Three ways of using stereo vision for traffic light recognition

Conference Paper · June 2017

DOI: 10.1109/IVS.2017.7995756

3 authors, including:

Julian Müller
Daimler
8 PUBLICATIONS **51** CITATIONS

SEE PROFILE

Three Ways of using Stereo Vision for Traffic Light Recognition

Andreas Fregin[1], Julian Müller[1,2] and Klaus Dietmayer[2]

Abstract— **This paper introduces three methods to improve traffic light recognition by using the stereo camera color image in tandem with the disparity image. The first method is object candidate filtering by analyzing the disparity values inside an object candidate.**

The second method applies the relative positioning filter. Using the depth measurement obtained from the disparity image as well as the intrinsic and extrinsic calibration, a three dimensional distance from an object to the vehicle can be calculated. Based on known real world traffic light locations, the filter is able to suppress thirty to seventy percent of false positives while only decreasing the detection rate by one percent. This result shows the huge potential for range filtering in traffic light recognition in general.

The third method is the hypothesis size enhancement when re-projecting a hypothesis into the image. This process is enabled by a real world traffic light model in conjunction with the depth measurement. It is shown that all true positive hypotheses will benefit from this technique, resulting in a massively better overlap with the ground truth labels. When evaluated framewise, re-projection can improve detection rate by up to fifteen percent.

This work primarily enhances traffic light hypotheses obtained by a baseline detector and thus requires a hypothesis disparity value. As a further contribution this paper presents different methods for determining a hypothesis-wide disparity and evaluates the differences in quality and quantity.

I. INTRODUCTION

Recently, the topic of the first level-four autonomous vehicle has been ubiquitous in the intelligent vehicles community. Among several recognition tasks, traffic light recognition is an essential component of autonomous driving. While car-to-infrastructure communication systems have gained popularity over the last years, a traffic light in most cases still is designed for purely visual detection, without any radio connection to the vehicle. Therefore, the research in traffic light recognition is focused on monocular image processing. In this paper a stereo camera provides a disparity image along with the regular color image. We introduce three different methods of enhancing a baseline detector by using depth information.

II. RELATED WORK

In scientific literature researchers started describing systems for recognition tasks in automotive environments as early as in 2001. Gavrila et al. [1] started using a stereo camera inside a research vehicle. The disparity image is used for free-space and pedestrian detection in their work. Traffic lights are detected using a color classifier.

A first contribution to traffic light detection using stereo vision was made by Lindner et al. in 2004 [2]. They used the depth measurement to suppress false positive hypotheses by generating a second hypothesis purely based on stereo vision and a real world traffic light model and comparing it with the hypothesis obtained from a color detector. Another filtering technique was presented in [3] by limiting the detection range. They used a monocular camera and a fixed height-over-ground assumption for calculating a hypothesis' distance over ground. Their approach is similar to our relative positioning filter described in Section VI, but they only filtered within a single dimension. Our filter has an upper and lower bound in three dimensions.

In 2013 Daimler navigated their Bertha-Memorial-Route autonomously using several different sensors [4]. A stereo camera was also part of the setup but due to the limited horizontal field of view, traffic light recognition was performed by a wider-angle monocular camera. Another research vehicle equipped with a stereo camera along with several other sensors is described in [5]. The authors performed traffic light recognition as well, but there is no information whether and how the task relied on stereo vision.

The only public available[1] dataset using a stereo camera that includes traffic light annotations was published in 2015 by Philipsen et al. [6]. In their work, the authors identify the requirement for stereo vision in traffic light recognition as the main reason for using a stereo camera to record their sequences. The same research group also mentioned the rare use of stereo vision for traffic light recognition research in their well-researched work published in 2016 [7].

Based on the work referenced above, it appears the use of a camera image along with a distance measurement for traffic light recognition was rarely described in literature. Some publications only vaguely describe the use, no publication at all evaluated the benefits of depth information for traffic light research.

III. BASELINE SYSTEM

This work relies mainly on four components: First, a stereo camera consisting of two synchronously triggered monocular cameras and a frame-grabber for calculating disparity. The rectified images as well as a disparity image are the resulting output. Secondly, a real world dataset consisting of numerous

[1]Daimler AG, Research and Development, Wilhelm-Runge-Str. 11, 89081 Ulm, Germany

[2]Ulm University, Institute of Measurement, Control and Microtechnology, Albert-Einstein-Allee 41, 89081 Ulm, Germany

[1]An .edu email address is required for registration, wherefore we could not verify the dataset and its quality.

traffic light image sequences recorded by a camera-equipped research vehicle in German cities. Finally, a detection module performs color segmentation on the camera images and a hypothesis generator estimates an upright bounding rectangle from a color segment.

A. Stereo Camera

We use a prototypical camera consisting of two two-megapixel imagers triggered by an FPGA processing hardware to acquire the images. Applying the established SGM algorithm originally proposed in [8] and augmented by the gravitational constraint in [9], the processing hardware generates bayer-patterned raw images, rectified full-resolution gray images, and a disparity image with a half-megapixel resolution. Because the image resolutions of raw/rectified images and the disparity image differ, mapping is necessary when the disparity information from a corresponding color image pixel is needed.

The cameras are equipped with lenses that achieve approximately 60 degrees horizontal field of view. Due to image resolution limits, the maximum detection distance for traffic lights is around 100 meters. The minimum detection distance may be affected by the horizontal and vertical field of view, depending on the location of a traffic light during a particular approach.

B. Dataset / Ground truth

A large number of image sequences were recorded while approaching traffic lights with a camera-equipped research vehicle in 2015. The sequences are from ten large cities in Germany, resulting in images with several thousand different traffic lights. While we plan to release this dataset in future, at the time of this work only a limited number of sequences from a few cities already include hand-annotated traffic lights labels. They are annotated as upright bounding rectangles. Furthermore, information such as the number of lights, traffic light state and relevance for the current driving route are annotated. For this work, we use a sub-set of 2345 images with around 5200 annotated traffic lights. More details can be found in Table I.

TABLE I

DATABASE STATISTICS: FOR THIS PAPER A LARGE DATABASE OF HAND-ANNOTATED TRAFFIC LIGHTS WITH OVER 2300 IMAGES IS USED.

Number of	Count
Images	2345
Traffic lights	5192
Images with relevant traffic lights	2268
Total relevant traffic lights	4093

For a better understanding of the quality of the available raw data, Figure 1 shows a typical example of an annotated traffic light in the raw color image as well as the corresponding region in the disparity image and in a colorized three dimensional point cloud.

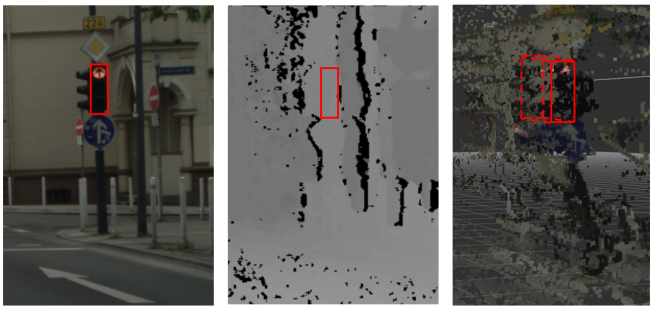

Fig. 1. Example of an annotated traffic light (left) as well as its corresponding region in the disparity image (center) and a colorized 3D point cloud (right).

C. Baseline Color Detector

A color detector is used as the baseline detector to quickly generate traffic light candidates. Based on the annotated traffic lights, this detector learned the characteristic colors of red, yellow and green traffic lights. A lookup table is generated in which the probability for each class is stored, allowing different operating points to be set simply by biasing the probability of one or several classes. With one lookup-operation per pixel the detector can create a color segmented image that includes the classes *traffic light-red*, *traffic light-yellow*, *traffic light-green* and background. Since the lookup table includes the probabilities for all classes, the certainty for the winning class can be determined. Figure 2 shows two different operating points for the detector applied on an example image. The left image shows an operating point that produces very few color segments, whereas an operating point that produces many traffic light color segments is shown in the right image.

D. Hypotheses Generator

Prior knowledge is applied to obtain a traffic light hypothesis from a color segment. In Germany, traffic lights are installed vertically, typically consisting of three lamps, where the top is red (state: *stop*), the center lamp is yellow (state: *prepare to stop*) and lowest lamp is green (state: *go*). This prior knowledge in combination with the size, position and color of a segment in the segmented image allows a traffic light hypothesis circumscribing the traffic light housing to be generated.

E. Evaluation Metrics

To evaluate the performance of the detector, the generated hypotheses and the ground truth labels are used to calculate the number of true positives (TP) and false positives (FP). A TP is defined as a frame with a hypothesis on at least one relevant label. In addition, the number of misses (false negatives, FN) for frames with none of the relevant labels is determined. The calculated numbers are used to generate a receiver-operating-characteristic (ROC). A regular ROC-curve uses the detection rate obtained by

$$TP_{rate} = \frac{TP}{TP + FN} \tag{1}$$

Fig. 2. Examples of different operating points of the color segmentation. Pixels in gray are classified as background, whereas pixels in red, yellow and green are classified as their corresponding traffic light color. The brightness of a pixel shows its certainty for the winning class: a darker pixel has a lower certainty than a brighter one. On the left, a very conservative operating point is chosen. It will create almost no false positive segments but also detects only very few traffic lights. In the right image, an operating point that creates numerous color segments is shown.

and the false positives rate obtained by

$$FP_{rate} = \frac{FP}{FP + TN}. \quad (2)$$

However, because the proposed baseline detector does not generate hypotheses for the background-class (true negative hypotheses, TN), the frame-wise mean of false positives (FP/frame) is used as the Y-axis instead. The evaluation takes redundancy and the relevance of traffic lights into account and thus follows the proposed TLF_R metric described in [10].

To generate the ROC-curve detection, requirements have to be defined. As both hypothesis and ground truth label are upright bounding rectangles, the intersection-over-union (IoU) between both rectangles can be used. An IoU greater than zero is a hypothesis that overlaps with a ground truth and thus can be seen as a detection (blue curve in Figure 3). However, when using neural networks or boosted cascade classifiers as the hypothesis verification modules, a hypothesis with an IoU slightly above zero will most likely not be classified as a true positive. In our experience, a minimum IoU of 0.5 is a more reasonable choice 3).

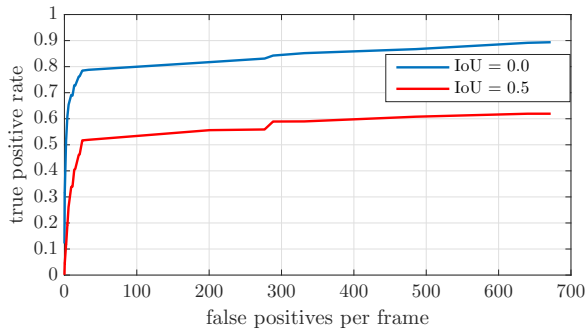

Fig. 3. ROC-plots for two different definitions of the of the minimum intersection-over-union of the baseline color detector: The blue curve shows the TP-rate for IoU greater zero, whereas the red curve shows the TP-rate for IoU equal or greater 0.5.

Figure 3 illustrates a serious decrease in true positive rate for a required IoU of 0.5. This finding is the key motivation for using stereo vision together with the baseline detector. In the following sections, methods for recovering the lost detection percentage as well as reducing the false positive rate are introduced.

IV. HYPOTHESIS DISPARITY

The distance between vehicle and a potential traffic light can be calculated from the disparity image of a stereo camera. However, an object in a camera or disparity image is just a cluster of pixels belonging to that object. Different methods are applied to calculate the disparity of an object (Figure 4):

A. Mean Hypothesis Disparity

A common method for obtaining a hypothesis-wise disparity is determining the mean disparity value of all pixels belonging to the hypothesis. It must be noted that an upright rectangle may not end properly at an objects edge, but may include the objects background as well. In general, traffic lights can be described by upright bounding boxes very well. Depending on the orientation of the camera in a vehicle and the viewing angle onto an object, the mentioned characteristic may occur even on traffic lights. To avoid including background pixels a hypothesis can be shrunken in its size by a few percent.

B. Mean Disparity on Selected Pixels

Depending on the size of a hypothesis and the overall number of hypotheses per frame, calculating the mean disparity can become a computational expense. A traffic light is an upright object of limited real world dimensions, the frontal view of the traffic light should thus have a relatively homogeneous disparity region from the perspective of a front-looking camera. This observation is the motivation to only sample a few selected pixels from the hypothesis to calculate the mean disparity.

C. Center Disparity

A special case of selecting a few pixels to calculate the mean disparity is distance estimation based on a single pixels disparity. For this method, the hypothesis center offers the best choice as it has a high probability of belonging to the object. It cannot be guaranteed that the center pixel of each hypothesis is a valid value (e.g. stereo shadow). Our algorithm in this case will try to use a direct neighbor pixel instead.

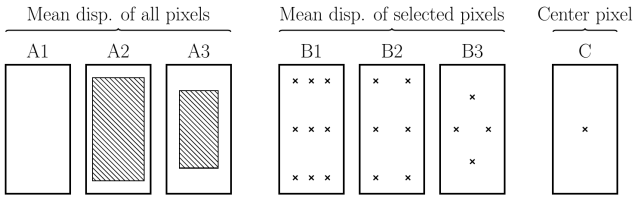

Fig. 4. To calculate the mean disparity of a hypothesis, seven different methods were implemented and evaluated. A box symbolizes a label and the shaded area the pixels used in the mean disparity calculation. Below, calculating the mean using all pixels of a hypothesis as well as two shrunken labels (methods *A*), methods *B* take only nine, six or four selected pixels into account, whereas for method *C* the hypothesis disparity is estimated from a single pixel at the center of the hypothesis.

D. Evaluation

We had neither a high precision vehicle localization system nor a high resolution map with high precision traffic light locations while recording the sequences. Thus, no measured ground truth distances are available and the different proposed methods are compared to one another. Below the absolute disparity deviation, the IQR is also calculated and drawn as a box-plot (Figure 5). A small deviation indicates a high homogeneity of the disparity values on a hypothesis. The figure shows the results of the proposed calculation methods. For methods *A* either the whole label was used (*A1*) or labels were reduced to 80 percent (*A2*) and 60 percent (*A3*) of their original sizes, respectively.

Figure 4 illustrates selected patterns used for the mean disparity calculation. For methods *B*, three selection patterns have been evaluated. The first selects nine pixels, the second six pixels and the last selects four pixels. Whenever a selected pixel has an invalid value, it is ignored. Method *C* is of particular interest as it is by far the fastest method. As shown in Figure 5, all methods are very similar. Method *A3* was used as the reference because a shrunken label is robust against outliers while using a higher number of pixels than methods *B*. Method *C* has the highest deviation from the reference. It can be observed that even method *C* has an IQR value that is below the expected disparity accuracy of 0.3 pixels.

V. HOMOGENEITY VERIFICATION

As shown in the previous section, different methods for obtaining a hypothesis disparity can be used. For traffic

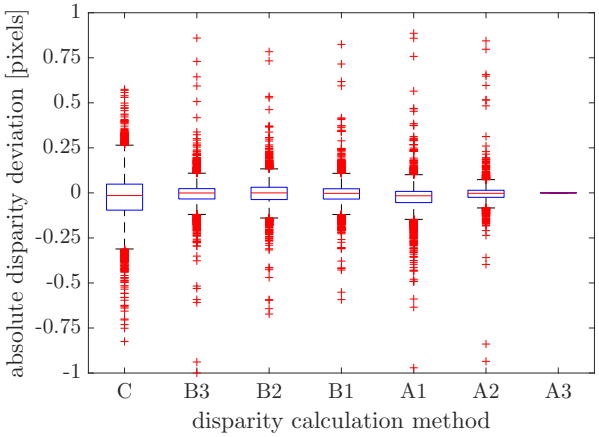

Fig. 5. Comparison of seven methods for obtaining a mean disparity for a hypothesis. Since no ground truth disparity is available, the methods can only be compared to one another. Method A3 is used as the reference as it is expected to provide the most reliable results. All examined methods show an expected median error of zero. Methods A1, A2 as well as B1-B3 deliver comparable results whereas the center pixel method has the highest IQR value. However, the IQR value of all methods is smaller than the expected disparity accuracy (∼0.3 pixels). Thus, the center pixel method is used in our system as it has by far the lowest computational cost.

lights, a homogeneous disparity can be assumed. This observation can be used for a hypothesis verification step. Inappropriate hypotheses can be filtered by their disparities standard deviation. This verification step cannot be combined with the center pixels disparity since no standard deviation can be computed with this method. A further related verification filter adds up invalid disparity values. While calculating a mean disparity the algorithm has to be aware of invalid values anyway, returning the number of invalid disparity values therefore adds virtually no complexity.

For the evaluation of the proposed verifier, the generated hypotheses from 25 operating points of the color detector were split into false positives and true positives. The filters are applied and the frame-wise detection metric for relevant traffic lights TLF_R is applied. Since verification is a filtering operation, it will add no additional detections (true positives). A decrease in false positives per frame by up to twenty percent as well as a small decrease in the number of detected frames (1-2 percent) is achieved. As can be observed in Figure 7, the number of false positives per frame still is relatively low. Therefore it depends on the specific operating point whether homogeneity filters are useful for a recognition system or not. In general, the more hypotheses a particular detector creates, the more effective these verifiers are.

VI. RELATIVE POSITIONING FILTER

Additional prior knowledge can be obtained from the law for placing traffic lights in Germany [11]. Traffic lights are placed along the streets or over the streets at a stipulated minimum height. Also, the maximum installation height can be guessed as traffic lights are mounted at a height so that

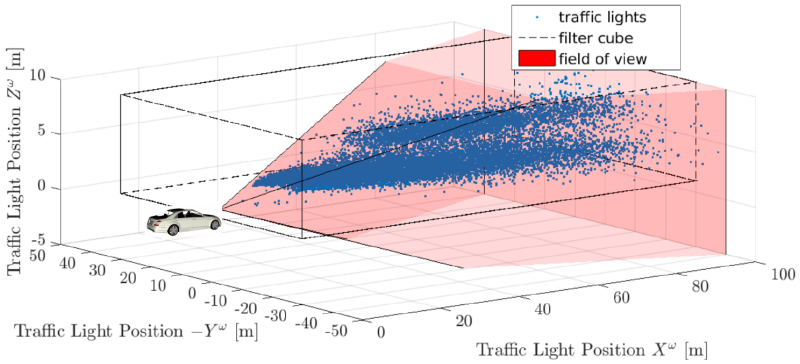

Fig. 6. 3D illustration of all annotated traffic lights in the vehicle coordinate system as well as the field of view of the camera. It can be observed that traffic lights occur in a wide range in the longitudinal direction (X^ω-axis) and in a much smaller range in the lateral direction (Y^ω-axis). Two modes can be interpreted in height over ground (Z^ω-axis), belonging to traffic lights over the sidewalks and above the road, respectively.

pedestrians and bikes (for traffic lights placed at sidewalks) or large vehicles such as buses and trucks (for traffic lights above the streets) are not affected. For this work, a large database is available as well. With the mean disparity of a label and the pinhole camera model, the distance in X^ω, Y^ω and Z^ω direction from the traffic light to the camera can be calculated. Since the pose of the camera in the vehicle coordinate system is also known, these points can be transformed into vehicle coordinates. When analyzing the 3D range as shown in Figure 6, it can be seen that all traffic lights are located inside a compact range. This knowledge can be used to filter all hypotheses by their relative position to the vehicle. As presented in the previous section, the generated hypotheses from twenty-five operating points of the color detector are used. They are divided into false positives and true positives before applying the filter on both groups. For evaluation, the TLF_R metric is used again. The results are shown in Figure 7. The number of false positives can be strongly reduced as this filter suppresses thirty to over seventy percent of false positives while only losing a very small percentage of detected objects. Since the proposed method is a filter operation, no additional detections are obtained.

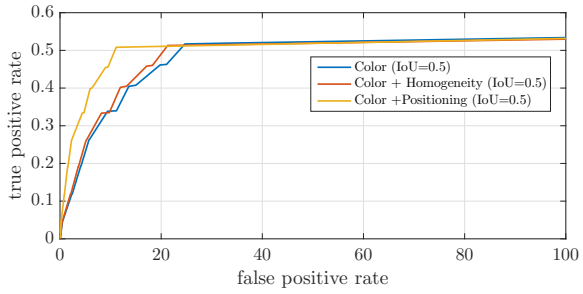

Fig. 7. ROC curves of the baseline detector (blue) and after applying the proposed range filter (yellow). The filter suppresses 30 - 70 % of false positives (depending on the operating point of the detector) while only loosing around 1 % detection rate. The homogeneity-based filter (red) reduces the number of false positives by around 8 %.

VII. HYPOTHESIS SIZE IMPROVEMENT

The distance information in conjunction with a real-world model of an object can be used further to re-project the object into the camera image (Figure 8). This method is especially useful when the primary detector only detects a single feature of the object. In this case, a potentially error-affected object-hypothesis must be estimated from this feature-hypothesis. This problem can clearly be observed with the baseline color detector. The feature *traffic light lamp* is detected and a full traffic light hypothesis is estimated from only the size and position of this feature in conjunction with a 2D model. Upon the estimation process an incomplete detected lamp leads to an insufficient hypothesis which can clearly be observed in an IoU histogram as shown in Figure 9: Many hypotheses are actually detected but their IoU is not sufficient for later verification steps (blue bars). With a well-fitting real world model and a relatively precise depth measurement, the IoU of these hypotheses may increase when applying the re-projection step.

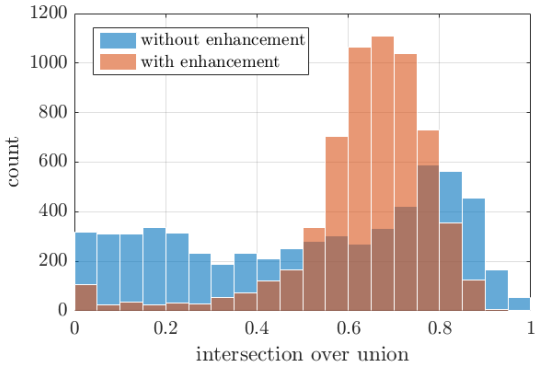

Fig. 9. Overlap histogram of the baseline detector at operating point 15. Blue bars: Many hypotheses are detected but as the object hypotheses generation step strongly relies on the size of the detected region, the amount of generated hypotheses is insufficient. By using the distance information in conjunction with a real world object model, the re-projected hypotheses lead to significantly improved IoU (red bars).

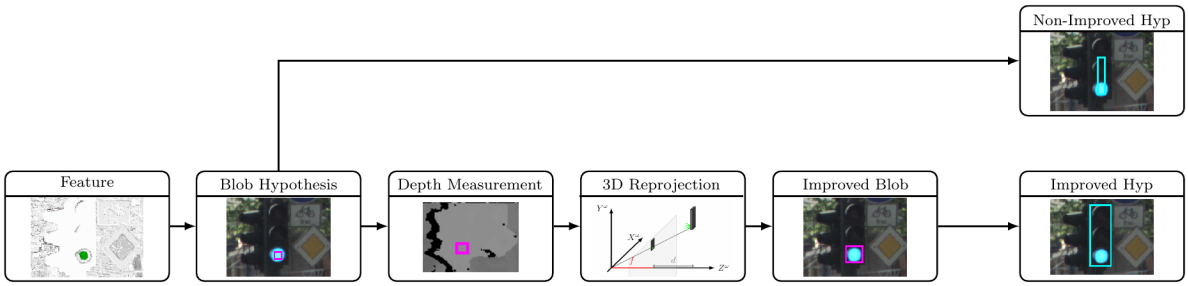

Fig. 8. Two-stage traffic light detection. Initially a color feature detector segments the input image into the four classes red, yellow, green and background. A blob hypothesis is created, circumscribing the detected color segment. Without depth improvement, traffic light hypotheses are created based on the blob hypothesis. Inaccuracies in blob hypotheses directly influence the created traffic light hypotheses. The proposed enhancement method uses the depth measurement at the blobs' location as well as a traffic light model to determine the blob in 3D. A re-projection of the 3D blob results in an accurate blob hypothesis. Based on this result, an improved traffic light hypothesis can be created.

Furthermore, primary detector configuration may be much more restricted if only a very small percentage of the object or an objects feature has to be detected. Thus, the re-projection step not only increases the detection rate but may also lead to a decrease in the number of hypotheses. A color detector would in principle only need to detect a single pixel in the middle of a traffic light lamp if the re-projection step is powerful enough to generate a sufficient traffic light hypothesis by measuring the pixels depth.

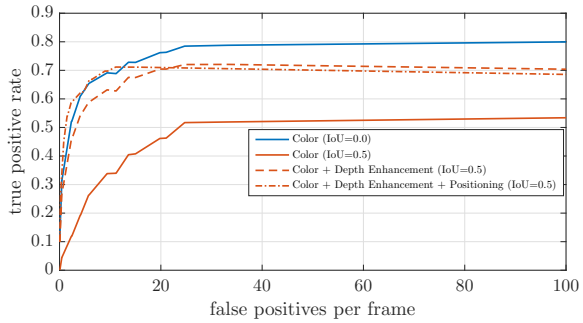

Fig. 10. ROC curves of the baseline detector and the two-stage-detector of color detection and model and depth-based re-projection, as well as positioning filter. Compared to the reasonable overlap criterion (IoU-0.5), the two-stage-detector achieves a significantly better detection rate. At some operation points, the IoU-0.0 curve can even be reached.

For evaluation of the proposed two-step detector consisting of color detection and re-projection, another true positives IoU histogram is generated (Figure 9). A shift towards higher IoUs can clearly be observed. The final comparison is the ROC-curve of the re-projected hypotheses compared to the ROC-curves obtained by the baseline detector (Figure 10). The higher IoUs translate directly into higher detection rates. For some operating points of the detector, the gain in detection rate is as high as thirty percentage points. Even the IoU-0.0-curve can be reached at some operating points. Still, the maximum detection rate does not reach the IoU-0.0 curve. There are several reasons for this: A noisy disparity measurement, an imperfect real world traffic-light model, mistakes during the labeling process, and naturally varying aspect ratios of the labeled ground truth that cannot be modeled.

VIII. CONCLUSION

Three methods for using the disparity image of a stereo camera in tandem with a color detector for traffic light recognition have been introduced. Because the disparity information is used on an object hypothesis, a disparity value for the hypothesis as a whole has to be determined. Seven different methods were introduced for this purpose. An evaluation with a real world dataset showed that even the simplest method - the hypothesis center pixels disparity - can be used. The standard deviation of the hypothesis disparity as well as an outliers-count was used to compare the different methods. Two filters were derived from this step. They can suppress up to twenty percent of false positives while only decreasing the detection rate around one percent. Since the baseline detector used in this work does not generate many hypotheses per frame, it depends strongly on the specific operation point of the detector whether the use of the filters is recommended. Especially since the introduced three dimensional distance filtering is capable of suppressing more than half of all false positive hypotheses in some operating points, this filter shows far greater potential for decreasing the number of false positives.

The third method for using the disparity image, the re-projection based on depth and real world traffic light model, showed especially good results for generating hypotheses with high ground truth overlaps. From our experience, the achieved overlaps will be a significantly better starting point for subsequent verification steps. Because this technique could be utilized for any objects with a real-world model of known size, it could be used for other recognition tasks as well.

REFERENCES

[1] D. M. Gavrila, U. Franke, C. Wohler, and S. Gorzig, "Real time vision for intelligent vehicles," *IEEE Instrumentation & Measurement Magazine*, vol. 4, no. 2, pp. 22–27, 2001.

[2] F. Lindner, U. Kressel, and S. Kaelberer, "Robust recognition of traffic signals," *IEEE Intelligent Vehicles Symposium, 2004*, pp. 49–53, 2004.

[3] M. Diaz-Cabrera, P. Cerri, and J. Sanchez-Medina, "Suspended traffic lights detection and distance estimation using color features," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, (Anchorage, AK), pp. 1315–1320, IEEE, 2012.

[4] U. Franke, D. Pfeiffer, C. Rabe, C. Knoeppel, M. Enzweiler, F. Stein, and R. G. Herrtwich, "Making bertha see," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 214–221, 2013.

[5] D. Barnes, W. Maddern, and I. Posner, "Exploiting 3D semantic scene priors for online traffic light interpretation," *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2015-Augus, no. Iv, pp. 573–578, 2015.

[6] M. P. Philipsen, M. B. Jensen, M. M. Trivedi, A. Mogelmose, and T. B. Moeslund, "Ongoing work on traffic lights: Detection and evaluation," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, aug 2015.

[7] M. B. Jensen, M. P. Philipsen, A. Mogelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–16, 2016.

[8] H. H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 2, pp. 807–814, 2005.

[9] S. K. Gehrig, H. Badino, and U. Franke, "Improving sub-pixel accuracy for long range stereo," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 16–24, 2012.

[10] A. Fregin and K. Dietmayer, "A closer look on traffic light detection evaluation metrics," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 971–975, IEEE, nov 2016.

[11] A. V. Forschungsgesellschaft fuer Strassen- und Verkehrswesen, ed., *Beispielsammlung zu den Richtlinien fuer Lichtsignalanlagen : RiLSA*. Köln: FGSV-Verlag, 2010.