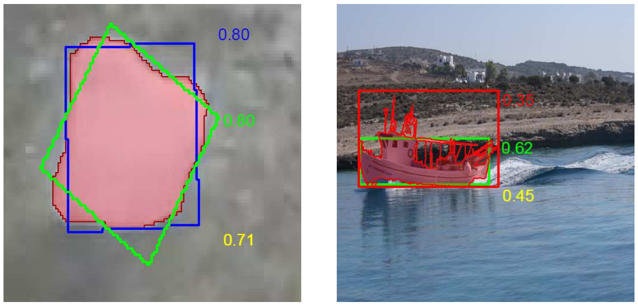
The Benefits of Evaluating Tracker Performance using Pixel-wise Segmentations

Tobias Böttger Patrick Follmann
MVTec Software GmbH
{boettger,follmann}@mvtec.com

Abstract

For years, the ground truth data for evaluating object trackers consists of axis-aligned or oriented boxes. This greatly reduces the workload of labeling the datasets in the common benchmarks. Nevertheless, boxes are a very coarse approximation of an object and the approximation by a box has a large degree of ambiguity. Furthermore, tracking approaches that are not restricted to boxes cannot be evaluated within the benchmarks without adding a penalty to them. We present a simple extension to the VOT evaluation procedure that enables to include these approaches. Furthermore, we present upper bounds for trackers restricted to boxes. Moreover, we present a new measure that captures how well an approach can cope with scale changes without the need of frame-wise labels. We present a learning-based approach which helps to identify frames with heavy occlusion automatically. The framework is tested on the segmentations of the VOT2016 dataset.



(a) `bag` from VOT2016 [11] (b) `boat` from DAVIS [18]

Figure 1. In image (a), both oriented boxes have an identical IoU with the ground truth segmentation. Nevertheless, their common IoU is only 0.71. Restricting the ground truth to boxes may introduce an undesired bias in the evaluation. Furthermore, although the object detection (green) in image (b) has an overlap of 0.62 with the ground truth segmentation, its IoU with the ground truth axis-aligned bounding box is only 0.45 and would be considered a false detection in the standard procedure.

1. Introduction

Visual object tracking is a rapidly evolving research area with dozens of new algorithms being published each year. To compare the performance of the many different approaches, a vast amount of evaluation datasets and schemes are available. The most common are OTB [27] and VOT [11]. Both benchmarks use axis-aligned or oriented boxes as ground truth and estimate the accuracy with the Intersection over Union (IoU) criterion [23, 25].

Unfortunately, using boxes as ground truth has two inevitable disadvantages:

1. The approximation of an object by a box is very crude. Especially articulated objects such as humans or animals can not be approximated well by boxes and the choice of the best suitable box is highly ambiguous. For example, the bag in Fig. 1 (a) has multiple valid box approximations with the same overlap. Nevertheless, the Intersection over Union (IoU) between two of the valid choices is only 0.71. To counter the ambiguity of the boxes to some extent, the objects in the VOT dataset are labeled by multiple annotators to get an approximation of the box ambiguity.

2. It is difficult to evaluate approaches that are not restricted to oriented or axis-aligned boxes on ground truth boxes without introducing an unwanted bias in the evaluation results. For example, the ground truth segmentation in Fig. 1 (b) only has an IoU of 0.35 with the red ground truth bounding box but is a perfect approximation of the object itself.

Especially the later point is of increasing concern. The recent advances of Fully Convolutional Networks (FCNs) for semantic segmentation [14, 21] have inspired approaches capable of tracking dense segmentations through image sequences in real-time. In One-Shot Video Object Segmentation (OSVOS), Caelles *et al*. [7] approximate the segmentations by bounding boxes to enable a comparison with the state-of-the-art bounding box tracker MDNET [16]. Furthermore, it is also difficult to evaluate the performance gains of approaches that approximate the object by general affine transformations (and not only through rotated boxes) with the current VOT and OTB ground truth data [6].

To address these problems, a number of densely segmented ground truth datasets has started to emerge [13, 18, 26]. Nevertheless, no evaluation protocol that enables a fair comparison of tracking approaches restricted to boxes and those which are not exists. We propose to use the IoU of the object segmentation and the tracker proposal as an accuracy measure. By furthermore calculating the optimal axis-aligned and oriented box for each segmentation, it is possible to obtain reliable upper bounds for all approaches restricted to boxes. The optimal boxes can be efficiently computed with the approach proposed in [5]. The main contributions of this work include:

- We introduce reliable upper bounds for trackers restricted to axis-aligned and oriented boxes for all VOT2016 sequences.
- We present a measure which captures how well an approach can capture scale change without the need of frame-wise labels. The measure is not correlated to the current accuracy and robustness measures and tested on a collection of the trackers submitted to VOT2016.
- We show how heavy occlusion occurring in the VOT dataset can be detected automatically without the need of manual annotation.

By proposing to use the segmentations to calculate the accuracy directly, we remove the restriction of the evaluation scheme to approaches restricted to boxes. The code to generate the theoretical trackers and the upper bounds for VOT2016 will be made available to the community[1].

The rest of the paper is organized as follows. In section 2, we examine the existing literature and motivate why boxes are not sufficient to measure tracker accuracies. In section 3, we present the upper bounds for VOT, the new scale measure and the learning-based scheme for detecting occlusions and our evaluations. Section 4 concludes the paper and gives an outlook on our further work.

2. Related Work

Very recently, a number of approaches have emerged where the current bounding box groundtruths were not sufficient [4, 6, 7]. Caelles *et al.* [7] introduce the above mentioned OSVOS, which essentially tracks a segmentation through an image sequence. The base of their CNN is pretrained on ImageNet for image labeling. The network is then trained on the binary masks of DAVIS [18], to learn a generic notion of how to segment objects from their background. When applied, the network is first fine-tuned on the first frame of the sequence (annotation mask and image). Then the network is tested on the successive frames. The approach allows to weight between accuracy and runtime and runs at max 10 fps on 480 × 854 sized images. To compare their approach to a current state-of-the-art tracker, they computed the bounding box of their segmentation and compared it to the ground-truth. Even though they were able to outperform the state-of-the-art, the restriction of their results to a bounding box introduced a negative bias. Similarly, Böttger *et al.* [6] proposed a sub-pixel precise tracking scheme which was evaluated on a handful of rigid objects from the OTB and VOT2016 dataset. Although the visual results indicate an increase of the accuracy, the bounding box ground truths were insufficient to show a performance gain on a quantative scale.

	Φ_{IoU}
VOT2015	0.577
VOT2016	0.651
`box-no-scale`	0.512
`box-axis-aligned`	0.722
`box-rot`	0.760

Table 1. The IoU of the theoretical tracker `box-axis-aligned`, `box-rot`, `box-no-scale` and the VOT2016 and VOT2015 ground-truths [11] with the VOT2016 segmentations [26].

In general, the accuracy of tracking approaches is computed as the Intersection over Union (IoU) of the box ground truths and the tracker predictions. Many other measures to compute the accuracy of trackers have been proposed [11, 12, 17, 22, 25, 27]. To unify the evaluation of trackers, Čehovin *et al.* [23, 25] provide a highly detailed theoretical and experimental analysis of the most popular performance measures and show that many of the above measures are highly correlated. The appealing property of the IoU measure is that it accounts for both position and size of the prediction and ground truth simultaneously. Hence, all of the common tracking benchmarks [11, 15, 27] use the Intersection over Union (IoU).

When moving to densely segmented ground truths, the IoU of the dense segmentations and the tracker predictions will generally not result in a measure between 0 and 1. To counter this, very recently, three new theoretical trackers that deliver upper bounds for approaches restricted to axis-aligned or oriented bounding boxes on dense segmentations have been proposed [5]. The authors use the boxes with the best possible IoU for a segmentation to normalize the IoU. The optimal boxes are essentially upper bounds for evaluating approaches restricted to boxes on densely segmented ground truth.

In our paper, we make use of the three theoretical tracker and show how they can be used to compute reliable upper bounds for the VOT2016 evaluation scheme. Furthermore,

[1] http://www.mvtec.com/company/research

they enable the calculation of a new measure which captures how well an approach can capture scale change. The theoretical trackers also enable the detection of frames where occlusions occur within the dataset and can be used as a valuable indicator for degenerated ground truths.

3. Evaluation Scheme

We present an extension of the 2016 VOT evaluation scheme [12]. In a first step, we propose to use dense segmentations as ground truth. A dataset for VOT2016 has recently been proposed [26]. This removes the ambiguity of fitting boxes to non-rectangular objects. Furthermore, it removes the need to label objects by different annotators to get an estimate of the variation of the respective ground truth box.

Nevertheless, since a great majority ($\approx$ 95%) of the current approaches is restricted to axis-aligned bounding boxes, we propose to incorporate the three theoretical trackers from [5] into the evaluation process. The trackers determine upper bounds for the accuracy of tracking approaches restricted to axis-aligned, oriented and scale-variant bounding boxes, respectively.

We restrict our complete evaluation to a handful of VOT2016 tracker which are openly available. We did not restrict the selection to the top-performing trackers, but tried to select a diverse set of trackers both in terms of their ranking and their feature selection.

3.1. Upper bounds for VOT

The concept of theoretical trackers was first introduced by Čehovin *et al*. [25] as an "*excellent interpretation guide in the graphical representation of results*". In their paper, they use perfectly robust or accurate theoretical trackers to create bounds for the comparison of the performance of different trackers. We propose to use the three theoretical tracker proposed in [5] to obtain upper bounds for the accuracy of trackers that underlie the box-world assumption.

Using the IoU of theoretical trackers Φ_{opt}, the relative Intersection over Union (rIoU) of a box $\mathcal{B}$ with a dense segmentation $\mathcal{S}$ is computed as,

$$\Phi_{rIoU}\left(\mathcal{S},\mathcal{B}\right) = \frac{\Phi_{IoU}(\mathcal{S},\mathcal{B})}{\Phi_{opt}(\mathcal{S})}, \qquad (1)$$

where Φ_{IoU} is the Intersection over Union (IoU),

$$\Phi_{IoU}\left(\mathcal{S},\mathcal{B}\right) = \frac{|\mathcal{S}\cap\mathcal{B}|}{|\mathcal{S}\cup\mathcal{B}|}. \qquad (2)$$

Here, Φ_{opt} is the best possible IoU a box can achieve for the segmentation $\mathcal{S}$. In comparison to the usual IoU (Φ_{IoU}), the rIoU measure (Φ_{rIoU}) truly ranges from 0 to 1 for all possible segmentations. The computation of the box that achieve an IoU of Φ_{opt} is explained in more detail in [5].

The rIoU measure is useful to estimate how well box-based schemes can do on a given sequence. It should not be used to compare a segmentation-based scheme to a box-based scheme. For comparing segmentation-based schemes to a box-based schemes the IoU of the tracker prediction and the ground truth segmentation itself is already a valid measure. Objects are generally not boxes and hence any approximation a tracking scheme makes is an error which should be visible in the accuracy measure. Nevertheless, the rIoU is a useful indicator of how much room for improvement any box-based scheme has for the given segmentations. When making the transition from bounding box to segmentation-based ground truths this is a useful tool to bring the currently existing approaches into perspective.

By computing $\Phi_{opt}(\mathcal{S})$ for a complete sequence and different parameterizations of $\mathcal{B}$, three theoretical trackers can be obtained. Given the segmentation $\mathcal{S}$, the first tracker returns the best possible axis-aligned box (`box-axis-aligned`), the second tracker returns the optimal oriented box (`box-rot`) and the third tracker returns the optimal axis-aligned box with a fixed scale (`box-no-scale`). The scale is initialized in the first frame with the scale of the box determined by `box-axis-aligned`. While the first two tracker cannot be efficiently optimized globally, the `box-no-scale` can efficiently be computed by exhaustively computing the IoU at all possible positions.

The theoretical trackers are essentially upper bounds for the IoU all trackers restricted to boxes can obtain on densely segmented ground truth. All trackers restricted to boxes in general can not obtain an IoU higher than that of `box-rot` for all possible segmentations. Those trackers which are further restricted to axis-aligned boxes can not exceed `box-axis-aligned`. Finally, trackers which cannot capture scale changes, such as the KCF [10] tracker, are also exceeded by `box-no-scale`.

The upper bounds are valuable indicators of how much potential a given tracking approach still has. In Table 1, the IoU of the VOT box ground truths and the theoretical tracker are listed to bring them into perspective. As displayed, no box-based tracker can obtain an IoU of over 0.76 for the complete dataset. The relative IoU measure is displayed for the complete dataset for the VOT2016 ground truth boxes and a handful of the trackers submitted to the VOT challenge in 2016 in Table 2. It becomes apparent that there is still sufficient room for improvement for the existing trackers.

3.2. Measuring Scale-Changes

To capture how well an approach can cope with scale changes, we propose a new measure. The measure builds on the fact that scale changes within a sequence result in a significant drop in the IoU of the `box-no-scale` tracker

	Φ_{IoU}	Φ_{rIoU}
Φ_{opt}	0.72	1.0
VOT2016	0.65	0.89
CCOT [9]	**0.41**	**0.56**
ANT [24]	0.26	0.37
DSST [8]	0.24	0.32
L1APG [2]	0.18	0.25
STAPLE [3]	0.33	0.46
DFST [19]	0.27	0.37
DPCF [1]	0.29	0.41

Table 2. Comparison of different tracking approaches and their average absolute (Φ_{IoU}) and relative IoU (Φ_{rIoU}) for the VOT2016 [11] segmentations. Φ_{opt} denotes the theoretical tracker `box-axis-aligned` from [5].

curve, while the `box-axis-aligned` tracker remains unaffected, as can be seen nicely in Fig. 2. The derivative of the difference of `racing` curve (γ) is visualized in Fig. 3 and is a reliable indicator of the scale-change and the foundation of the presented scale measure. Whenever it exceeds a threshold, it is assumed that the scale is changing. We used a threshold of 0.5×10^{-3} to suppress minor scale changes and to compensate for noise in the segmentations. Moreover, we smooth all three derivatives with a Gaussian function with $\sigma = 3$.

For the frames that are identified as changing scale, we calculate the scale score s. For each of these frames, we compare the change of the size of the tracker predictions to that of the `box-axis-aligned` tracker. If both have the same direction, we assume the tracker is successfully registering a scale change. To make the approach as independent from the accuracy measure as possible, we do not regard the magnitude of the size changes, but merely their sign. Please note, the change of the size of the ground truths boxes or segmentations to estimate the "*ground truth*" scale change could equally be used. Nevertheless, we chose to use the `box-axis-aligned` tracker to obtain an estimate of the scale score for two reasons. First of all, the segmentations themselves are very noisy and secondly, by using the `box-axis-aligned` scale change, it is possible to bring the tracker scale scores into relation to the scale score of the VOT2016 ground truth boxes.

The scale score s for a sequence is computed as

$$s = \frac{1}{n}\sum_{i}^{n} \delta_{\mathrm{sgn}(\mathrm{size}(\mathcal{T}_i)'),\mathrm{sgn}(\mathrm{size}(\mathcal{G}_i)')}, \tag{3}$$

where $\mathcal{T}_i$ is the tracker prediction at frame i, $\mathcal{G}_i$ is the ground truth box, segmentation or one of the scale adaptive theoretical tracker (`box-axis-aligned` and `box-rot`) at frame i, sgn is the signum function and $\delta_{i,j}$ is the Kronecker delta, which is 1 if the variables are equal, and 0 otherwise:

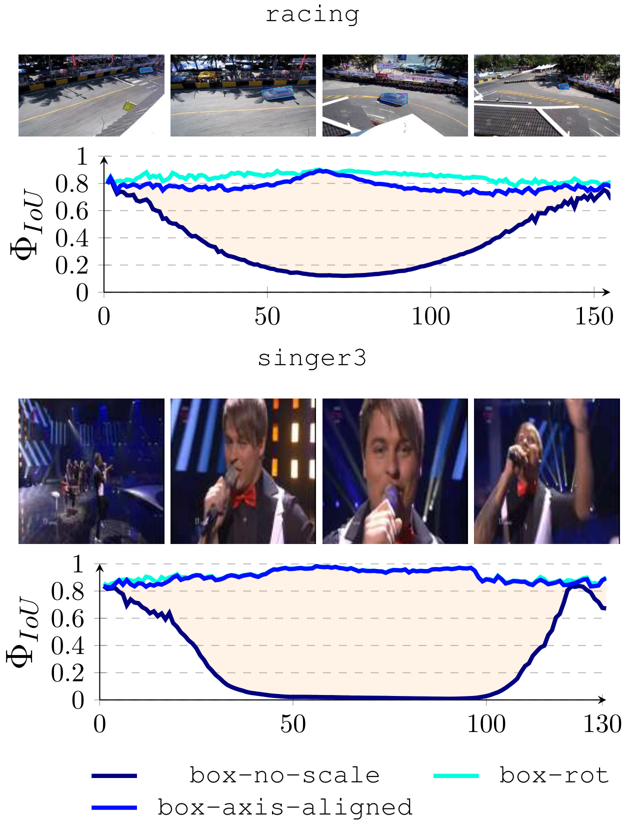


Figure 2. `racing` and `singer3` from VOT2016 [11]. The increasing gap between the `box-no-scale` and the other two theoretical trackers indicates a scale change. The best possible IoU is never above 0.80 for the top sequence and never above 0.90 for the bottom sequence.

$$\delta_{i,j} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{4}$$

The size of the region $\mathcal{R}$ is denoted as $\mathrm{size}(\mathcal{R})$ and its derivative as $\mathrm{size}(\mathcal{R})'$. The derivative is approximated by central differences.

Trackers that do not estimate the scale have a scale score of 0 ($\mathrm{size}(\mathcal{T}_i)' = 0 \ \forall \in n$) and a perfect scale adaptive tracker has a score of 1. Please note, no frame-wise labels are required and the scale score is, by construction, uncorrelated to the accuracy or robustness overlap. We computed the scale score without reinitialization on tracker failure and ignored the frames where the tracker had completely failed (hence $\Phi_{IoU} = 0$).

We evaluated the new measure for all sequences in the VOT2016 challenge for a handful of top trackers and for the ground truth themselves. The results are displayed in

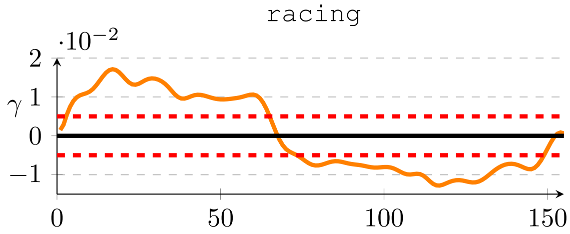


Figure 3. The derivative of the difference between the `box-no-scale` and `box-axis-aligned` tracker from Fig. 2. The magnitude of the curve is a reliable indicator for scale change. To suppress minor scale changes and to compensate for noise in the segmentations, we require the magnitude to exceeds the fixed threshold of 0.5×10^{-3} (visualized as the red-dotted lines).

	s
`box-axis-aligned`	1.0
`box-no-scale`	0.0
VOT2016	0.81
CCOT [9]	**0.54**
DSST [8]	0.36
ANT [24]	0.30
L1APG [2]	0.31
STAPLE [3]	0.37
DFST [19]	0.32
DPCF [1]	0.36

Table 3. The scale score (3) of the theoretical tracker `box-axis-aligned`, the VOT2016 ground-truths [11] and a collection of trackers from the literature. Even the top performing tracker (CCOT) has a relatively low score (e.g. in comparison to the VOT2016 ground truths).

Table 3. Since all the trackers in the VOT2016 challenge were restricted to boxes, we use the size change of the `box-axis-aligned` tracker to compute $\text{sgn}(\text{size}(\mathcal{G}_i)')$. We retained from using the segmentations of VOT2016 directly, since they are quite noisy. Nevertheless, the extension of the scale measure to segmentation-based trackers is straight-forward.

In general, the scale adaptation appears to be a problem of current state-of-the-art approaches. The scale score for all trackers is significantly lower than that of the VOT2016 ground truths. There are many examples where the `box-no-scale` tracker is able to outperform all of the tested axis-aligned tracker in terms of the IoU. A few striking examples are displayed in Fig. 4. The scale score s of the respective sequences is well below 0.3 for all trackers.

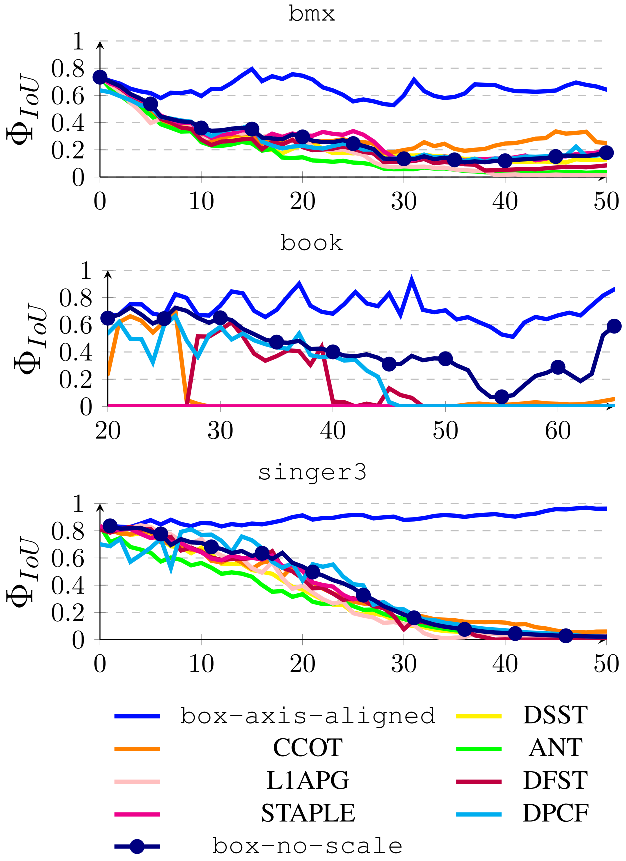


Figure 4. `bmx` (top), `book` (middle) and `singer3` (bottom) from VOT2016 [11]. None of the tested trackers can cope with the scale change. This becomes apparent since none of them can seriously outperform the `box-no-scale` tracker and all fail early on in the sequences.

3.3. Detecting Label Errors

The manual generation of densely segmented ground truths is a tedious task. Hence, many datasets use a semi-supervised approach such as GrabCut [20] or OSVOS [7] to generate the labels. Nevertheless, it is inevitable that label errors may be generated by semi-supervised approaches. To help identify obvious errors automatically, we propose to use the derivative of the difference of the `box-no-scale` tracker and the `box-axis-aligned` tracker (see Fig. 3). When ever the magnitude exceeds a certain threshold the object is either undergoing a very rapid scale change, or the segmentation is degenerating. Since most tracking sequences have a reasonable frame-rate and object deformation, degenerated labels tend to have a much more extreme derivative than natural deformation. We applied the approach to the complete VOT2016 segmentations and where able to identify all of the extremely degenerated frames. A collection of the detected degenerations and the deriva-

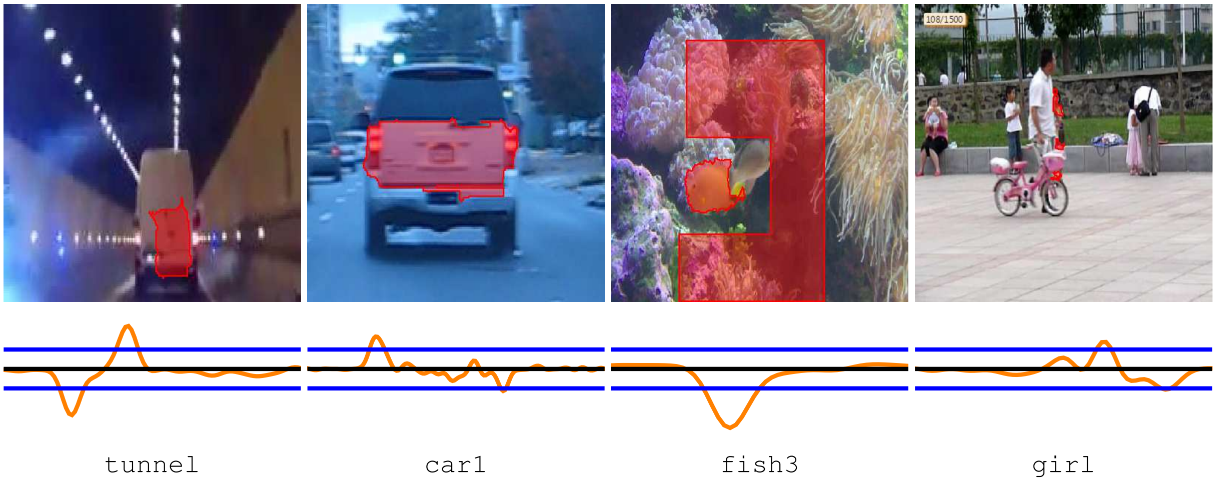


Figure 5. Detecting label errors on VOT2016 [11] and the respective derivative of the difference between the `box-no-scale` tracker and the `box-axis-aligned` tracker is displayed. In general, a large derivative is a good indicator of a degenerated label. There are a few false positives though. For example, in the right image, the girl is merely being occluded and the segmentations are good. Nevertheless, the derivatives are a useful tool for a semi-supervised labeling.

tives of the respective difference of the `box-no-scale` and `box-axis-aligned` IoUs are displayed in Fig. 5.

We also tried to use the derivative of the size of the segmentations itself as a measure to detect degenerations. Nevertheless, since the size also incorporates all size changes due to natural object deformation (such as a walking pedestrian) it is much more noisy than the box approximations of `box-no-scale` and `box-axis-aligned`. Please note, the small resolution of the VOT dataset makes the pixel-wise ground truth annotation a very difficult task in general.

3.4. Detecting Occlusions

In many cases, trackers fail in moments when objects are occluded. One of the reasons is that the underlying models cannot be updated to handle the abrupt change of the object appearance without loosing robustness in general situations. Furthermore, the majority of trackers only handle short-term tracking and thus do not re-detect the object once it is fully lost. Of course, this behavior influences the IoU or rIoU measures on a whole sequence a lot. It is therefore helpful to be aware of occlusions occurring in a sequence to evaluate tracker performance and failure with more detail.

Here, we present a learning-based approach to detect frames with occlusions using only the dense segmentations and the theoretical trackers mentioned above. As is indicated in Fig. 6, frames with occlusions are often characterized by significant bumps in both the area of the segmentation as well as the IoUs of the theoretical trackers.

To learn the occlusions, we use a fully-convolutional network (FCN) [21] with a simple structure, as depicted in Fig. 6. As input, we took the derivative of the smoothed functions of the area of the segmentation and the area of `box-rot`[2]. We also include the IoU values of `box-axis-aligned` and `box-rot`, normalized to have zero-mean and standard deviation one. Thus, we use four time-series as input to the network. We padded all four smoothed functions constantly by ten frames to the left of the first frame and the right of the last frame. The first layer has kernels of shape $4 \times 1 \times 21$ (depth $\times$ height $\times$ width). It is followed by two 1×1-convolutions that replace the usual inner-product layers for classification. Between conv1 and

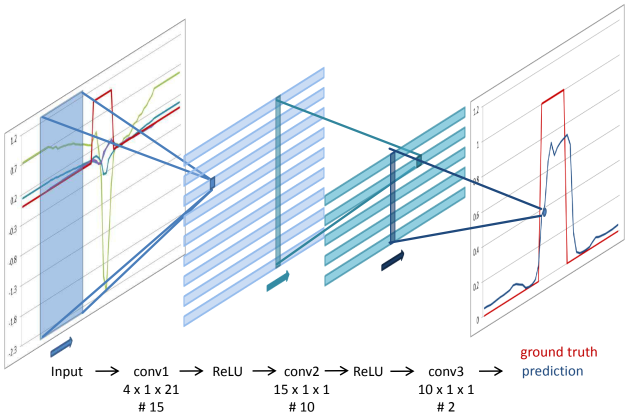


Figure 6. Scheme of our FCN model for occlusion detection: The input are 4 time series consisting of area derivatives and IoU values of segmentations and theoretical tracker outputs (see text for details). Ground truth is overlayed in red, where values of 1 indicate occlusion. The input values and prediction shown here were calculated on sequence `motorbike` of DAVIS [18] (best viewed in color and digitally with zoom).

[2] All area values were divided by the maximum area of the sequence before 0.5 was subtracted.

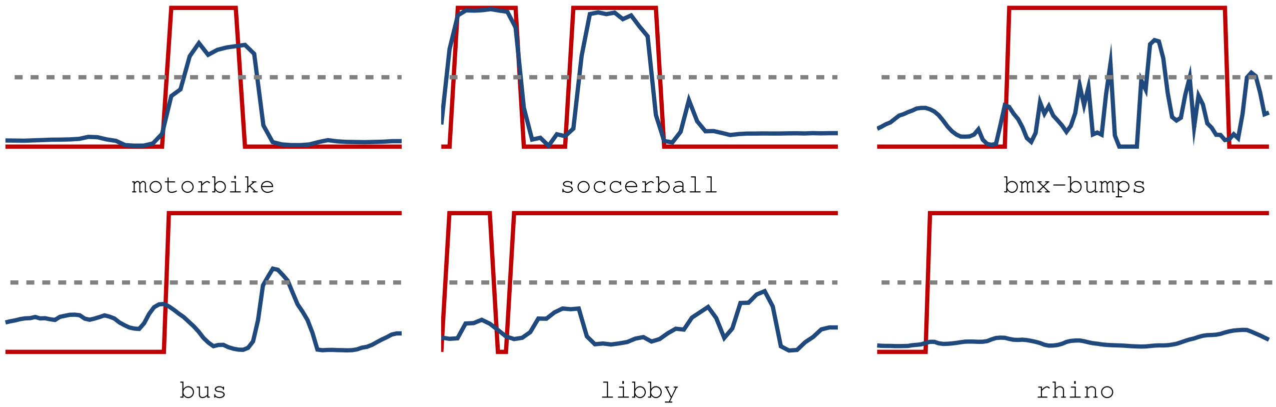


Figure 7. Occlusion prediction on DAVIS [18]. The red line represents the ground truth for occlusion (1 indicates occlusion is present), the blue line is the prediction of our CNN, where the softmax output was thresholded at 0.5 as indicated by the dashed horizontal line. All shown sequences are from validation or test set and were not seen during training. In cases, where the occlusions are severe, the results are good (e.g. motorbike, soccerball). Occlusions with long duration (bmx-bumps, bus), can be detected succesfully only to some extent. For sequences where the occlusion is present during almost the whole scene or the occlusion is only minor, the results are less promising (libby, rhino).

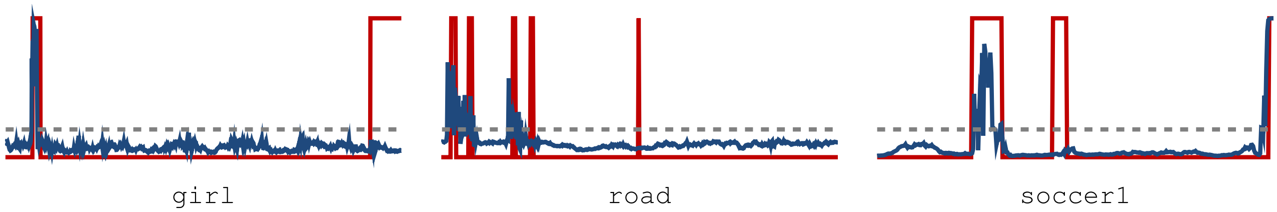


Figure 8. Occlusion prediction on VOT2016 [11]. The red line represents the ground truth for occlusion (1 indicates occlusion is present), the blue line is the prediction of our CNN, where the softmax output was thresholded at 0.2 as indicated by the dashed horizontal line. All shown sequences are from validation or test set and were not seen during training. Heavy occlusions (e.g. girl, road) can be detected in most cases, partial occlusions as those in soccer1 are more difficult.

conv2, as well as conv2 and conv3, we put a ReLU nonlinearity.

We divided the sequences of DAVIS [18] randomly into training, validation and test splits, assuring that each of the sets has approximately the same number of sequences with and without occlusion. We use the DAVIS data since the ground truths are labeled very accurately and the noise within the data is significantly lower than for the VOT segmentations. All sequences were manually labeled. The mean occlusion-rate per sequence is 13.3%, 25.2% and 23.8% for training, validation and test set, respectively. We trained our model for 150 epochs on the 16 sequences of the training set and evaluated after each epoch on the validation set. In each iteration, the data of a whole sequence was put into the FCN simultaneously. Thus, 16 iterations are one epoch. We used stochastic gradient descent with a constant learning rate of 0.01, momentum of 0.9 and weight-decay of 0.001. To avoid overfitting, we stopped early after 99 epochs to obtain a mean accuracy per sequence of 78.4% and 76.4% on the validation and test set, respectively.

Although these numbers appear to be low, the model returns a lot more valuable results than a model that always returns “no occlusion” as result (and would still have a high test accuracy). Some promising sequences and some failure cases out of the validation and test set are visualized in Fig. 7. Generally, severe occlusions can be detected quite well. Nevertheless, the model has its difficulties for sequences with a low number of frames, where the occlusion is only minor, or where the occlusion is present during almost all frames. In sequences with long occlusions, such as bmx-bumps or bus, the occlusion is only predicted partly. For the annotator, this still has the benefit that he gets a hint where to look.

When training the model longer, there was severe overfitting to the training sequences, leading to a high number of false positives in the validation and test sets. This leads to the assumption that more training data is necessary in order to get a model that generalizes better. Nevertheless, the scheme is a first step towards automatically obtaining frame-wise occlusion information for sequences without manually labeling them. Please note, the training data currently only consists of 16 sequences for DAVIS.

We also trained and evaluated our model on the VOT sequences. The hyperparameters of the model and solver settings were similar. The mean sequence occlusion-rate was 7.8%, 11.0% and 4.8% on the training, validation and test set, respectively. Each of the sets consisted of 20 sequences, eight of them containing at least one occlusion. The extension of the approach to the VOT data is difficult. This is mainly due to the high level of noise within the VOT segmentations. These are inevitable since the objects within the VOT dataset are partly very small and the images themselves tend to have a very low resolution. Hence, without smoothing the data, the IoU curves are very noisy, as well.

Early stopping after 71 epochs, we could obtain a mean accuracy per sequence of 89.1% on the validation sets. Reducing the confidence threshold for an occlusion from 0.5 to 0.2 slightly increased the mean accuracy to 89.2% and lead to 94.9% mean accuracy per sequence on the test set. This model rarely predicts occlusion which leads to a relatively high number of false negatives. In those cases it does predict occlusions, it is often right. False positives mainly occur at positions with high object deformation or at positions of label errors. A few examples are presented in Fig. 8.

4. Conclusion

In this paper, we proposed an extension of the 2016 VOT evaluation protocol [12]. We used the segmentations of the scenes to evaluate the accuracy directly. By incorporating three new theoretical tracker into the evaluation framework, it was possible to obtain reliable upper bounds for the performance of all trackers that are restricted to a box representation of the object. We presented a new measure that ranges from 0 to 1 and evaluates how well a tracker can adapt to scale changes. The measure does not require any framewise labels. Additionally, the derivatives required to compute the scale measure can be a good indicator for ground truth label errors. Furthermore, we presented a learning-based approach to automatically detect occlusions in tracking sequences. The method can help to reduce the workload of manually labeling the occlusions in new tracking sequences. In future work, we want to use the increasing amount of available high-quality segmentations in order to improve our presented occlusion detection model. Especially, the problem of automatically detecting partial occlusions or frames where the object is partially outside the image.

References

[1] O. Akin, E. Erdem, A. Erdem, and K. Mikolajczyk. Deformable part-based tracking by coupled global and local correlation filters. *Journal of Visual Communication and Image Representation*, 38:763–774, 2016. 4, 5

[2] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust L1 tracker using accelerated proximal gradient approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1830–1837, 2012. 4, 5

[3] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr. Staple: Complementary learners for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1401–1409, 2016. 4, 5

[4] T. Böttger and C. Eisenhofer. Efficiently tracking extremal regions in multichannel images. In *International Conference on Pattern Recognition Systems (ICPRS)*, 2017. 2

[5] T. Böttger, P. Follmann, and M. Fauser. Measuring the accuracy of object detectors and trackers. In *German Conference on Pattern Recognition*, 2017. 2, 3, 4

[6] T. Böttger, M. Ulrich, and C. Steger. *Subpixel-Precise Tracking of Rigid Objects in Real-Time*, pages 54–65. Image Analysis: 20th Scandinavian Conference, SCIA, 2017. 2

[7] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5

[8] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*, 2014. 4, 5

[9] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Computer Vision - European Conference on Computer Vision*, pages 472–488, 2016. 4, 5

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015. 3

[11] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, L. Čehovin, T. Vojír, and G. Häger. The visual object tracking VOT2016 challenge results. In *Computer Vision - European Conference on Computer Vision Workshops*, pages 777–823, 2016. 1, 2, 4, 5, 6, 7

[12] M. Kristan, J. Matas, A. Leonardis, T. Vojír, R. P. Pflugfelder, G. Fernández, G. Nebehay, F. Porikli, and L. Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016. 2, 3, 8

[13] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - European Conference on Computer Vision*, pages 740–755, 2014. 2

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[15] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 2

[16] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016. 1

[17] T. Nawaz and A. Cavallaro. A protocol for evaluating video trackers under real-world conditions. *IEEE Transactions on Image Processing*, 22(4):1354–1361, 2013. 2

[18] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. J. V. Gool, M. H. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In

IEEE Conference on Computer Vision and Pattern Recognition, pages 724–732, 2016. 1, 2, 6, 7

[19] G. Roffo and S. Melzi. Online feature selection for visual tracking. In *In Conf. The British Machine Vision Conference*, 2016. 4, 5

[20] C. Rother, V. Kolmogorov, and A. Blake. ”grabcut”: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, Aug. 2004. 5

[21] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. 1, 6

[22] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2014. 2

[23] L. Čehovin, M. Kristan, and A. Leonardis. Is my new tracker really better than yours? In *IEEE Winter Conference on Applications of Computer Vision*, pages 540–547, 2014. 1, 2

[24] L. Čehovin, A. Leonardis, and M. Kristan. Robust visual tracking using template anchors. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2016. 4, 5

[25] L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited. *IEEE Transactions on Image Processing*, 25(3):1261–1274, 2016. 1, 2, 3

[26] T. Vojir and J. Matas. Pixel-wise object segmentations for the VOT 2016 dataset. Research Report CTU–CMP–2017–01, Center for Machine Perception, Czech Technical University, Prague, Czech Republic, January 2017. 2, 3

[27] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 1, 2