Fast and Robust Cyclist Detection for Monocular Camera Systems

Wei Tian and Martin Lauer
Institute of Measurement and Control, KIT, Engler-Bunte-Ring 21, Karlsruhe, Germany

1 MOTIVATION

In recent years vision-based detection becomes more and more important among automobile industries. One of the most successful examples is pedestrian detection, which is broadly applied in driver assistance systems. In comparison to that, cyclist detection hasn't attracted much interest from most researchers, although cyclists are ranged in the same level with pedestrians in the group of vulnerable road users (Gandhi and Trivedi, 2007).

However, cyclist detection is a also an important task. Not only because cyclists are as vulnerable as pedestrians but also they share the same road with other vehicles with a comparable velocity, which makes them much easier to get involved in accidents and to suffer from severe injuries and even fatalities. In addition, the available systems of cyclist detection are mostly radar-based. They are sensitive to all the objects in surroundings so that more false positives can be generated, reducing the robustness of the system. In the mean while, vision based pedestrian detection has already achieved high precision (Dollár et al., 2012). Therefore, designing a vision-based detection system specially tailored to cyclists is necessary.

2 RESEARCH PROBLEM

For implementation of such a system several problems remain to be solved. At first, the appearance of cyclists in the image obviously varies according to the viewpoints. This requires extra algorithms to deal with multi-view detection and the precision should be maintained as well. This problem is rarely taken into consideration in pedestrian detection. Secondly, due to small bodies, cyclists can be easily occluded by other objects, e.g. cars. To capture partially occluded cyclists is also challenging. Thirdly, collision prediction requires knowledge of the cyclists' trajectories. This knowledge allows us to estimate the risk of an accident which can be used to implant active protection systems or to send warning signals to car drivers and cyclists. At last, the detector should be able to run in real time. This is very important for application on low cost hardware, as cost reduction is always the issue of industries. So this system should provide a solution to these points.

3 STATE OF THE ART

In the last decade enormous progress has been seen in the area of vision-based object detection. Viola et al. in (Viola and Jones, 2004) designed a cascade detector integrated with Haar-like features and achieved a real time speed. Dalal et al. proposed HOG (histogram of oriented gradients) feature in (Dalal and Triggs, 2005), improving the detection precision significantly. Based on that, Felzenszwalb et al. introduced DPMs (deformable part models) in (Felzenszwalb et al., 2008). By applying part filters and allowing displacements between them, this method has achieved the best results at one time. As another alternative, Wojek et al. combined HOG with Haar-like, shapelet (Sabzmeydani and Mori, 2007) and shape context (Mori et al., 2005) in (Wojek and Schiele, 2008) and proved that it performs better than individual features. In (Walk et al., 2010) Walk et al. added color and motion information to this method and improved the precision in one more step. Unlike them, Dollár utilized different features to assemble integral channels (Dollár et al., 2009). By using a cascade detector, he achieved a comparable result and a much faster speed. As extension, Benenson et al. in (Benenson et al., 2012) made an even faster detector with the help of differently scaled models and stixels from stereo images. These methods are mainly focused on detection of pedestrians.

As for cyclist detection, Rogers et al. in (Rogers and Papanikolopoulos, 2000) modeled bicycles by two circles. Cho et al. applied DPMs in combination with an extended Kalman filter to detect and track bicycles (Cho et al., 2010). However, bicycle detection makes little sense in some situations, e.g. parked bicycles without riders, because it is the cyclist not the bicycle that needs protection. Instead, Li et al. used

Tian W. and Lauer M..
Fast and Robust Cyclist Detection for Monocular Camera Systems.

HOG-LP to detect crossing cyclists (Li et al., 2010). As cyclists from other directions are not concerned, the usability of their method is limited.

4 OUTLINE OF OBJECTIVES

In this work we are aiming at building a vision-based detection system for cyclists. To solve the multi-view problem, we divide the cyclist's viewpoints into several subgroups and for each group a detector is built. Occluded cyclists are only partially visible, which means that part detectors are required. Based on the detector results, hypotheses about cyclists are made after probability analysis. As a fact, time cost for detection is strongly dependent not only on the detection algorithm but also on the image size. So we integrate ROI (region of interest) extracion into our framework. With its help, only interesting regions in the image are concerned, so that detection can further speed up. Moreover, tracking algorithms in 2-D and even in 3-D environment are integrated to improve the detection results and to estimate the trajectory. At last, we are also planing to extend our framework to detect general object classes.

5 METHODOLOGY

This section consists of four parts. The first part describes our selected detectors. The second part introduces the methods to deal with partial occlusions. The third one focuses on monocular camera based ROI extraction. Finally, we briefly discuss about tracking algorithms.

5.1 Detector Structure

5.1.1 Unimodel vs. Multi-model

To search for objects in an image, the mostly used method is sliding window detection, which is also applied in our project. In this method, a window with the same size as the training samples is shifted over the image. Hypotheses about objects are made based on features which are calculated inside the window. Since only objects within this window can be detected, it is important to choose the window size appropriately.

For simplicity, most researchers use only one fixed window size to detect pedestrians, since the aspect ratio is almost constant, independent of the pose and the appearance. Unlike that, the appearance of other objects, e.g. cars or cylists, varies significantly in different viewpoints. In (Ohn-Bar and Trivedi, 2014) Ohn-Bar solved this problem for cars by building detectors for each visual subcatgory and achieved good results. So in this project we do it in a similar way. For training we also choose the KITTI dataset (Geiger et al., 2012), which is one of the benchmarks for object detection and contains rich images of traffic scenes. Instead of using additional features in (Ohn-Bar and Trivedi, 2014), positive samples here are sorted directly according to their 3-D orientations. We do it in this way, because, not only the procedure can be simplified but also errors from false classifications can be avoided.

In this paper we divide positive samples into eight equidistant orientation groups, each with a range of 45°, as shown in Figure 1. Since sliding window detection is used, we also scale samples of each group to their rounded average aspect ratios. E.g. 0.5 is chosen for group II and VI, 1.0 for group IV and VIII and the others have an aspect ratio of 0.75. The minimal height of each sample is 80 pixels.

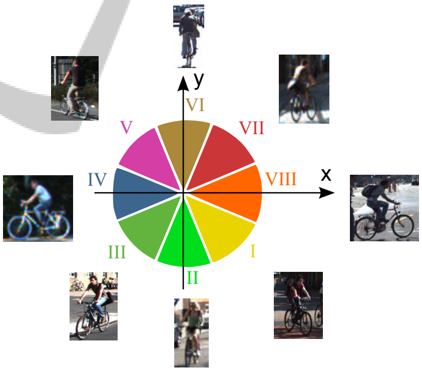

Figure 1: Division of positive samples into eight equidistant orientation groups I to VIII.

5.1.2 SVM vs. DT

Support vector machines (SVM) are often used together with HOG features. SVMs can obtain high precision but at a great time cost. On the contrary, decision trees (DT) often run at fast speed, even though the classification power is often weak in practice. So our aim is to build a detector which combines advantages of both kinds of classifiers.

Here we choose a cascade as the detector structure. Unlike (Dollár et al., 2009), we do some modifications to it. As shown in Figure 2, it consists of n stages of DTs and one SVM. The front DTs can filter out lots of negative samples very quickly, so that a fast speed is achieved. The last SVM guarantees a high precision and its influence on speed is mere. Because it is located at the last stage, only a few classifications are done there.

Since only gray images are used in this paper, i.e. color information is not available, we choose HOG as features instead of integral channels in (Dollár et al., 2009) (Benenson et al., 2012). For classification, HOG vectors are calculated for each image patch. The SVM makes use of the whole vector. Instead, DTs only deal with some specific vector elements. To select the number n, two points should be concerned. On one side, too few DT stages can increase the burden on SVM and further reduce the detection speed. On the other side, with too many DT stages, only a few negative samples can be obtained for training SVM, which makes it very easy to be overfit. According to our experience, 2 to 4 is a reasonable number.

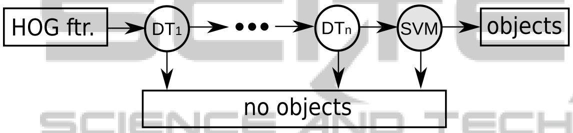

Figure 2: Structure of the cascade detector. It consists of n stages of decision trees DT_1 to DT_n and one support vector machine SVM. For each image patch a HOG feature vector is calculated. While the SVM makes use of the whole vector, the DTs only consider some specific vector elements, so that a fast detection speed is achieved.

5.2 Detection with Partial Occlusion

Up to now we focused on detecting cyclists with full bodies. As for occluded cyclists, we can only recognize them by capturing their visible parts with part detectors.

To decide which part should be detected, we take the same approach as (Felzenszwalb et al., 2008). At first, we calculate HOG features for each sample and use them to train a SVM. Then a rectangular filter with the same size as the part detector is convoluted with the weights. The maxima, regarded as with high energy, are selected as locations for part detectors.

For each group in Figure 1, we train m part detectors with the same structure as Figure 2. For efficiency, we only run part detectors on image regions without hypotheses of full body detectors. Based on results of part detectors we can estimate the presence of cyclists according to the method proposed by Shu in (Shu et al., 2012).

Here we interpret the score $s(p_i)$ of part i at location $p = (x,y)$ as

$$s(p_i) = c_{p_i} + d_{p_i} \cdot f_d(d_x, d_y), \tag{1}$$

where c_{p_i} is the convolution value of the part filter and (d_x, d_y) is the relative displacement to its anchor. $f_d(d_x, d_y) = (d_x, d_y, d_x^2, d_y^2)$ denotes the deformation and d_{p_i} is the coefficient vector, which can be learned from a latent SVM as in (Felzenszwalb et al., 2008). So the total score from part filters can be interpreted as

$$score_p = b + \sum_{i=1}^{n} s(p_i), \tag{2}$$

where b is a bias factor. But this equation is only valid for completely visible cyclists. For occlusion, only the visible parts are interesting to us. If we take an assumption that visible parts always have high part scores, the problem can be converted to search a set S_m of the most confident parts, which maximize the score. Then Equation (2) can be rewritten as

$$score_p = b + \arg\max_{S_m} \frac{1}{|S_m|} \cdot \sum_{i \in S_m} \frac{1}{1 + exp(A(p_i) \cdot s(p_i) + B(p_i))}, \tag{3}$$

where $|S_m|$ is the set cardinality, in (Shu et al., 2012) it is equal to 3. A and B are parameters, which can be learned by the sigmoid fitting method in (Platt, 1999). This equation can be considered as calculating the average score of a subset of parts. As occluded parts always have a smaller score, the score reaches its maximum only if all parts are visible in the subset.

Since cyclists are divided into 8 aspects, there are totally $8m$ part detectors. To apply them on images one after another is extremely time consuming and unnecessary. Because at one location only one part type can exist. Hence, we decide to assemble part detectors from one aspect, e.g. with the help of random forests (Breiman, 2001). Then we only have 16 detectors.

5.3 ROI Extraction from Monoscopic Image

For further improving detection performance, we would like to integrate ROI extraction to our framework. Here the ROI extraction is based on the fact that all cyclists are on the ground and that the height, the pitch and the roll angle of the camera are known. Hence, the task is to find regions in the image, which are corresponding to detected objects standing on the ground. As one solution, Sudowe et al. in (Sudowe and Leibe, 2011) introduced the mathemtical derivation from the size of an object in the real world to its location in the image.

Here we also prefer this geometric method to extract ROIs, because no further sensors are available and precise locating of object is required, which is important to predict collisions. But unlike Sudowe's approach we will reveal the relationship between the size and the location of an object in the image by a regression method.

At first we observe objects in the real world. Here we build two horizontal planes with a height $z = 0$ and $z = S_{obj}$ respectively. The first one corresponds to the ground plane and the other is a horizontal plane just above the head of the objects. As shown in Figure 3, both planes consist of grid points. For each point from one plane we associate it with the point from the other plane with the same horizontal coordinates to make one pair. So between each point pair, the distance is constant and equals object's height S_{obj}. Here we take for example $S_{obj} = 2$m.

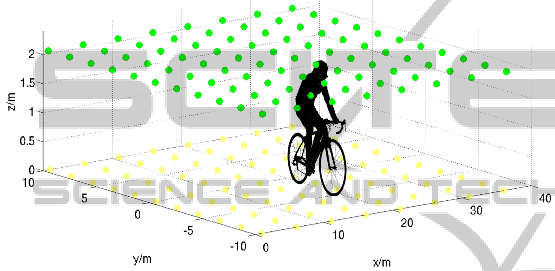

Figure 3: Two planes of grid points. The ground plane is presented in yellow with a height of $z = 0$ and the green one is the plane just above object's head with a height of $z = S_{obj} = 2$m.

Then we project the grid points from both planes into the image (Figure 4) and store their coordinates in matrix $\mathbf{U_0}$ and $\mathbf{U_{S_{obj}}}$ respectively. After that we calculate the distance between each point pair again, which corresponds to object's height in the image, and store them in vector $\mathbf{h}$. According to (Sudowe and Leibe, 2011), the size and the location of an object in the image have a nearly linear dependence,

$$\mathbf{h} = \mathbf{U_0} \cdot \mathbf{B}, \tag{4}$$

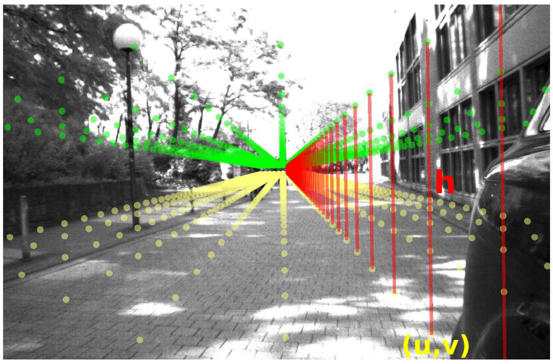

Figure 4: Calculate the distance h between each point pair and store it. Location (u, v) is the coordinates of the corresponding ground point.

where $\mathbf{B}$ is a parameter matrix and can be obtained by some regression method, such as the least squares algorithm. For simplicity, Equation (4) can be rewritten as

$$h = \begin{bmatrix} u & v & 1 \end{bmatrix} \cdot \mathbf{B} \tag{5}$$

where h is object's height with respect to the ground point $[u, v, 1]^T$ in the image.

If the roll angle of the camera is zero, the size of an object is only dependent on the vertical coordinate of its bottom v. Therefore, the ROIs for objects can be restricted to a region between two horizontal lines (Figure 5), which is in accordance with the results in (Sudowe and Leibe, 2011).

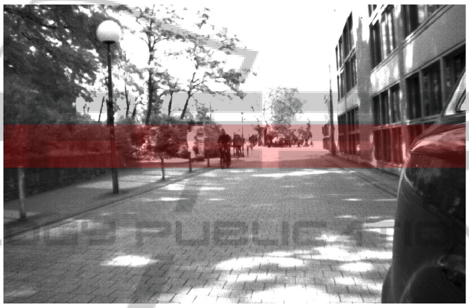

Figure 5: Example for ROI extraction by our geometric method. This image is captured in the KIT. The roll angle of the camera is zero and the detection window has a height of 80 pixels. The height of the object in the real world is assumed to be 2m. Red region represents the ROI of valid objects.

In the same way we can obtain the dependence between the size and the location of an object in scaled images, so that the geometry based ROI extraction can be extended to multi-scale detection.

5.4 Tracking of Cyclists

To predict collisions, not only the presence of cyclists but also their trajectories should be known, which requires to track cyclists along image sequences. Here we use a simple model with constant velocity. The state vector is

$$\mathbf{x} = [x, y, v_x, v_y, h, w]^T, \tag{6}$$

where (x, y) is coordinates of the bottom left corner of the cyclist, (v_x, v_y) denotes the velocity in the image, and (h, w) represents the size. The acceleration and size variations are modeled as noise. With the help of a Kalman filter (Kalman, 1960), the state of the cyclist can be predicted in the next image. To associate the predictions and the detections, we use overlapping areas between them. It means only the overlapping rate between a prediction and a detection is greater than a predefined threshold, an association will be created.

As the number of detections in an image can be different from the number of predictions, we will solve the association problem by JPDA (Joint Probabilistic Data Association) algorithm (Bojilov et al., 2003). The predicted state is updated by its associated detecion. Single detections are regarded as new objects (cyclists) and single objects without detection are directly taken into next prediction. This procedure can be seen in the right part of Figure 6.

Because detectors can fail for some images, errors accumulate for those objects in prediction steps. To avoid it, we add other trackers, such as optical flow, which consists of several feature points. As shown in the left part of Figure 6, each detected cyclist is assigned with a Kalman filter and an optical flow tracker. The feature points of the optical flow tracker are updated for each frame. If an associated detection exists, the feature points are checked and only valid ones remain. Otherwise the state of the cyclist is estimated based on the optical flow tracker, so that the tracking results can be further stabilized.

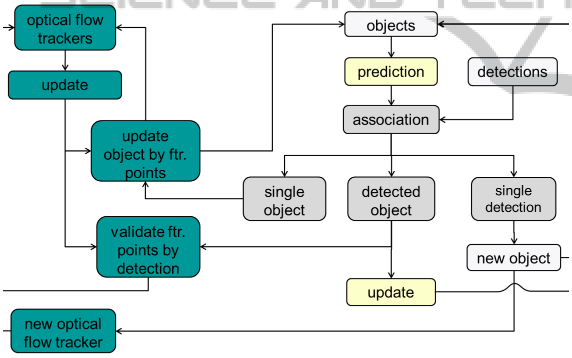

Figure 6: Tracking algorithm by Kalman filter (right side) and optical flow (left side). Each detected cyclist is assigned with a Kalman filter and an optical flow tracker. For each frame, the states of objects are predicted and the feature points of the optical flow tracker are also updated. The predicted objects are updated by their associated detecions. Single detections are regarded as new objects (cyclists) and single objects without detection are updated by their optical flow trackers, so that the tracking results can be further stabilized.

6 STATE OF THE RESEARCH

At the moment we have already built cascade detectors to detect cyclists. The geometry based ROI extraction method is also integrated. To explore the detector's performance, we have captured a video with scenes both on campus and in nearby urban areas. There are totally 45000 images, consisting of lots of objects, such as cars, pedestrians and cyclists. The images have a size of 1312 × 1082 pixels. In comparison, we also trained a cascaded DPM detector for cyclists with the codes provided by (Felzenszwalb et al., 2010). We ran both our detector and the DPM detector on the test data and the recall-precision-curves are plotted in Figure 7. Additionally, we also evaluated computational efficiency of each detector (Table 1).

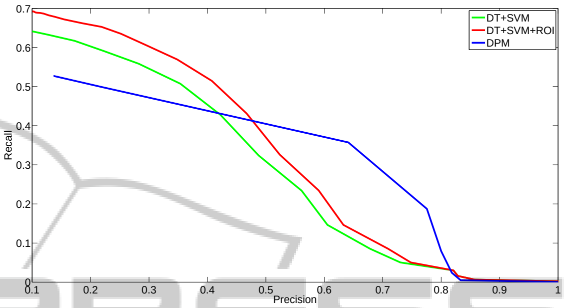

Figure 7: Recall-precision-curves for our detectors with (red) and without ROI extraction (green) and the DPM detector (blue).

Table 1: Time cost of our detectors with (DT+SVM+ROI) and without ROI extraction (DT+SVM) and the DPM detector.

	DPM	DT+SVM	DT+SVM+ROI
time(s)	2.2	0.28	0.09

It can be seen that the precision of our detector can increase at most 10% if the ROI extraction method is integrated. In comparison with DPM, our detector has a higher recall at low precision ranges but its precision becomes worse if the recall value decreases. As far as the real situation is concerned, false negatives are even worse than false positives. Because protection measurements can not be activated, if cyclists are not recognized, leading to reduced survival probabilities in accidents. As for time cost, our detector runs at a speed of almost 11 fps. In comparison, DPM takes 2.2s for one image. Obviously, our detector is more suitable for real time applications, even though it is not as good as the DPM detector in high precision areas.

7 EXPECTED OUTCOME

For a more precise impression of our detector's performance, we would like to do test on more datasets as well as to compare it with other standard detectors. So far we are only concerned about detection of completely visible objects. In the next step, we will work on detection of partially occluded objects. To stabilize detection along image sequences, the tracking algorithm also works in progress. For a precise estimation of a cyclist's trajectory, we would like to project its 2-D movement into 3-D coordinates. This can only be done with the help of additional sensors, e.g. lidar. Moreover, we also want to extend our framework for the recognization of general object classes. In the future, we would like to see our system applied on vehicles, which makes contribution to protection of cyclists.

REFERENCES

Benenson, R., Mathias, M., Timofte, R., and Van Gool, L. (2012). Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2903–2910.

Bojilov, L., Alexiev, K., and Konstantinova, P. (2003). An accelerated imm jpda algorithm for tracking multiple manoeuvring targets in clutter. In *Numerical Methods and Applications*, volume 2542, pages 274–282. Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cho, H., Rybski, P., and Zhang, W. (2010). Vision-based bicycle detection and tracking using a deformable part model and an ekf algorithm. In *13th International IEEE Conference on Intelligent Transportation Systems*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1.

Dollár, P., Tu, Z., Perona, P., and Belongie, S. (2009). Integral channel features. In *BMVC*.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2010). Discriminatively trained deformable part models, release 4.

Gandhi, T. and Trivedi, M. (2007). Pedestrian protection systems: Issues, survey, and challenges. *Intelligent Transportation Systems, IEEE Transactions on*, 8(3):413–430.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Li, T., Cao, X., and Xu, Y. (2010). An effective crossing cyclist detection on a moving vehicle. In *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, pages 368–372.

Mori, G., Belongie, S., and Malik, J. (2005). Efficient shape matching using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1832–1837.

Ohn-Bar, E. and Trivedi, M. M. (2014). Fast and robust object detection using visual subcategories. In *Computer Vision and Pattern Recognition Workshops-Mobile Vision*.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Rogers, S. and Papanikolopoulos, N. (2000). Counting bicycles using computer vision. In *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*, pages 33–38.

Sabzmeydani, P. and Mori, G. (2007). Detecting pedestrians by learning shapelet features. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Shu, G., Dehghan, A., Oreifej, O., Hand, E., and Shah, M. (2012). Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821.

Sudowe, P. and Leibe, B. (2011). Efficient use of geometric constraints for sliding-window object detection in video. In *Computer Vision Systems*, volume 6962, pages 11–20. Springer Berlin Heidelberg.

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

Walk, S., Majer, N., Schindler, K., and Schiele, B. (2010). New features and insights for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1030–1037.

Wojek, C. and Schiele, B. (2008). A performance evaluation of single and multi-feature people detection. In *Pattern Recognition*, volume 5096, pages 82–91. Springer Berlin Heidelberg.