23^{rd} *Computer Vision Winter Workshop*
Zuzana Kúkelová and Júlia Škovierová (eds.)
Český Krumlov, Czech Republic, February 5–7, 2018

Improving Traffic Sign Detection with Temporal Information

Domen Tabernik, Jon Muhovič, Alan Lukežič, and Danijel Skočaj
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, 1000 Ljubljana
domen.tabernik@fri.uni-lj.si

Abstract. *Traffic sign detection is a frequently addressed research and application problem, and many solutions to this problem have been proposed. A vast majority of the proposed approaches perform traffic sign detection on individual images, although a video recordings are often available. In this paper, we propose a method that exploits also the temporal information in image sequences. We propose a three-stage traffic sign detection approach. Traffic signs are first detected on individual images. In the second stage, visual tracking is used to track these initial detections to generate multiple detection hypotheses. These hypotheses are finally integrated and refined detections are obtained. We evaluate the proposed approach by detecting 91 traffic sign categories in a video sequence of more than 18.000 frames. Results show that the traffic signs are better localized and detected with a higher accuracy, which is very beneficial for applications such as maintenance of the traffic sign records.*

1. Introduction

The problem of traffic sign detection and recognition has received a considerable amount of attention in the computer vision community [14, 54]. Several transportation related problems can benefit from latest developments in traffic sign detection and recognition including automation of traffic signalization records maintenance services. However, most solutions nowadays are driven by driver-assistant systems [44] and autonomous vehicle [33] applications, and as such often lack properties for addressing automation of maintenance of traffic signalization records. In particular, addressing traffic sign detection and recognition for this task introduces two distinct requirements that are often a secondary objective in other applications: (a) detection and recognition of a large number of traffic sign categories and (b) high-precision localization. While large-scale detection and recognition of traffic signs has been addressed in our previous work using a deep learning approach [45], in this work we focus on addressing the localization accuracy of traffic sign detection applicable to the automation of the traffic signalization records maintenance services.

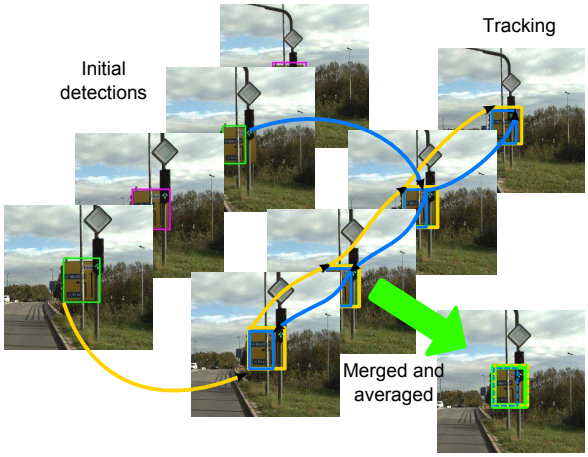

Figure 1. Detections are improved by visual temporal information.

Most traffic sign detection and recognition solutions that are driven by driver-assistant systems [44] and autonomous vehicle [33] applications often neglect high-precision localization accuracy. In those applications it is more important to accurately recognize the presence of a traffic sign and take appropriate action. However, high-precision location is not crucial for such decision making. On the other hand, in the automation of road maintenance services a high-precision location is often needed to correctly assess the quality, or correctly find the position and orientation of a traffic sign.

Most traffic sign detection and recognition systems often rely on a per-frame based information,

but an important cue for accurate localization can be found in visual temporal information. Some existing methods utilize this information to improve detection and recognition, but they focus on driver-assistant systems and are limited by online processing [39, 38]. For automation of road maintenance services, offline processing enables further improvements as strong traffic sign detection and recognition can be coupled with extensive forward and backward search of trajectories for all detections.

In this work we address the issue of high-precision localization of traffic signs in videos for road maintenance service applications. We propose a three-stage system with a per-frame visual detection module in the first stage, generation of new redundant region proposals using visual temporal information in the second stage, and location refinement from multiple detection hypothesis in the last stage (see Figure 1). As we are not limited by online processing we use a powerful deep learning approach in the detection module and extensive refinement of detection with visual tracking in the temporal module. In the temporal module we apply correlation-based visual tracking on each detection from the first stage and create trajectories in both temporal directions. With multiple overlapping trajectories we provide a strong cue to improve two aspects of traffic signs detection. First, we improve the recall rate as visual tracking recovers detections missed by the visual detector, and second we significantly improve localization accuracy of all detections by averaging multiple overlapping trajectories. We demonstrate this with a large-scale detector on 91 traffic sign categories and evaluate our approach on a sequence with more than 18,000 frames captured in a city and country-side setting.

The remainder of this paper is structured as follows: in Section 2 we review several related works, in Section 3 we provide more detail on our three-stage approach using visual detection and tracking. We perform an experimental evaluation in Section 4 and conclude in Section 5.

2. Related work

2.1. Traffic sign detection and recognition

An enormous amount of literature exists on the topics of traffic sign detection (TSD) and recognition (TSR), and several review papers are available [34, 13, 47]. Various methods have been applied for TSD and TSR. Traditionally hand-crafted features have been used, like histogram of oriented gradients (HOG) [48, 32, 20, 19], scale invariant feature transform (SIFT) [14], local binary patterns (LBP) [11], GIST [35], or integral channel features [32], whereas a wide range of machine learning methods have been employed, ranging from support vector machine (SVM) [12, 11, 51], logistic regression [35], and random forests [11, 51], to artificial neural networks in the form of an extreme learning machine (ELM) [20].

Recently, like the entire computer vision field, TSD and TSR have also been subject to CNN renaissance. A modern CNN approach that automatically extracts multi-scale features for TSD has been applied in [50]. In TSR, CNNs have been used to automatically learn feature representations and to perform the classification [40, 46]. In order to further improve the recognition accuracy, a combination of CNN and a Multilayer Perceptron was applied in [3], while an ensemble classifier consisting of several CNNs is proposed in [4, 22]. A method that uses CNN to learn features and then applies ELM as a classifier has been applied in [52], while [15] employed a deep network consisting of spatial transformer layers and a modified version of inception module. It has been shown in [43] that the performance of CNN on TSR outperforms the human performance on GTSRB benchmark dataset. Both stages of the recognition pipeline were addressed using CNNs in [54]. They applied a fully convolutional network to obtain a heat map of the image, on which a region proposal algorithm was employed for TSD. Finally, a different CNN was then employed to classify the obtained regions.

2.2. Discriminative correlation filter tracking

Visual tracking has been a very active research field. Recent visual tracking benchmarks [25, 26, 49] show that significant progress has been made in the last few years. On these benchmarks, most of the top-performing trackers are discriminative correlation filters (DCFs). They have been primarily used for object detection [18] and later introduced in visual tracking by Bolme *et al*. [2].

The MOSSE tracker [2] was based on grayscale templates, but recently correlation filters have been extended to multi-dimensional features like Colornames [9] and HOG [16, 7, 27] which significantly improved tracking performance. Henriques *et al*. [17] introduced a kernelized version of correla-

tion filters, Danelljan *et al*. [7] and Zhang *et al*. [53] investigated scale change estimation with correlation filters. To improve target localization, correlation filters were combined with color segmentation probability map [1].

Several improvements have been recently made on filter learning. Kiani Galogahi *et al*. [23] addressed the problem that occurs due to learning circular correlation from small training regions. The learning cost function was reformulated in [8] to penalize non-zero filter values outside the target bounding box. A constrained filter learning method was proposed in [29] which combines filter learning with color segmentation and allows enlarging the filter capture range.

Part-based correlation filter methods were proposed to improve target localization during partial occlusion and deformation. A tracking method for modelling the target structure with multiple parts using multiple correlation filters was presented in [28]. Lukežič *et al*. [30] proposed to treat the parts' correlation filter responses and their constellation constraints jointly as an equivalent spring system.

Recent advances in deep convolutional neural networks have been reflected in correlation filter visual tracking. Large performance boosts were reported using deep convolutional features [6, 10, 36, 31] in discriminative correlation filters, but at a cost of significant speed reduction.

2.3. Temporal information in traffic sign detection

A handful of papers also included temporal information for traffic sign detection. Rong *et al*. [38] performed detection of direction traffic sign and extracted text. They use temporal fusion of text region proposals to reduce redundant computation in consecutive frames. Šegvić *et al*. [39] proposed two-stage approach to improve detection by applying temporal and spatial constraints to the occurrences of traffic signs in video. However, they use hand crafted haar features and apply them only to a small number of categories. Both methods also work in online mode and therefore do not exploit temporal information in both directions.

3. Method

In this section we describe our three-stage approach for traffic sign detection and recognition as depicted in Figure 2. In the first stage we perform a per-frame detection and recognition with the region proposal generator and deep learning recognition similar to [45]. In the second stage we extract visual temporal information based on initial detections from the first stage and obtain new redundant region proposals. We re-verify them with deep learning recognition module from the first stage and in the last stage we integrate multiple verified detection hypotheses into final detections. We describe all three stages in more detail in the following subsections.

3.1. Initial region proposals and recognition

We obtain initial detections using the Faster R-CNN network [37] that generates region proposals and performs recognition in a unified convolutional network. Region proposals are generated with a so-called Region Proposal Network (RPN), that takes an input image and produces a set of rectangular object proposals, each with an objectness score. Recognition of regions is then performed with a Fast R-CNN network, that classifies the proposed regions into the set of predefined categories. Fast R-CNN network also performs bounding box regression and non-maxima suppression to further refine the quality of the proposed regions and retain only the best detections. The entire system is highly efficient since RPN and Fast R-CNN share their convolutional features. Recognition of multiple region proposals has minimal overhead with extra computation for each region proposal only on the last fully-connected layers. This way Faster R-CNN enables rapid detection

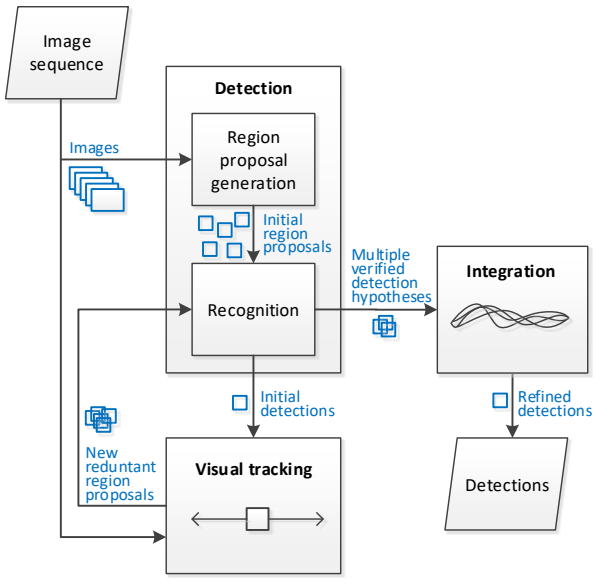

Figure 2. Processing pipeline.

and recognition in the test phase.

We apply Faster R-CNN to individual image frames and for each input image the Faster R-CNN outputs a set of object bounding boxes, where each box is associated with a category label and a softmax score in the interval $[0, 1]$. All detections with the score above a certain threshold are then outputted to the next stage.

3.2. Visual tracking

In the second stage a modified version of the DPT tracker [30] is used to exploit temporal information in our traffic sign detection framework. In the rest of the text we just use *tracker* to denote the modified version of [30]. The DPT tracker is a two-stage correlation filter based tracker consisting of the coarse and mid-level layers. The coarse layer combines correlation filter response with the color segmentation and the mid-level layer consists of four parts represented by the correlation filters. Part-based tracking methods are typically used to address non-rigid object deformation. Since traffic signs are rigid objects we only use the coarse layer of the DPT tracker [30]. The tracker uses HOG features [5] to represent the target and a color segmentation method [24] is used to improve the localization of the correlation filter. For more details on the tracker we refer the reader to [30].

With DPT we track all initial detections from the first stage that are large enough to initialize the tracker. The region is tracked forward and backward through the image sequence to create a hypothesis sequence for the location of the traffic sign in neighboring frames. We stop the tracker when certain criteria are met. In backward tracking we stop the tracker when the object size becomes too small or the tracking quality score becomes too low, while in forward tracking we stop the tracker when the object reaches the edge of the image, *i.e*. the traffic sign exits the camera's field of view. This produces a set of new redundant region proposals that are then re-verified by the recognition module. In Faster R-CNN, region re-verification can be implemented efficiently since only last fully-connected layers need to be re-computed. Multiple detection hypotheses are finally outputted to the next stage, each consisting of a tracked sequence and its re-verification score. The start and the end of the output sequence are additionally trimmed based on the quality of the recognition scores from re-verification.

3.3. Multiple sequence hypotheses integration

By tracking separate detections in the second stage we acquired a number of hypothetical sequences for each physical traffic sign that appears in the image sequence. To achieve a consensus about the traffic sign position in each image, the hypotheses are merged.

Merging was performed by finding overlapping sequences with the same class. Because more than one instance of a traffic sign with the same class can appear simultaneously (sometimes even very close together) a sufficient overlap between regions also had to occur in order to join two sequences. For each of the overlapping frames, the overlap between the mean region of the main sequence and the region of the candidate sequence was calculated. If the overlap was large enough we added the candidate sequence to the main one. Repeating this process until no more sequences could be added produced a final sequence containing multiple region proposals for each frame. Regions in each frame were then averaged to obtain the final region of a detection.

Several different scoring types can be assigned to final detected regions which is used asses classifier performance. In the evaluation we consider several scoring approaches composed of different combinations of scores:

Initial detection score: We match a verified detection hypothesis with the initial detection from the first stage based on their overlap. We only match detections of the same class and use a high IoU overlap of 0.6 to ensure we find the correct detection. If no initial detections are present in that frame, we use linear interpolation between scores of matched initial detection on neighboring frames, or we use the score from a matched initial detection from closest frame if we cannot interpolate (*e.g*. for first and last regions in the sequence). Note that when using this score the re-verification score is still used to trim the beginning and the end of the sequence.

Re-verification score: We use the score from the re-verification of the region using Fast R-CNN recognition module.

Tracking score Original DPT tracker score $\tilde{s}_t$ ranges between $[0, \infty]$, which we normalize to

range between $[0, 1]$:

$$s_t = 1 - \frac{1}{\tilde{s}_t}.$$

We then use the maximum score s_t from all merged sequences.

4. Experimental results

We evaluated our approach on the detection and recognition of 91 traffic sign categories. Evaluation is performed on a per-frame basis and we observed two important metrics: (a) mean average precision (mAP) over all 91 categories and (b) distribution of overlaps with the ground truth regions. Mean average precision shows us how the detection improved including the detection of previously missing traffic signs. The distribution of overlaps reveals the improved accuracy of localization.

4.1. Evaluation dataset

Evaluation was performed on a dataset with over 18,000 image frames, around 40,000 annotated instances and 115 traffic sign categories. The sequence contains images of 1920×1800 pixels in size taken on a path around the city of Ljubljana and its surrounding area. Images were captured with a vehicle-mounted camera at 1 frame per second and only images where any traffic signs are present were included in the dataset. Although the dataset contains 115 categories, we only used the 91 categories that intersect with the available training categories for the detector.

4.2. Implementation details

In this section we provide further implementation details of the Faster R-CNN, the DPT visual tracker and our tracker integration process used in our experiments.

The Faster R-CNN

The Faster R-CNN module was trained on an independent dataset. We used the DFG dataset from [45] and extended it to over 7000 images with more then 200 traffic sign categories overall. We only used categories with at least four real-world training samples and additionally used data augmentation with segmented real-world training samples similar as in [45]. With augmentation we ensured at least 100 training samples for each traffic sign category.

We used the Matlab implementation [37] of Faster R-CNN that is based on Caffe framework [21]. We additionally modified this implementation with online hard-negative mining during the training as described in [41]. For the deep learning model we used a pre-trained VGG-16 network [42], which has 13 convolutional layers and 3 fully-connected layers. We used the same parameters as in [45] with the exception of using full HD images during the training instead of image downscaling.

When thresholding detections for output we defined threshold individually for each category. Individual thresholds are computed as a threshold at the ideal F-measure on the Faster R-CNN training set. The same individual class thresholds are also used in tracking integration. However, for outputting detection in the first stage we use a maximum threshold of 0.4 to assure a larger set of initial detections if the threshold based on ideal F-measure would be higher.

The DPT visual tracker

Given the feature size limitation of VGG16 on minimum size for tracking a detection threshold size was set to 20 pixels (based on the smaller of the dimensions). The tracker's parameters were then set differently given the direction of tracking. As the traffic signs generally get smaller when tracking backwards and bigger while tracking forwards, the maximum and minimum scales were adjusted. We used five scales and a scale step size of 1.02 for both directions, while for backward tracking we limited minimal scale to a factor of 0.05 and for forward tracking we limited maximal scale to a factor of 7.

Visual tracking integration

Two parameters are used to control the merging of tracked sequences. The tracked sequence was terminated if the verified score fell below 20% of the class' individual threshold. Furthermore, when determining whether two sequences represent the same physical sign, a large overlap was required to connect the two sequences. We used IoU overlap threshold of 0.6.

4.3. Experiments

We first perform experiments with different scoring types, then we compare our approach against the baseline Faster R-CNN considering different region sizes and improvements in localization precision.

Table 1. Results using our temporal information with different scoring types. We report mAP at 0.5 IoU overlap over all 91 categories, and consider only regions greater than 40 pixels in size.

Scoring function	**mAP**
Detection	90.25
Re-verification	89.79
Detection $\times$ tracking	**90.62**
Re-verification $\times$ tracking	89.84

Evaluating scoring function

We evaluate four different scoring functions based on different combinations of score types as described in Section 3.3:

- Initial detection score
- Initial detection multiplied with tracking score
- Re-verification score
- Re-verification multiplied with tracking score

Results are reported in Table 1 using mean average precision (mAP) over all 91 categories. Comparing different scoring functions we see minimal difference, with the best performing function combined from the initial detection multiplied by normalized tracking score with mAP of 90.62%. Comparing scoring from the matched initial detection against the re-verification score we notice a slight improvement when using initial detection scores. This seems counter-intuitive as re-verification should provide more accurate recognition. However, re-verification may not work properly in all cases. For instance, Faster R-CNN recognition module does not work best for smaller regions (as shown by experiments in the next subsection), and interpolation from neighboring scores provides more accurate scores in those cases. This difference is minor, with improvement measuring only 0.5% over all categories. Multiplying with tracking score also improves results but only for less then 0.5%. Nevertheless, we use the best performing scoring function in all remaining experiments.

Comparing with baseline Faster R-CNN

Next, we compare our approach using temporal information against the baseline Faster R-CNN without using any temporal information. Results in mAP over all 91 categories are reported in Table 2. We can observe up to 5% improvements in mAP when using our approach. Improvements stem partially from detection of missing traffic signs as can be observed in Figure 5. First stage Faster R-CNN often misses some detections which we can now successfully recover using temporal information.

Table 2. Results at different minimal region sizes using scoring with initial detections and tracking using 0.50 IoU overlap. We report mAP over all 91 categories.

Temporal information	**Min region size**		
(mAP)	**30 px**	**40 px**	**50 px**
With (ours)	**85.05**	**90.62**	**91.67**
Without	79.76	86.79	89.52
Improvement	*+5.29*	*+3.83*	*+2.15*

Table 3. Results with different overlap thresholds and region sizes. We report mAP over all 91 categories.

Minimal region size	**Temporal information**	**IoU Overlap threshold** 0.50	0.60	0.70	0.80
30 px	With (ours)	**85.05**	**83.96**	**78.88**	**55.30**
	Without	79.76	75.44	62.20	29.44
40 px	With (ours)	**90.62**	**89.95**	**83.25**	**59.48**
	Without	86.79	83.31	70.52	34.16
50 px	With (ours)	**91.67**	**91.06**	**85.36**	**62.83**
	Without	89.52	87.76	77.23	38.79

Region size

We limit detection and groundtruth sizes that we use in evaluation since Faster R-CNN uses VGG16 network with 32-times reduction of resolution at highest layers. This reduces the amount of information for smaller regions. We ignore detections and groundtruths below certain threshold size in the evaluation, and evaluate with three different size thresholds at 30, 40 and 50 pixels in size (the smallest side must be larger then the threshold).

Results as reported in Table 2 show that performance for our approach and baseline drop with smaller region sizes, however, our approach still achieves by 5% better mAP than the baseline which does not consider temporal information. When considering regions larger then 30 pixels in size our approach achieves mAP of 85.1% while baseline achieves mAP of 79.8%. Considering larger regions with 40 or 50 pixels in size the mAP improves to 90.6% and 91.7% respectively for our approach, and to 86.8% and 89.5% respectively for baseline. Results indicate that baseline Faster R-CNN performs poorly at smaller regions, but our approach with temporal information helps to improve that. At bigger

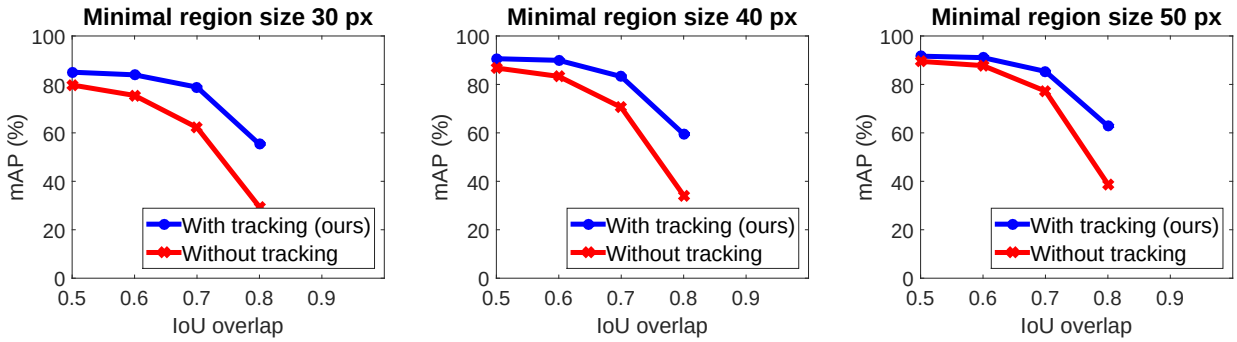

Figure 3. Results at different overlaps and minimal region sizes comparing against Faster R-CNN baseline without temporal information.

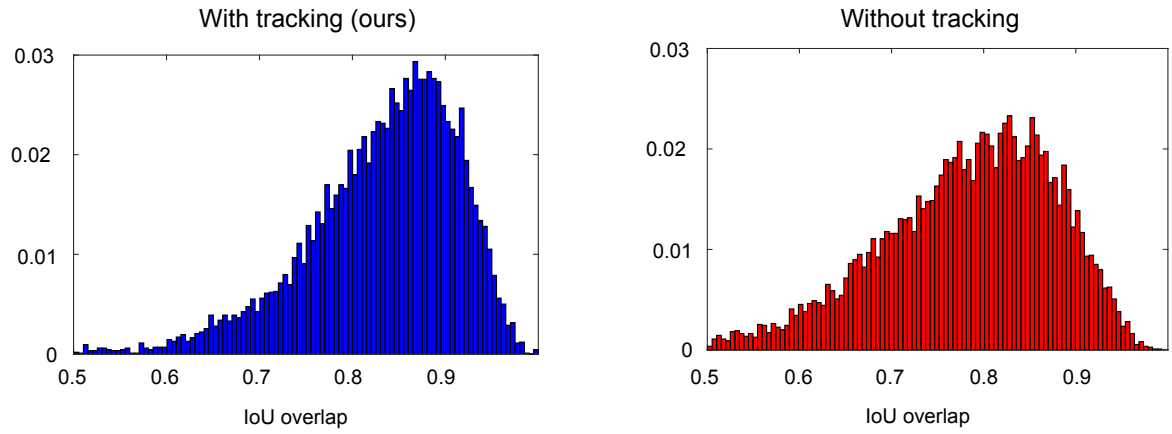

Figure 4. Distribution of true positive detection overlaps with groundtruth regions when considering minimal region size of 30 pixels. Distributions for our approach with temporal information on the left and for baseline Faster R-CNN without using temporal information on the right. Our method has detections concentrated at higher overlaps.

regions our approach also improves mAP over baseline by around 2-3%.

4.3.1 Localization precision analysis

Next, we focus on analyzing accuracy and precision of localization. We analyze localization by observing performance at different IoU overlaps of true positive detections. Results with different minimal region sizes over different overlaps are reported in Figure 3 and Table 3. Looking at mAP for different overlaps we observe a drop of performance with higher overlaps. Baseline performance drops to below 40-30% at overlap of 0.80, however, when using temporal information we retain mAP performance between 55% and 60%. At lower overlaps our approach also retains better performance compared to the baseline. This improvement is well observed when looking at the difference between red and blue lines in Figure 3. More specific numbers are also reported in Table 3.

Observed difference in performance is indicative of improved localization precision in our approach.

Table 4. Mean detection overlaps with groundtruth at different minimal region sizes. We used scoring with initial detections and tracking, and considered only correct detections with 0.50 IoU overlap.

Temporal information (mean overlap)	**Min region size**		
	30 px	**40 px**	**50 px**
With (ours)	**83.31**	**84.08**	**84.73**
Without	78.42	79.91	81.26
Improvement	+4.89	+4.16	+3.47

This conclusion is further supported when looking at the distribution of overlaps reported in Figure 4. We plot distributions of true positive detection overlaps with the groundtruth. Both distributions show concentration around IoU overlap of 0.80, however, with baseline method the distribution is wider and contains more samples at overlaps lower than 0.70. In our approach the temporal information helps to improve localization as less detections have poor overlaps of IoU below 0.60 and most are concentrated at IoU overlaps closer to 0.85. Looking at the mean of

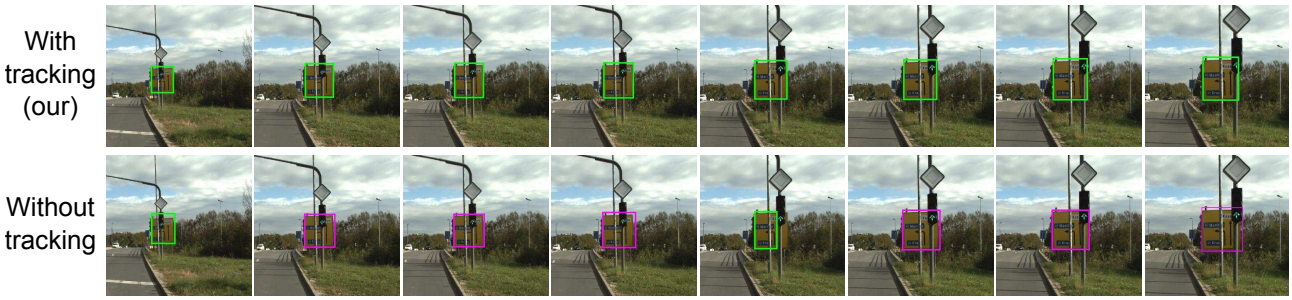

Figure 5. Example of detection recovery and location refinement. Missing detections are painted with magenta.

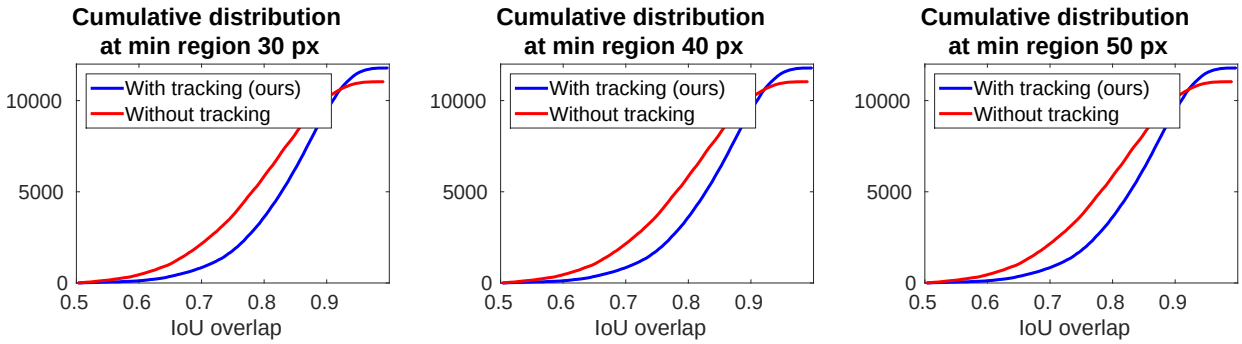

Figure 6. Cumulative distribution of overlaps.

each distributions in Table 4 reveals similar improvement with mean overlaps improved by around 4%, however, that mean value is skewed by the long tail of overlaps towards 0.50 and improvement is slightly underrated. Plotted distributions in Figure 4 show the distribution peak closer to 0.90 for our approach and around 0.80 for baseline. This difference is better observed in Figure 6 where we plot cumulative distribution of overlaps for both approaches in the same graph. Those graphs reveal our approach reduces the number of detections below overlap of 0.80 from over 50% of detections in baseline to around 30% of detections in our approach.

5. Conclusion

In this work we proposed a novel approach to improve detection of traffic signs using visual temporal information. We proposed a three stage approach with Faster R-CNN as detection and recognition algorithm in the first stage, generation of new redundant region proposals from visual temporal information using DPT tracker in the second stage and integration of multiple detections hypothesis from tracked sequences into final detections in the last stage. We evaluated our approach on 91 traffic sign categories using database sequence with over 18,000 full HD image frames and 40,000 traffic sign instances. We showed two important improvements when using visual temporal information. Compared to baseline Faster R-CNN we improved detection rate of traffic signs. We achieved this for frames with poor or missing detections where temporal information can extract cues from neighboring frames and recover the presence of a traffic sign. Moreover, we demonstrated improved localization accuracy using visual temporal information as we reduced the number of detections with poor IoU overlap. We achieved this by tracking initial detections in both temporal directions and merging them to refine the traffic sign location.

In our future work we plan on extending the approach to even larger-scale detections with all 200 traffic sign categories. Tracking may prove even more important for some difficult traffic sign categories that were not present in our sequence, such as small road marking signs or certain large directional signs with varying content. Moreover, the detection of smaller regions still remains problematic. We plan on addressing this issue by improving the first stage with features that have a smaller reduction of resolution in higher layers.

Acknowledgements This work was in part supported by ARRS research project L2-6765 and ARRS research programme P2-0214. We would also like to thank to Slovenian company DFG Consulting d.o.o., in particular Domen Smole, Simon Jud and mag. Tomaž Gvozdanović, for providing the dataset and for their help in annotating the dataset.

References

[1] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr. Staple: Complementary Learners for Real-Time Tracking. In *Comp. Vis. Patt. Recognition*, pages 1401–1409, 2016. 3

[2] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Comp. Vis. Patt. Recognition*, pages 2544–2550. IEEE, 2010. 2

[3] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. A committee of neural networks for traffic sign classification. In *IJCNN*, pages 1918–1921, 2011. 2

[4] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012. 2

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition*, pages 886–893, 2005. 4

[6] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. ECO: Efficient Convolution Operators for Tracking. In *Comp. Vis. Patt. Recognition*, pages 6638–6646, 2017. 3

[7] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Discriminative Scale Space Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(8):1561–1575, 2017. 2, 3

[8] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Learning Spatially Regularized Correlation Filters for Visual Tracking. In *Int. Conf. Computer Vision*, pages 4310–4318, 2015. 3

[9] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive Color Attributes for Real-Time Visual Tracking. In *Comp. Vis. Patt. Recognition*, pages 1090–1097, 2014. 2

[10] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: learning continuous convolution operators for visual tracking. In *Proc. European Conf. Computer Vision*, pages 472–488. Springer, 2016. 3

[11] A. Ellahyani, M. E. Aansari, and I. E. Jaafari. Traffic Sign Detection and Recognition using Features Combination and Random Forests. *IJACSA*, 7(1):6861–6931, 2016. 2

[12] J. Greenhalgh and M. Mirmehdi. Real-Time Detection and Recognition of Road Traffic Signs. *Transactions on Intelligent Transportation Systems*, 13(4):1498–1506, 2012. 2

[13] A. Gudigar, S. Chokkadi, and R. U. A review on automatic detection and recognition of traffic sign. *Multimedia Tools and Applications*, 75(1):333–364, 2014. 2

[14] M. Haloi. A novel pLSA based Traffic Signs Classification System. *CoRR*, abs/1503.0, 2015. 1, 2

[15] M. Haloi. Traffic Sign Classification Using Deep Inception Based Convolutional Networks. *CoRR*, abs/1511.0, 2015. 2

[16] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In *Proc. European Conf. Computer Vision*, pages 702–715, 2012. 2

[17] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):583–596, 2015. 2

[18] C. F. Hester and D. Casasent. Multivariant technique for multiclass pattern recognition. *Applied Optics*, 19(11):1758–1761, 1980. 2

[19] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel. Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. Ieee, 8 2013. 2

[20] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely Connected Convolutional Networks. 2017. 2

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093*, 2014. 5

[22] J. Jin, K. Fu, and C. Zhang. Traffic Sign Recognition With Hinge Loss Trained Convolutional Neural Networks. *Transactions on Intelligent Transportation Systems*, 15(5):1991–2000, 2014. 2

[23] H. Kiani Galoogahi, T. Sim, and S. Lucey. Correlation Filters With Limited Boundaries. In *Comp. Vis. Patt. Recognition*, pages 4630–4638, 2015. 3

[24] M. Kristan, V. S. Kenk, S. Kovačič, and J. Perš. Fast Image-Based Obstacle Detection From Unmanned Surface Vehicles. *IEEE Transactions on Cybernetics*, 46(3):641–654, 2016. 4

[25] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and G. Fernandez. The Visual Object Tracking VOT2017 Challenge Results. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[26] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 2

[27] Y. Li and J. Zhu. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In *Proc. European Conf. Computer Vision*, pages 254–265, 2014. 2

[28] S. Liu, T. Zhang, X. Cao, and C. Xu. Structural Correlation Filter for Robust Visual Tracking. In *Comp. Vis. Patt. Recognition*, pages 4312–4320, 2016. 3

[29] A. Lukežič, T. Vojí\'r, L. Čehovin Zajc, J. Matas, and M. Kristan. Discriminative Correlation Filter with Channel and Spatial Reliability. In *Comp. Vis. Patt. Recognition*, pages 6309–6318, 2017. 3

[30] A. Lukežič, L. . Zajc, and M. Kristan. Deformable Parts Correlation Filters for Robust Visual Tracking. *IEEE Transactions on Cybernetics*, PP(99):1–13, 2017. 3, 4

[31] C. Ma, J. B. Huang, X. Yang, and M. H. Yang. Hierarchical Convolutional Features for Visual Tracking. In *Int. Conf. Computer Vision*, pages 3074–3082, 12 2015. 3

[32] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool. Traffic sign recognition - How far are we from the solution? In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. Ieee, 8 2013. 2

[33] A. Mogelmose. *Visual Analysis in Traffic & Re-identification*. PhD thesis, Faculty of Engineering and Science, Aalborg University, 2015. 1

[34] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund. Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey. *Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012. 2

[35] D. Pei, F. Sun, and H. Liu. Supervised Low-Rank Matrix Recovery for Traffic Sign Recognition in Image Sequences. *IEEE SPL*, 20(3):241–244, 2013. 2

[36] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M. H. Yang. Hedged Deep Tracking. In *CVPR*, pages 4303–4311, 2016. 3

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 3, 5

[38] X. Rong, C. Yi, and Y. Tian. Recognizing text-based traffic guide panels with cascaded localization network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9913 LNCS:109–121, 2016. 2, 3

[39] S. Šegvić, K. Brkić, Z. Kalafatić, and A. Pinz. Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle. *Machine Vision and Applications*, 25(3):649–665, 2014. 2, 3

[40] P. Sermanet and Y. LeCun. Traffic sign recognition with multi-scale Convolutional Networks. In *IJCNN*, pages 2809–2813, 2011. 2

[41] A. Shrivastava, A. Gupta, and R. Girshick. Training Region-based Object Detectors with Online Hard Example Mining. 2016. 5

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[43] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks*, pages 1453–1460. Ieee, 7 2011. 2

[44] R. Timofte, V. A. Prisacariu, L. J. V. Gool, and I. Reid. Combining Traffic Sign Detection with 3D Tracking Towards Better Driver Assistance. In *Emerging Topics in Computer Vision and its Applications*. 2011. 1

[45] P. Uršič, D. Tabernik, R. Mandeljc, and D. Skočaj. Towards large-scale traffic sign detection and recognition. In *Computer Vision Winter Workshop*, 2017. 1, 3, 5

[46] V. Vukotić, J. Krapac, and S. Šegvić. Convolutional Neural Networks for Croatian Traffic Signs Recognition. In *CCVW*, pages 15–20, 2014. 2

[47] S. B. Wali, M. A. Hannan, A. Hussain, and S. A. Samad. Comparative Survey on Traffic Sign Detection and Recognition: a Review. *Przeglad Elektrotechniczny*, 1(12):40–44, 2015. 2

[48] Y. Wang and D. Forsyth. Large multi-class image categorization with ensembles of label trees. In *IEEE International Conference on Multimedia & Expo*, 2013. 2

[49] Y. Wu, J. Lim, and M. H. Yang. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1834–1848, 9 2015. 2

[50] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu. Traffic sign detection based on convolutional neural networks. In *IJCNN*, pages 1–7, 2013. 2

[51] F. Zaklouta and B. Stanciulescu. Real-time traffic sign recognition in three stages. *Robotics and Autonomous Systems*, 62(1):16–24, 2014. 2

[52] Y. Zeng, X. Xu, Y. Fang, and K. Zhao. Traffic Sign Recognition Using Deep Convolutional Networks and Extreme Learning Machine. In *IScIDE*, volume 9242, pages 272–280, 2015. 2

[53] M. Zhang, J. Xing, J. Gao, and W. Hu. Robust visual tracking using joint scale-spatial correlation filters. In *Proc. Int. Conf. Image Processing*, pages 1468–1472. IEEE, 2015. 3

[54] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, and W. Liu. Traffic sign detection and recognition using fully convolutional network guided proposals. *Neurocomputing*, 214:758–766, 2016. 1, 2