce the amount of data needed to be processed. Accordingly, all subsequent processing stages, like visual tracking should benefit from the fused input stream. However, due to an inevitable non-optimal fused image, it is not clear if the general expectation of increased performance is met. Towards this end, we evaluate the effect of current state-of-art visual trackers on fused data streams provided by current state-of-the-art fusion strategies.

The evaluation is done on typical visual tracking sequences from the Camel [9] data set. Since image fusion relies on synchronized and well-registered camera, the data set is carefully further edited to ensure best possible fusion results by still capturing different visual attributes as displayed in Fig. 1. Thereby, we can identify approaches and occurring visual attributes which can benefit from a fused input stream. This study differs from the 2015 VOT challenge [15], by only evaluating the appearance changes from one spectrum to another, by using synchronized input streams.

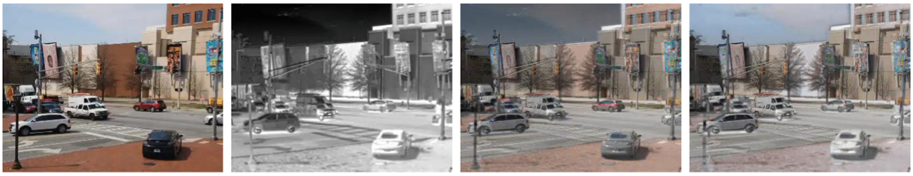

Fig. 1. Synchronized frames of a video stream from Camel [9] in the (from left to right) visual, infrared, addition and l1-norm subsets.

In the following Sect. 2 a short introduction of the selected trackers and fusion strategies are presented. Afterward, we describe the evaluation process and examine the quantitative results in Sect. 3 and Sect. 4 concludes the paper.

2 Visual Trackers and Fusion

2.1 Reference Trackers

Due to the potential wide range of industrial tracking-based applications, visual tracking is a very popular research area and several publicly benchmarks exist. A selection of current single-object visual trackers achieving top-ranks on most-widely used visual tracking benchmarks [14,24] is considered for the experiments presented in this paper. Furthermore, the selected tracker have to rely on an FCNN for feature extraction and operate model free. For a more detailed description and categorization, we refer to the original papers and to the corresponding benchmark papers. Following those criteria, we chose to work with:

Re^3 [10] (Real-Time Recurrent Regression Network) performs feature extraction using the CaffeNet architecture (Re-implementation of AlexNet [16] in Caffe) without the fully connected part. A regression layer is used to output the location of the object in the frame and the size of the object. The tracker uses two LSTM layers for remembering appearance changes and motion information of the target.

MBMD [27] is the winner of 2018 VOT long term challenge [14] and is composed of a bounding box regression network for identifying potential locations of the target and a verification network, identifying the target and the potential locations. Feature extraction for the regression is performed by an SSD-MobileNet [13] network and feature extraction for the verification stage is performed by a pre-trained VGG-16 [22].

DaSiamRPN [28] is the winner of the VOT real-time challenge 2018 [14] and runner up in the long term part of the challenge and is an extension of SiamRPN [18]. Feature extraction is performed using the FCNN of Bertinetto et al. [5]. In addition, it uses a local-global search region strategy for target redetection and a distractor-aware component for catching target appearance variations and to discriminate the target against the surrounding.

MemTrack [26] is based on a dynamic memory network [17] architecture. Feature extraction is performed by the FCNN used in [5]. A soft attention mechanism [25] locates potential regions of the targets in the search area. An LSTM selects an appropriate template from stored ones based on the output of the attention mechanism, combined with a reference template, creating a residual template. Afterwards, the residual template is used for finding the location of the target in the search area.

SiamMask [23] is a recently state-of-the-art introduced tracker which produces in addition to a rotational bounding box a binary mask, classifying pixel-wise belongingness to the target. Feature extraction is performed by a variation of ResNet-50 [12]. Using a depth-wise cross correlation [4] on feature spaces and a region proposal network [21]. By examining the potential targets locations, the actual target is found and a binary mask is created.

2.2 Fusion Strategies

Image fusion strives to preserve positive characteristics of individual channels, in addition to reducing the amount of data needed for processing. Due to inevitable errors induced by image fusion and depending on specific conditions, it is unclear how beneficial this approach is for tracking applications. Before analyzing the effects, we present the main concepts on image fusion methods and the selected strategies for this study.

In multi-modal fusion, different sensors i.e. visible, infrared, are used in the process of image acquisition. The fusion process can be applied on pixel, object or on decision levels. However, in this paper, we examine only pixel level fusion, which can traditionally be divided into transformation domain methods and spatial domain methods [7,11]. These techniques can involve around simple transformations e.g. averaging, adding, subtracting, on pixel intensities, or more complex transformations e.g. Laplace pyramid, wavelet pyramid. Unlike these traditional strategies, a shift to deep learning is occurring, which are now able to achieve state-of-the-art performance. Therefore, for the experiments of this study we select the deep neural network (DNN), DenseFuse [19].

The authors of DenseFuse propose a novel deep learning architecture using convolutional layers and dense blocks. The DNN is composed of three major components. The first component is an encoder made from one convolutional layer, that extracts rough features, followed by three dense convolutional layers, enabling the network to preserve mid and deep level features better. Allowing in addition to improve information flow and diminish the overfitting problem during training. The second component is a fusion layer incorporating two fusion strategies, i.e addition strategy presented originally in DeepFuse [20] and an l1-norm strategy. This layer integrates into one feature map the pertinent features extracted by the encoder from the source images i.e from the visible and infrared images. The third part of the DNN is a decoder which re-constructs the fused visible-infrared image using convolutional layers [19].

Originally, DenseFuse fuses a grayscale image with an infrared image, but it also handles visible images, in splitting the image into separate channels, that are then passed through the DNN and fused separately with the infrared image. The final result is a combination of the three newly created fused images. Figure 1 displays the fused image using both strategies from DenseFuse, and the original visible and infrared images.

3 Evaluation

The goal of this section is to provide empirical results and discuss the benefits and detriments of a fused approach, which can lead to non-optimal output.

3.1 Data Set to Subsets

For this study, we employed the Camel [9] data set which captured 26 annotated video streams paired in the visible and infrared domain. The video streams were taken in an urban environment and captured during day and night time. Similar visual attributes from popular data set for visual object tracking challenges [8,14] are present, i.e. in-plane rotation, illumination change, scale variation, occlusion and camera motion. The data set contains 765 annotated objects in the visible domain and 787 in the infrared domain, where four different classes are present, i.e bicycle, person, vehicle and dog. In order to reduce registration errors, only a reduced subset of sequences are considered for the evaluation. The criteria are set as follows:

- We kept sequences having an Intersection over Union (IoU) over 0.7 between ground-truth bounding boxes in the visible-infrared domain and lasting at least 30 consecutive frames.

- Furthermore, in order to ensure adequate sequence length, a short drop of the IoU under 0.7 is accepted if its only for less than 10 frames and still above 0.5. Whereas an IoU under 0.5 stops the recording until the condition from stage one is valid again.

Based on the newly created domain subset, we created the two fused subsets with the available fusion strategies from DenseFuse (i.e. addition and l1-norm subsets). The ground truth annotation for the fused subsets, is simply adapted by averaging the ground truth bounding boxes from both domains, ensuring us to keep a minimal valid bounding box around the target.

The resulting 4 subsets, visible (VIS), infrared (IR), addition (Add) and l1-norm (l1) subsets used for the evaluation, contain 438 sequences, with a median sequence length of 107 frames, a median target width ratio of 0.1 and height ratio of 0.17. Example images from the four evaluation subsets are displayed in figures of Subsect. 3.5.

Although, we use a state-of-art DNN for image fusion, we can not prevent errors generated by the DNN during the fusion process, i.e. noise, registration difference between visible and infrared images. For example, the images in Fig. 2 depict a non-optimal registration between the source images.

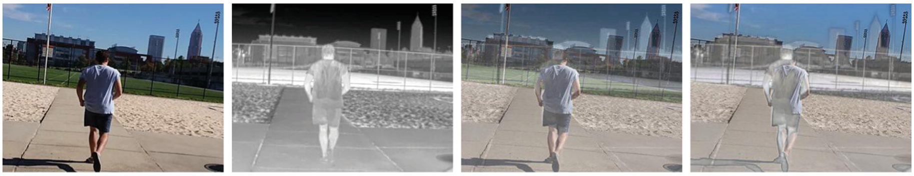

Fig. 2. Example images from Camel [9], showing a registration problem between the synchronized frames in (from left to right) the visual and infrared domain subsets, and the final results in the addition and l1-norm subsets

In contrast to the most closely related investigations on the VOT-IR dataset, here by selecting the Camel data set, we ensure to find synchronized and same attributes at the same locations, effectively studying only the impact of a spectral change on the FCNN of the trackers.

3.2 Evaluation Metrics

Methodologies from one challenge to another differ in the evaluation process as well as the performance measures. In this paper two measures are mainly used to rank the performance: Firstly, accuracy which measures the overlap during successful tracking periods, secondly robustness, which is the number of times the tracker lost the target. We rank the tracker accordingly to their average IoU and average robustness on the whole subsets. For the evaluation, a target is considered lost when the IoU between ground truth and predicted bounding box is under 0.5 for 10 consecutive frames. If the target is lost, we initialize the tracker again on the next frame. For easier comparison between the tracker results, we combine accuracy and robustness together in on score.

3.3 Evaluation on Subsets

Since we use video streams differing domain wise, and use original implementation of the trackers, the change in performance can mainly by assigned to the extracted features. Their performances on the subsets are displayed in Fig. 3 and Table 1 resumes the results in one score.

Based on these results, the Re3, and MBMD trackers show better scores in accuracy and robustness on the visual subset compared to the infrared subset. Both also perform better on the addition subset, gaining in accuracy and robustness, whereas the usage of the l1-norm subset shows a performance drop.

The MemTrack and SiamMask trackers responded interestingly with a better score on the infrared subset than on the visible subset, but regardless of the fusion strategy employed, both achieve better results on the fused subsets. With the MemTrack achieving a slightly better score on the addition subset, and the SiamMask tracker on the l1-norm subset.

In contrast to previous trackers, the DaSiamRPN tracker, does not react as positively as expected. Indeed, the best performance is achieved on the visible subset, even though the addition subset score is close to the visible subset. The worst score for this tracker is achieved when applying the fused L1-norm subset.

Aside from the DaSiamRPN tracker, all trackers benefit from a fusion between the visible and the infrared spectrum at the input stage as shown in Table 2. We also notice that, even though the FCNN part of the tracker are different from each other, some react similarly, as Re3 with MBMD or MemTrack with SiamMask. For instance, the MemTrack and SiamMask respond very well to both fusion strategies, and Re3 displays a similar score variation as MBMD on VIS-Add.

Table 1. Average score (combination between average accuracy and robustness) for each tracker on every subsets.

Subset	DaSiamRPN	Re3	MemTrack	MBMD	SiamMask
VIS	**0.998**	0.976	**0.941**	0.883	**0.995**
IR	0.973	**0.935**	0.971	**0.878**	0.999
Add	0.981	**0.999**	**0.976**	**0.905**	1.015
l1	**0.972**	0.975	0.975	0.883	**1.016**

Table 2. Relative performance variation between a domain subsets (DS) and a fused subsets (FS) depending on the tracker.

DS-FS	DaSiamRPN (%)	Re3 (%)	MemTrack (%)	MBMD (%)	SiamMask (%)
VIS-Add	−1.69	2.40	3.76	2.46	2.00
IR-Add	0.89	6.84	0.61	3.09	1.57
VIS-l1	−2.64	−0.03	3.64	0.02	2.12
IR-l1	−0.08	4.30	0.49	0.64	1.69

3.4 Fused Subsets Versus Domain Subsets

Based on Table 3 we notice that the lowest score distribution occurs on the infrared subset, regardless the tracker. Indicating that the infrared subset is the most difficult one to extract an appearance model from the surrounding, which is coherent since the trackers were originally trained on visible images.

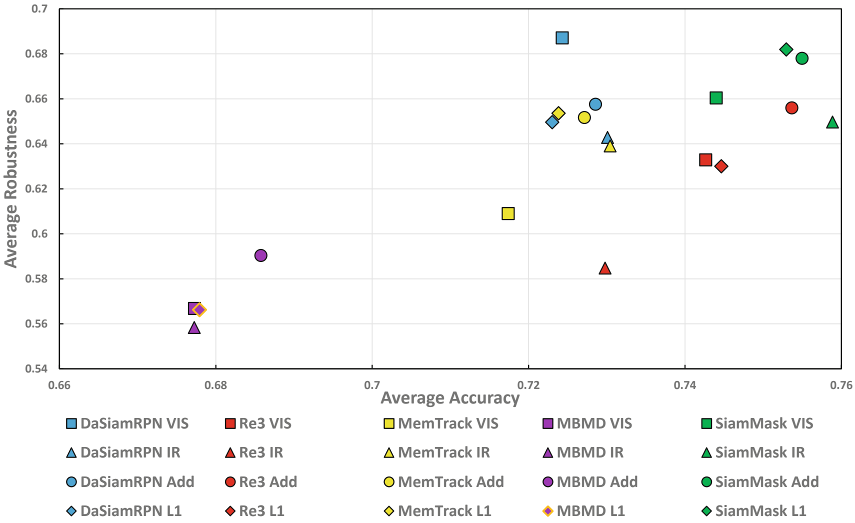

Fig. 3. Average accuracy and robustness given by reference trackers on the subsets. Tracker reaching the top-right corner of the graphic display better performance.

DaSiamPRN responds strongly on 48% video streams of the visible subset and poorly on 22%, with a standard score deviation of 0.229. Whereas, 13% video streams from the additional subset enable a high response from the tracker and only 4% a low response, with a standard score deviation of 0.196. Although, the DaSiamRPN responds stronger on a higher number of video streams from the visible subset, it also responds poorer on a higher number in comparison to the additional subset. Thus, using a fused subset enabled the tracker to track more robustly on a wider sequences diversity, even if using the fused subset did not manage to outperform the score from the visible subset.

Re3 and MBMD score better on 35% and 39% of the video streams from the additional subset, albeit Re3 also responds strongly to the same number of video streams in the visible subset. We note that both respond poorly on a low number of video streams from the addition subset, and achieving a low standard score deviation on the fused subsets. Indicating that using fused input stream enhances their performance on a larger number of streams and allows a more robust approach.

Oddly enough, MemTrack and SiamMask show a strong response on 43% video streams of the infrared subset and a poor response on 39% of it, albeit they originally were trained for visible imagery. Yet, both respond with a low amount of strong and poor scores on the video streams from the fused subsets. The standard score deviation for both trackers is also lower in the fused subsets compared to the domain subsets, even though the MemTrack has the lowest deviation on the infrared subset. Results from SiamMask indicate that using video streams from the fused subset reduce the standard score deviation, thus making the tracker more stable and also improving the results on a variety of sequences. Whereas the MemTrack, having a lower standard score deviation on the infrared domain, performs still better on the fused subsets, suggesting that the fused subset did not necessarily increase stability, but increased overall scores of the tracker on the sequences that gave previously poor scores on the domain subsets.

Table 3. Highest and lowest score distribution of trackers on video streams from the subsets and standard score deviation of the trackers on the subsets

Subset		DaSiamRPN	Re3	MemTrack	MBMD	SiamMask
VIS	Percentage of highest scores	**48%**	**35%**	30%	13%	22%
	Percentage of lowest scores	22%	26%	35%	30%	35%
	Standard score deviation	0.229	0.249	0.207	0.219	0.231
IR	Percentage of highest scores	17%	22%	**43%**	13%	**43%**
	Percentage of lowest scores	**52%**	**57%**	**39%**	**39%**	**39%**
	Standard score deviation	0.204	0.252	0.192	0.223	0.222
Add	Percentage of highest scores	13%	**35%**	17%	**39%**	9%
	Percentage of lowest scores	4%	9%	9%	4%	17%
	Standard score deviation	0.196	0.236	0.194	0.203	0.207
l1	Percentage of highest scores	22%	9%	9%	22%	13%
	Percentage of lowest scores	22%	9%	17%	26%	9%
	Standard score deviation	0.195	0.224	0.197	0.201	0.209

In most cases, the usage of fused subsets proved to be beneficially for stability, as the scores where more balanced and enhanced overall as a whole, rather than individual sequences.

3.5 Special Case Analysis

In this section, we look at video streams that give high score variations from a domain subset to a fused subset, regardless if positive or negative. Albeit the trackers react very differently to the video streams, there are special cases where a general tendency can be observed.

Synchronized frames of the video stream number 3 from Camel [9] are shown in Fig. 4. The video stream has a score increase on the fused subsets compared to the infrared subset, but a lower score compared to the visible subset. In these video stream, the potential targets belong to the same class and are clustered together, thus making the tracking process more challenging in the infrared domain because they all look alike in the infrared spectrum. Whereas, in the visible domain, color and texture are used to discriminate the target against the surrounding, easing the tracking process. But, depending on the fusion strategy, color and texture are removed to some extent and whitened, and useful features from the visible domain are kept to a degree, allowing a performance increase compared to the infrared domain.

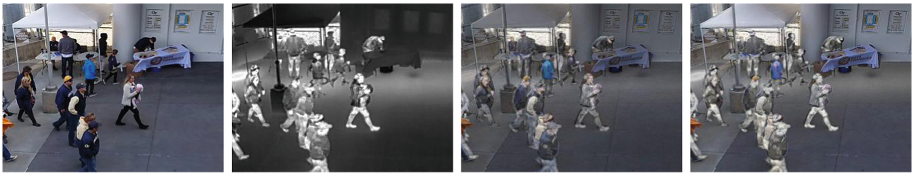

Fig. 4. Synchronized frames of video stream 03 from Camel [9] in the (from left to right) visual, infrared, addition and l1-norm subset, which favours the visible domain due to the presence of color and texture in a crowded scene. (Color figure online)

Matching frames from video stream 9 from Camel [9] are displayed in Fig. 5. Contrariwise to the previous example, these video stream shows a performance gain on the fused subset compared to the visible subset and a drop compared to the infrared subset. Most of the potential targets (i.e. cars) undergo a sudden change in luminosity when passing under the shadows, which is a difficult attribute to handle for visual trackers. Whereas, on the infrared subset, the potential targets are still very clear, and since they are also clearly apart from each other, no clutter attribute is present to increase the difficulty of the tracking process in the infrared domain. By fusing both streams, the whitened target benefits from a constant white color that does not fade away under the shadows, in contrary to colors and textures.

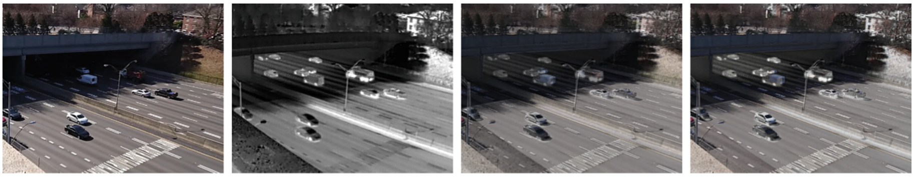

Fig. 5. Synchronized frames of video stream 09 from Camel [9] in the (from left to right) visual, infrared, addition and l1-norm subsets, which favours the infrared domain due to the sudden illumination change. (Color figure online)

Presented in Fig. 6, are four synchronized frames from video steam number 15, which is recorded during night time in Camel [9]. Trackers show better performances on the fused subsets version of this stream than on the domain subsets.

Because of the gloomy environment, tracking in the visible domain is very difficult and naturally tracking in the infrared domain is more suited under these conditions. However, a fused version of both domains, shows a more robust alternative to the visible subset and an increase in accuracy compared to the infrared subset.

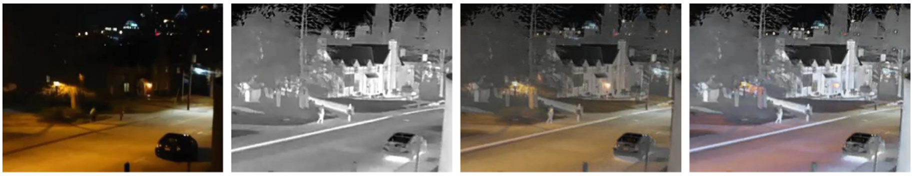

Fig. 6. Synchronized frames of video stream 15 from Camel [9] in the (from left to right) visual, infrared, addition and l1-norm subsets, favouring the fused subsets.

Synchronized frames of the video stream number 7 from Camel [9] are showed in Fig. 7. Trackers show better results on the domain subsets than on the fused versions. Due to the environment, a fused approach makes tracking more difficult since the whitening of the targets blends better with the white background and the snowy weather. Thus, tracking in the visible domain under these conditions is easier, since the tracking process can rely on color and texture features. Also, when using the infrared domain, the target is even clearer to discriminates against the surrounding background as shown in Fig. 7.

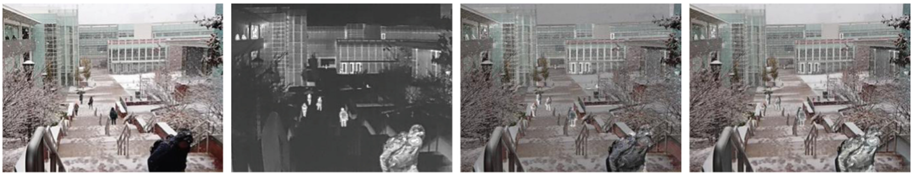

Fig. 7. Synchronized frames of video stream 07 from Camel [9] in the (from left to right) visual, infrared, addition and l1-norm subsets, where neither of the fused subset outperforms a domain subset. (Color figure online)

4 Conclusion

We present an evaluation of state-of-the-art visual trackers applied on visible, infrared, and fused input streams. In contrary to using one domain for tracking, where specific attributes to the domain can be difficult i.e. illumination change in the visible domain or clutter in the infrared domain to deal with, a fused approach at the input stage can be effective at handling those attributes. Indeed using early fused input streams indicate a tendency to enhance performance, enabling more robust and accurate tracking. In addition, allowing also to handle more robustly diverse type of sequences under various conditions and attributes. Depending on the fusion strategy, performances can improve or diminish. However an ideal fusion strategy would enable the tracker to perform on the fused subset as good as it would perform on an adequate domain subset version.

References

1. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
2. Battiato, S., Farinella, G.M., Furnari, A., Puglisi, G., Snijders, A., Spiekstra, J.: An integrated system for vehicle tracking and classification. Expert Syst. Appl. **42**(21), 7263–7275 (2015)
3. Becker, S., Scherer-Negenborn, N., Thakkar, P., Hubner, W., Arens, M.: An evaluation of background subtraction algorithms on fused infrared-visible video streams. In: 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–6, November 2015. https://doi.org/10.1109/DICTA.2015.7371229
4. Bertinetto, L., Henriques, J.A.F., Valmadre, J., Torr, P.H.S., Vedaldi, A.: Learning feed-forward one-shot learners. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, pp. 523–531, Curran Associates Inc., USA (2016). http://dl.acm.org/citation.cfm?id=3157096.3157155
5. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 850–865. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_56
6. Blake, A., Isard, M.: Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion. Springer, London (2012). https://doi.org/10.1007/978-1-4471-1555-7
7. Blum, R.S., Zheng, L.: Multi-Sensor Image Fusion and Its Applications. Signal Processing and Communications. Taylor & Francis, Boca Raton (2005). http://opac.inria.fr/record=b1105877
8. Fan, H., et al.: LaSOT: a high-quality benchmark for large-scale single object tracking. arXiv preprint arXiv:1809.07845 (2018)
9. Gebhardt, E., Wolf, M.: Camel dataset for visual and thermal infrared multiple object detection and tracking. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2018)
10. Gordon, D., Farhadi, A., Fox, D.: Re3: Real-time recurrent regression networks for visual tracking of generic objects. IEEE Robot. Autom. Lett. **3**(2), 788–795 (2018)
11. Goshtasby, A.A., Nikolov, S.G.: Image fusion: advances in the state of the art. Inf. Fusion **8**, 114–118 (2007)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. CoRR arxiv:abs/1704.04861 (2017)
14. Kristan, M., et al.: The sixth visual object tracking VOT2018 challenge results. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 3–53. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_1

15. Kristan, M., et al.: The visual object tracking vot2015 challenge results. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1–23 (2015)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates, Inc. (2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
17. Kumar, A., et al.: Ask me anything: dynamic memory networks for natural language processing. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 48, pp. 1378–1387, New York, USA, 20–22 Jun 2016. http://proceedings.mlr.press/v48/kumar16.html
18. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with Siamese region proposal network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Li, H., Wu, X.: DenseFuse: a fusion approach to infrared and visible images. IEEE Trans. Image Process. **28**(5), 2614–2623 (2019). https://doi.org/10.1109/TIP.2018.2887342
20. Prabhakar, K.R., Srikar, V.S., Babu, R.V.: DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4724–4732, October 2017. https://doi.org/10.1109/ICCV.2017.505
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR arxiv:abs/1409.1556 (2014)
23. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: a unifying approach. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
24. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
25. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
26. Yang, T., Chan, A.B.: Learning dynamic memory networks for object tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 153–169. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_10
27. Zhang, Y., Wang, D., Wang, L., Qi, J., Lu, H.: Learning regression and verification networks for long-term visual tracking. arXiv preprint arXiv:1809.04320 (2018)
28. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware Siamese networks for visual object tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 103–119. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_7