Sliding window object detection without spatial clustering of raw detection responses ⋆

Siniša Šegvić Zoran Kalafatić Ivan Kovaček

University of Zagreb Faculty of Electrical Engineering and Computing (e-mail: name.surname@fer.hr).

Abstract: Sliding window object detection has received remarkable attention in recent years due to great versatility and extraordinary detection performance. However, straightforward applications of the concept fail to meet criteria of real applications due to insufficient precision and inaccurate localization. Localization accuracy is especially important when the detection needs to be followed by recognition. In this paper, we present a detection approach which completely obviates the need for blind spatial clustering of nearby detection responses, which is known as a major factor of localization inaccuracy. The approach has been evaluated on traffic sign detection, where we consider the superclass of triangular warning signs. The obtained results confirm the viability of the approach and provide useful directions for future work.

Keywords: Road traffic, computer vision, traffic control, sign detection, detection algorithms, multitarget tracking, hypotheses.

1. INTRODUCTION

Object detection is an essential step in many computer vision applications and a prerequisite for successful object recognition. Therefore it is not surprising that it gains much attention in the computer vision community. Aside from the approaches that utilize some specific properties of objects to be detected, such as color or characteristic shape, many recent systems rely on generic, machine-learned binary classification in a sliding window: Viola and Jones (2004); Dalal and Triggs (2005); Zaklouta and Stanciulescu (2011). In this setup, the classifier trained on a large set of examples is applied to various image locations at different spatial scales and typically generates multiple responses that subsequently need to be grouped into distinct high-level detections. The grouping stage usually performs spatial clustering of raw detector responses and substitutes the obtained clusters with some form of average windows.

Although this kind of filtering certainly reduces the error by choosing the most probable object location, the final result is rarely perfectly aligned with the actual object position. This localization error can significantly affect the subsequent stage of object recognition. However, most of the papers on object detection report only on the detection recall and precision. In some studies (e.g. Dalal and Triggs (2005)) the detector performance is tested on cropped images of objects, while others evaluate their detectors on whole images with ground truth annotated in a separate file. The reported success of the detection is usually based on the relative area of overlap between the estimated bounding box and the ground truth reference.

⋆ This research has been supported by the University of Zagreb Development Fund, Research Centre for Advanced Cooperative Systems (EU FP7 #285939) and Croatian Science Foundation.

In our previous work on traffic sign detection and recognition (Šegvić et al. (2010)) we obtained very good detection rates (96%) by utilizing a standard cascaded boosted Haar classifier (Viola and Jones (2004)). The problem of relatively low precision was addressed in Bonači et al. (2011) by adding a stronger, specifically trained classifier that significantly reduced false positives. A neural network classifier was trained by a bootstrap procedure (Sung and Poggio (1998); Xiao et al. (2003)) and added as the last stage of the cascade. Nevertheless, we noted that the localization inaccuracy of detector responses degraded the performance of subsequent recognition process. Such effects were also discussed by Rodriguez et al. (2006), who demonstrated that localization accuracy of face detection can dramatically influence the performance of face verification applications. We tried to address this issue in Šegvić et al. (2011) by enforcing temporal consistency of the detections through video sequence. The procedure is based on combining the detections obtained from the boosted Haar detector with a differential tracker (Shi and Tomasi (1994)). However, the detector employed to initialize the tracker and to disambiguate the hypothesized tracks uses blind spatial clustering of multiple detection responses. In that way, some possibly very accurate detections are substituted by average counterparts, leading to decreased localization performance. In this paper we explore the assumption that raw, ungrouped detections have significant potential that could be exploited in an interplay with the tracker. To reach that goal, and simultaneously suppress the potential explosion of false positives, we propose to avoid the heuristic grouping of the detector responses.

The paper is organized as follows. We first describe the datasets used for training and evaluation of the system in Section 2. The proposed organization of object detection system is presented in Section 3. The Section 4 reports

on the evaluation of the proposed approach. The paper is concluded in Section 5.

2. DATASETS AND ASSUMPTIONS

The experimental evaluation has been performed in the context of traffic sign detection, on videos supplied by our industrial partner. The videos were acquired by a higher-level consumer-grade camera mounted on the top of a vehicle, along the Croatian local roads (cf. Fig.1 and Fig.3).

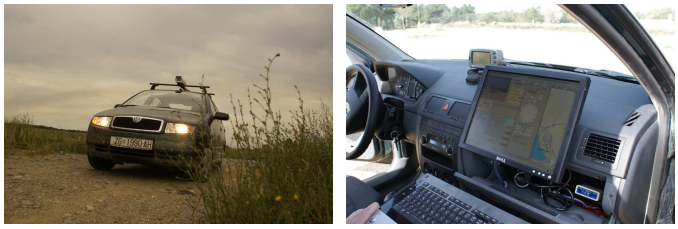

Fig. 1. The acquisition vehicle from outside (left), and inside (right).

To ensure proper testing and training, a large sample collection was acquired and carefully annotated. Each physical sign is annotated in four video frames at regular intervals (Fig. 2). The resulting collection contains about 7500 annotations of different sign classes.

Fig. 2. Each sign annotated in four frames as illustrated in the figure.

In this study we focus on the class of danger warning signs since they are most frequent (3000 of 7500 annotations in our dataset). Similar results can be expected for other ideogram-based signs. As described in Šegvić et al. (2011), we organize the 3000 annotated samples of danger warning signs into two datasets: (i) T2009, containing 2000 signs acquired with an interlaced camera; (ii) T2010, containing 1000 signs acquired with a progressive scan camera. The T2009 dataset (cf. Fig. 3(c)) is used for training, and T2010 (cf. Fig. 3(d)) for evaluation [1] .

3. THE PROPOSED APPROACH

The detection procedure based on cascading classifiers of increasing complexity has proven its performance in many applications. Therefore the proposed approach follows the same track by configuring the baseline sliding window detector for high recall. However, we try to omit the heuristic grouping and devise additional techniques to improve precision and localization accuracy. Since well trained baseline detectors generate relatively small number of false positives (e.g. less than 10 per image), these additional techniques can be computationally expensive without hurting overall performance.

[1] Both the datasets and our annotation tool can be freely downloaded from the web site of our research project:
`http://www.zemris.fer.hr/~ssegvic/mastif/datasets.shtml`
`http://www.zemris.fer.hr/~ssegvic/mastif/marker/marker.zip`.

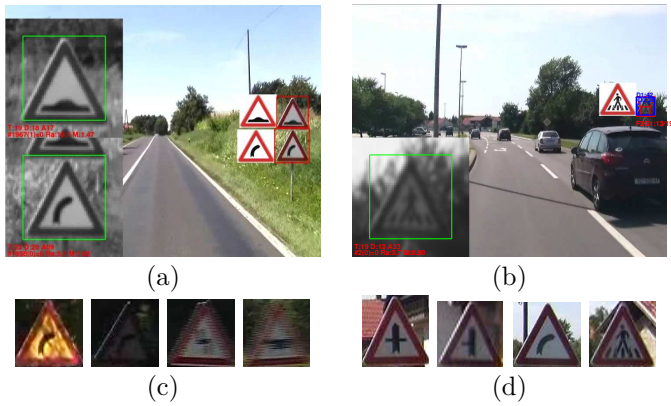

Fig. 3. Typically, the traffic signs leave the field of view when they are about 80×80 pixels large (a). However they may be smaller due to lateral displacement (b). Noisy pixels, motion blur (c) and unreliable colours are common (d).

The concept of heterogeneous classification cascades can be further pursued at the level of temporal detection sequences in video. Therefore we propose the following detection pipeline (cf. Fig.4): (i) baseline sliding window detection; (ii) introducing a strong classifier in the additional cascade stage to improve *precision*; (iii) enforcing temporal consistency to improve *localization accuracy* and further improve *precision*; (iv) enforcing learned contextual constraints to further improve *precision*. The last two stages operate on detection tracks – temporal sequences of traffic sign position, scale and appearance. This paper focuses on the first three steps of the presented pipeline.

3.1 Baseline detection

The baseline detection is performed by a boosted Haar cascade, however many other sliding window detectors would equally be applicable.. The classifier is composed of a cascade of increasingly more complex stages. Each stage consists of extremely simple classifiers implemented as single Haar features with associated polarity and threshold. These weak classifiers are composed by a boosting procedure into an ensemble forming a strong classifier.

Due to the increasing complexity of the cascade and its specific training, such classifier is able to quickly discard candidate image patches unlikely to contain the object of interest. As usually only a few image locations contain objects, this approach is extremely computationally efficient in sliding window detection. The complexity of each stage is tuned by training on increasingly harder examples (false positives from earlier stages).

By using this approach, we obtained encouraging recall of more than 95% signs detected in Šegvić et al. (2010). However, boosted Haar cascades alone do not provide enough performance for automated operation. The main concerns are: (i) unsatisfactory precision (50% or lower), due to poor generalizing over unseen negatives; (ii) localization inaccuracy, which can lead to poor recognition rates (Šegvić et al. (2010),Rodriguez et al. (2006)).

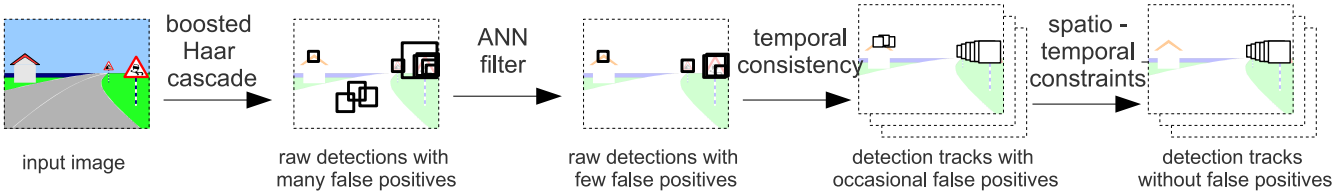

Fig. 4. The proposed detection pipeline (please see text for details).

3.2 Improving the detection by bootstrap training

As proposed in Bonači et al. (2011) we build a heterogeneous cascade for object detection in images. The baseline detector is used for fast rejection of easy negatives, while a stronger classifier is employed to deal with the hard cases. The additional classifier stage is implemented as a suitable artificial neural network using a HOG descriptor (Dalal and Triggs (2005)) as feature vector. Experiments have shown that the choice of the strong classifier is not critical since we obtained very similar results with an SVM classifier operating on the same feature vector.

The boosted Haar cascade is trained for maximum recall and reasonable precision on the T2009 dataset, while the additional ANN classifier is trained on the false positives of the Haar cascade applied to the background images from the same dataset. This kind of collecting hard negative examples is known as bootstrap training (Sung and Poggio (1998); Xiao et al. (2003)).

It is worth noting that the additional classifier stage must be applied *before* the spatial grouping step. Otherwise, the grouping may result in an alignment poorly represented in the learning set and consequently lead to drastic drop in detection rate.

By using the bootstrap filter we obtained a significant precision improvement, while only slightly harming the recall and slightly enhancing the localization accuracy. The resulting time penalty is typically very small (about 20 ms) since our boosted Haar cascades typically pass through less than ten false positives. Fig. 5 illustrates the effect of the ANN filter applied to the output of a boosted Haar cascade. However, some hard false positives still survive (cf. Fig. 6).

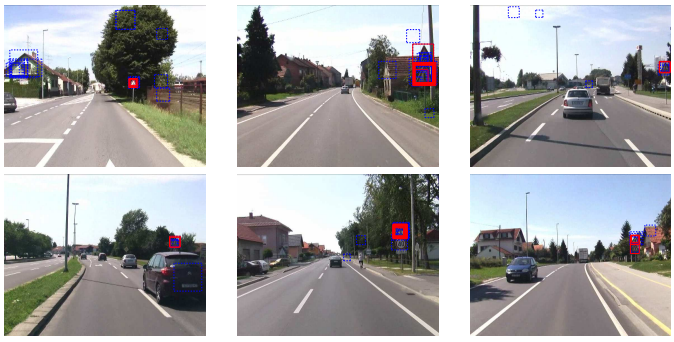

Fig. 5. The responses of a Boosted Haar cascade (blue), and the detections accepted by a suitably trained bootstrap filter (ANN+HOG, red).

Fig. 6. Some hard false positives remain even after applying the bootstrap filter.

3.3 Extracting temporally consistent detection tracks

All previously described approaches try to detect objects in individual images. However, image sequences carry dynamic information that can be utilized to improve the detection (Grabner et al. (2005)). The main idea of this component is to require the detection sequences to be temporally consistent, as proposed in Šegvić et al. (2011). To obtain that goal we track many detection hypotheses along the sequence, as shown in Fig. 7. During the tracking, the nearby detection responses are recorded as evidence which supports the hypothesis. When the tracked object leaves the field of view, we are able to pick the hypothesis which received most support from the detections.

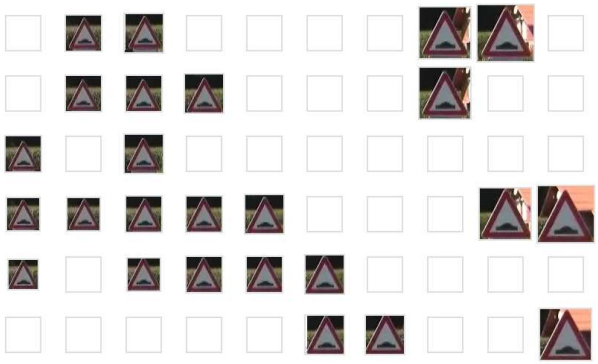

Fig. 7. Development of 6 distinct hypotheses (rows) in frames 9063-9072 (columns) of the test video. The lattice entries contain raw detections which confirm the corresponding hypothesis. Empty squares indicate frames in which the particular hypothesis has not been confirmed. Thus the most prominent hypothesis is the fourth from the top (7 confirmations).

This approach has several benefits in comparison to simple detection chaining. First, it is able to reject false positives which are either temporally inconsistent or large. Additionally, such approach is able to achieve better localization due to i) lack of blind spatial clustering, and ii)

integrating evidence from many frames. Finally, the strict tracking constraints improve the chances for distinguishing small objects from background clutter.

The proposed approach operates as shown in Fig. 8. New detection track hypotheses are seeded in the *interior* of each raw detection which happens to be displaced (cf. `thNewTrack`) from all active hypotheses (we do not track entire signs in order to avoid problems due to variable background). The hypotheses are maintained by combining the detector and the tracker (cf. `thResume`), somewhat in the spirit of particle filter. The tracks are receiving support from detector responses in their vicinity (cf. `thConfirm`). Overlapping hypotheses are grouped together into clusters (cf. `thNewTrack`) corresponding to distinct physical signs. When all hypotheses of a cluster are lost, the hypothesis with most evidence from raw detections is selected.

```
for each image do:
  # track existing hypotheses f_i
  track(image, f_i)

  # extract raw detections g_j
  detect(image, g_j)

  # determine distance matrix between
  # features f_i and detections g_j
  calculateDistance(f_i, g_j, M)

  # try to resume the tracking
  for each lost feature f_i:
    find the closest detection g_j
    if M[i][j] < thResume:
      try to resume f_i starting from g_j

  # seed new hypotheses
  for each detection g_j:
    find the closest feature f_i
    if M[i][j] > thNewTrack:
      start a new trajectory f_k at g_j
      if M[i][j] > thNewGroup:
        f_k.cluster=createNewCluster()
      else
        f_k.cluster=f_i.cluster

  # evaluate hypotheses
  for each feature f_i:
    find the closest detection g_j
    if M[i][j] < thConfirm:
      f_i.nConfirmations+=1
```

Fig. 8. The proposed algorithm for extracting temporally consistent detection tracks

In order to be able to collect a swarm of concurent hypotheses, we set both thresholds `thResume` and `thNewTrack` to `0.1` (this is much lower than in Šegvić et al. (2011)). Hence, the system generates much more hypotheses than before, and consequently improves the chances for early generation of well-localized hypotheses. As in our previous work, we characterize the distance between the hypothesized detection tracks f_i and the detections g_j as a normalized overlap between the two windows.

$$\mathrm{d}(f_i, g_j) = 1 - \frac{\mathrm{area}(f_i \cap g_j)}{\max(\mathrm{area}(f_i), \mathrm{area}(g_j))} \,. \qquad (1)$$

In exceptional cases when the windows are disjoint, we instead employ a scale-normalized Euclidean distance (Roth (2008)) with a penalization factor on scale difference.

4. EXPERIMENTAL RESULTS

We first present experiments on individual images which suggest that avoiding the spatial clustering of raw detections may favourably affect the detection performance. Subsequently, we present the results of a complete detection system featuring the proposed organization. We especially look at the localization accuracy as a performance indicator which has been often overlooked in previous work.

4.1 Measuring the localization error

We evaluate the localization accuracy by comparing the detection responses with hand-annotated groundtruth. The comparison is performed by honouring the inability of our basic detector to extract non-square detection candidates. We employ a slightly modified distance function (1), in which the height of the detection response is modified in order to match the aspect ratio of the annotation. This is equivalent to requiring that the detection needs to be correctly aligned with the bottom, left, and the right edges of the annotation bounding box.

4.2 Detection performance in individual images

In these experiments we evaluate the influence of filtering and spatial clustering to the following indicators of the detection performance: the detection recall (fraction of correctly identified traffic signs), the number of false positives (#FP) per image, as well as mean and median localization errors. The results are summarized in Table 1 for all four possible choices of employing the spatial clustering (grouping) and bootstrap filtering, or not. The table shows that filtering substantially reduces the incidence of false positives, and only slightly deteriorates the recall and the localization accuracy of unclustered detections. Spatial clustering also suppresses the false positives, however at a price of increased localization uncertainty (this effect is smaller when grouping filtered detections).

Table 1. Quality of raw detection responses.

configuration	recall	#FP/image	localization
no filter, no grouping	95%	9.5	0.070, 0.055
no filter, grouping	95%	2.4	0.171, 0.163
filter, no grouping	91%	2.5	0.095, 0.063
filter, grouping	91%	0.35	0.150, 0.139

These effects are presented in more detail in Fig. 9, as particular distributions of the localization error in the four distinct cases considered above. The rows contain distributions obtained without (top) and with (bottom) the bootstrap filter. The columns show distributions obtained without (left) and with (right) the spatial grouping. The increase of the localization error due to spatial clustering can be perceived by considering the transition from the left towards the right graphs in each of the two rows. On the other hand, by considering the transitions from the top towards the bottom, we see that filtering only slightly

disturbs the localization accuracy when no grouping is performed, while actually improving the localization of the detection clusters.

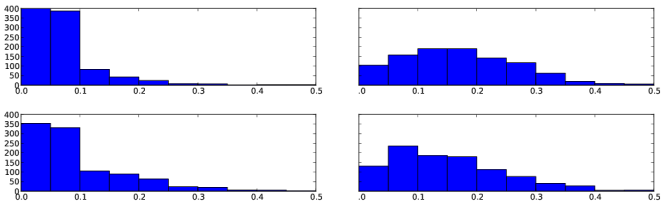

Fig. 9. Distribution of the localization error in the four cases from Table 1. The rows contain distributions obtained without (top) and with (bottom) the bootstrap filter. The columns show distributions obtained without (left) and with (right) spatial grouping.

4.3 Detection performance in video

Here we present the results of a complete detection system organized as proposed in Section 3. The proposed organization allowed us to achieve 100% recall at the system level: all physical signs from the video have been found in at least 5 consecutive frames. The present system typically produces only one false positive in 5000 images, which is considerably better than in our previous results. The only two false positives from a video of about 11000 frames are shown in Fig. 10.

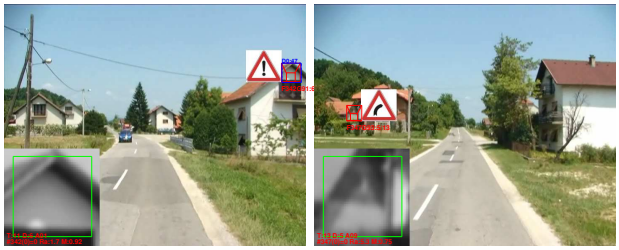

Fig. 10. The only two false positives extracted in a video of about 11000 frames.

We assess the localization accuracy by employing the methodology from the previous subsection. Fig. 11 shows the distribution of the localization error which has been estimated by comparing the extracted detection tracks with the groundtruth. We succeeded to associate 706 of the total 1037 annotations with individual detection track patches. In our view this is a very good result since most large annotations are located near the margin, where our tracker is not operable. In comparison with our previous system which employs grouping but not bootstrap filtering , the mean relative localization error has improved from about 0.12 to about 0.10 (Šegvić et al. (2011), cf. Fig 11). We note that the improvement in localization accuracy noticeably improved downstream recognition results [2] , although the detailed analysis of that effect is out of the scope of this paper.

As before, we have tried to compare the performance of our approach (filtering, no grouping) with other three configurations. Unfortunately, this was not an easy task to do, since our present system could not accomodate the two approaches without the filtering due to excessive memory constraints. We did have some success in making our system run in the case of filtering-grouping, however about 10% of physical signs have not been located, so that any comparisons would not be sensible.

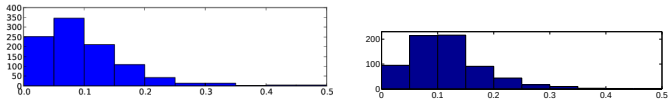

Fig. 11. Distribution of the localization error in the detection tracks extracted by the proposed approach (left). Our previous best result is shown for comparison (Šegvić et al. (2011), right).

5. CONCLUSION

This paper presented a novel organization of a moving object detection system based on binary classification in a sliding window. The proposed organization is suitable for simultaneously achieving the three main goals of object detection: high recall, high precision, and accurate localization. The main idea of the proposal is to omit blind spatial clustering of raw detections in order to preserve their localization accuracy. Instead, the raw detections are associated temporally by requiring consistent appearance throughout the sequence. This temporal grouping results in a redundant swarm of detection tracks corresponding to hypothesized sequences of object location in time. The final decision is deferred until none of the hypotheses from the swarm can be located in the current video frame. Thus, our approach is able to accumulate the evidence collected by processing all video frames in which the object of interest is within the field of view.

The presented experiments have been performed on a large test dataset of 1000 triangular warning traffic signs annotated by hand with tight bounding boxes. The test dataset is completely independent from the datasets employed to train the basic detector and the bootstrap filter. The experiments show that the responses obtained by blind spatial clustering are considerably worse localized than the best individual raw detections. The downgrade is especially noticeable for spatial clusters of basic detector responses (localization error increased by 100%), but significant effect is also observed for clustered boostrap filter outputs (localization error increased by 50%). The proposed organization succeeds to avoid spatial clustering while at the same time achieving excellent recall (all physical signs are properly detected) and competitive precision (about 1 false positive in 5000 frames).

The presented research is relevant for various applications of moving object detection. This is especially true in cases when the detected objects subsequently need to be identified, since the recognition is known to be sensitive to the localization accuracy (Rodriguez et al. (2006)). Interesting directions for future work include evaluating the proposed organization on semirigid and articulated objects such as faces and pedestrians. Future challenges include simultaneous detection of different sign classes and generic detection of table-like objects.

[2] Video presentation of our current results can be viewed from: `http://www.zemris.fer.hr/~ssegvic/pubs/syroco12.avi`

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contribution of Karla Brkić and other students in collecting the training and evaluation datasets.

REFERENCES

Bonači, I., Kovaček, I., Kusalić, I., Kalafatić, Z., and Šegvić, S. (2011). Addressing false alarms and localization inaccuracy in traffic sign detection and recognition. In *Proceedings of the Computer Vision Winter Workshop*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 886–893. doi:http://dx.doi.org/10.1109/CVPR.2005.177.

Grabner, H., Beleznai, C., and Bischof, H. (2005). Improving adaboost detection rate by wobble and mean shift. In *Proceedings of the Computer Vision Winter Workshop*, 23–32. Zell an der Pram, Austria.

Rodriguez, Y., Cardinaux, F., Bengio, S., and Marithoz, J. (2006). Measuring the performance of face localization systems. *Image and Vision Computing*, 24, 2006.

Roth, P.M. (2008). *On-line Conservative learning*. Ph.D. thesis, Graz university of Technology, Institute for Computer Vision and Graphics.

Shi, J. and Tomasi, C. (1994). Good features to track. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 593–600. Seattle, Washington.

Sung, K.K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Transactions on Pattern recognition and Machine Intelligence*, 20(1), 39–51.

Viola, P. and Jones, M.J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154.

Šegvić, S., Brkić, K., Kalafatić, Z., and Pinz, A. (2011). Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle. *Machine Vision and Applications*, 1–17. Accepted for publication.

Šegvić, S., Brkić, K., Kalafatić, Z., Stanisavljević, V., Ševrović, M., Budimir, D., and Dadić, I. (2010). A computer vision assisted geoinformation inventory for traffic infrastructure. In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 66–73. Madeira, Portugal.

Xiao, R., Zhu, L., and Zhang, H. (2003). Boosting chain learning for object detection. In *Proceedings of the International Conference on Computer Vision*, 709–715.

Zaklouta, F. and Stanciulescu, B. (2011). Real-time traffic sign recognition using spatially weighted hog trees. In *Proceedings of International Conference on Advanced Robotics*, 61–66. Tallinn, Estonia.