# Multi-band Hough Forests for Detecting Humans with Reflective Safety Clothing from Mobile Machinery

Rafael Mosberger[1], Bastian Leibe[2], Henrik Andreasson[1] and Achim J. Lilienthal[1]

***Abstract*— We address the problem of human detection from heavy mobile machinery and robotic equipment operating at industrial working sites. Exploiting the fact that workers are typically obliged to wear high-visibility clothing with reflective markers, we propose a new recognition algorithm that specifically incorporates the highly discriminative features of the safety garments in the detection process. Termed Multi-band Hough Forest, our detector fuses the input from active near-infrared (NIR) and RGB color vision to learn a human appearance model that not only allows us to detect and localize industrial workers, but also to estimate their body orientation. We further propose an efficient pipeline for automated generation of training data with high-quality body part annotations that are used in training to increase detector performance. We report a thorough experimental evaluation on challenging image sequences from a real-world production environment, where persons appear in a variety of upright and non-upright body positions.**

## I. Introduction

In this work we address the problem of human detection from heavy industrial machinery and autonomous robots that operate in environments shared with humans. While recent years have seen steady progress in the development of pedestrian detectors for cars, their adaption to industrial scenarios has received limited attention. A direct transfer of the employed methods has been found difficult, mainly because pedestrian detectors strongly focus on upright persons. In an industrial setting however, humans are also frequently observed in non-upright positions, due to the large variety of tasks being carried out (cf. Fig. 1). The combination with the additional challenges of varying illumination conditions, occlusions, and background clutter, makes human detection for industrial applications a very difficult task.

To address the specific requirements for industrial human detection, Mosberger et al. [1] introduced a vision-based approach that targets industrial environments where human workforce is required to wear reflective safety clothing (cf. Fig. 2c). The method relies on the reflective properties of the safety garments, and detects humans solely by the appearance of the reflective markers (cf. Fig. 2b and 2d) in images captured by an active near-infrared (NIR) stereo camera setup. The method has been shown to yield good detection rates, while being robust to strongly varying illumination conditions.

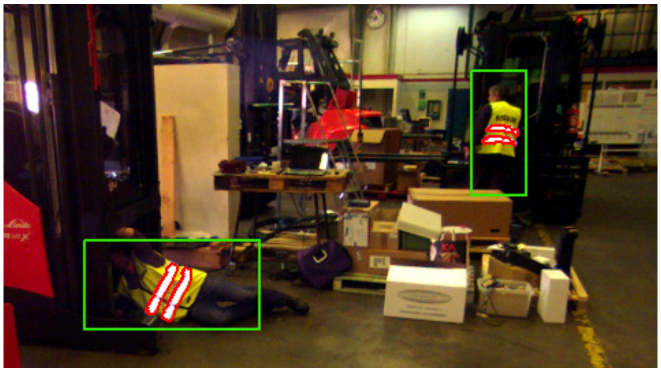


Fig. 1: Output of our multi-band Hough Forest human detector for a characteristic industrial scene, with detected reflectors marked red, and green bounding box detections showing size and aspect ratio estimated in the process.

However, by entirely focusing on the appearance of reflective markers, the approach in [1] lacks the ability of exploiting other features commonly used in vision-based human detection, such as gradient histograms [2] extracted from images taken in the visible spectrum. In consequence, the position and orientation of individual parts of the human body are not exploited and no attempt is made to estimate a full bounding box around a detected person. Yet, especially in an industrial setting, humans do not necessarily appear in upright position (cf. Fig. 1) and bounding box detections of correct size and aspect ratio are essential for various tasks, including safe navigation of autonomous or human-operated vehicles around human workforce.

We therefore propose a new recognition algorithm based on the popular Hough Forest framework, which not only detects industrial workers but also estimates position, size and aspect ratio of the bounding boxes (cf. Fig. 1). Termed *Multi-band Hough Forest*, the algorithm jointly uses multiple spectral bands by fusing the strongly contrasting natures of the NIR and RGB input images (cf. Fig. 2b–2d) in a combined appearance model.

Our article makes the following principal contributions: i) We introduce our Multi-band Hough Forest detector that uses the reflective patterns of the safety garments as well as the human appearance to detect and localize industrial workers and estimate their bounding boxes. Our proposed training procedure explicitly uses pixel-wise body part labels to increase the performance of the obtained detector with respect to the classical Hough Forest. ii) We propose a novel, efficient and automated pipeline for creating training material with high-quality body part annotations. The

[1] Rafael Mosberger, Henrik Andreasson and Achim J. Lilienthal are with the AASS Research Center, School of Science and Technology, Örebro University, S-70182 Örebro, Sweden firstname.lastname@oru.se

[2] Bastian Leibe is with Computer Vision Group, RWTH Aachen University, D-52074 Aachen, Germany leibe@vision.rwth-aachen.de

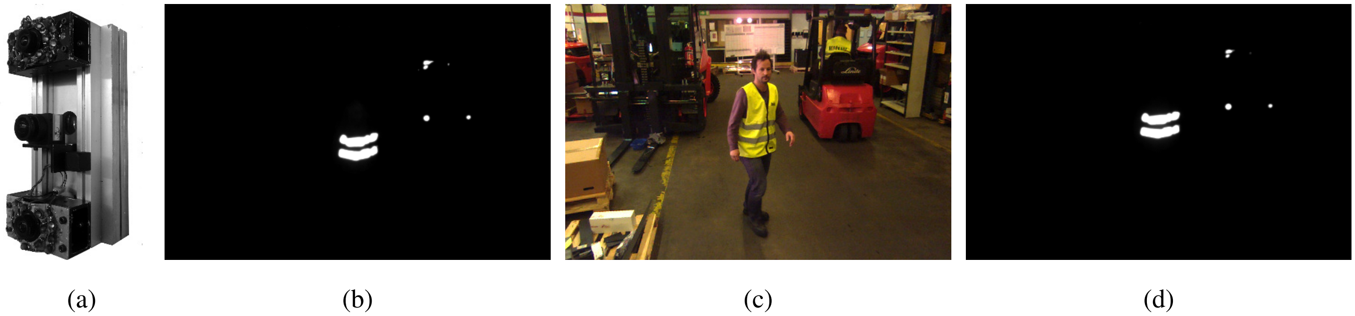


Fig. 2: Input to our multi-band Hough Forest detector for a typical industrial scene: (a) Camera setup with two active NIR camera units (top and bottom) and a standard color camera (center). (b,d) The images acquired by the two NIR cameras serve to detect reflective markers, estimate their depth, and generate regions of interest. (c) The same scene captured by the color camera. Both the NIR and the RGB images are jointly used to create an appearance model of the industrial workers.

pipeline allows us to efficiently learn the human appearance model from images recorded in a photo studio and apply it to a challenging real-world scenario. iii) We present a thorough experimental evaluation on challenging real-world sequences recorded at an industrial production site with workers appearing in upright and non-upright positions. We thereby investigate the design space of the proposed detector by illustrating the performance improvement due to each processing step of our method.

## II. Related Work

Vision-based pedestrian detection in the context of road safety has been extensively studied over recent years [3], [4], [5], and steady progress has lead to the integration of commercial pedestrian detectors into today's generation of cars. By contrast, the problem of detecting humans from machinery that operates at industrial work sites has not received the same interest. Even though the task shares many characteristics, a number of differences between urban traffic scenes and industrial workplaces make it difficult to directly apply existing pedestrian detectors in an industrial context. Specific approaches to vision-based human detection from industrial machinery include visible-light stereo vision [6], thermal imagery [7] or the combination of on-board and stationary camera modules [8].

A particular challenge which is usually not handled well are persons in non-upright positions. However, in the context of industrial safety, it is crucial to consider a large variety of potential body postures, following the number of tasks carried out by industrial workers. Yet, as existing pedestrian detectors are highly influenced by the characteristics of urban traffic scenes, they typically perform well on upright persons only. Deformable part models [9] have shown to handle higher degrees of articulation much better, but are computationally expensive. A different approach that considers the problem from an industrial perspective was proposed in [1], and is based on active NIR vision. By only observing patterns created by the reflective markers attached to the worker's safety clothing in the NIR image, the method yields good detection performance, but does not allow to perceive a person as such. It is therefore difficult to infer additional information about the body orientation.

To address this issue, we propose a system combining the robust active NIR approach for reflector detection with color vision. Our method detects and localizes persons and additionally estimates the aspect ratio of bounding boxes to obtain a notion of body orientation. However, training our detector on a variety of different body positions requires to collect and label a large amount of training material. We therefore propose a pipeline that allows to efficiently generate training material from images acquired in a well-controlled environment. Image annotations are automatically extracted and the training images are enhanced using a learned tone-mapping function. Our detector uses the popular Hough Forest framework [10]. Hough Forests are object appearance models based on the Implicit Shape Model (ISM) [11], where an object class is modeled by a large number of local prototypic patches of specific appearance and given relative location from a defined object center.

## III. Multi-band Hough Forest Detector

Our approach builds on the method described in [1] for detection of industrial workers wearing safety clothing with reflective markers. We propose to extend the camera setup by integrating a color camera and combine the advantages of both sensor types. NIR stereo vision (cf. Fig. 2b and 2d) allows us to efficiently detect reflective markers, infer their distance to the camera, and generate regions of interest (ROIs). The ROIs are then analyzed to either collect more evidence for the presence of a person, or, to discard types of reflectors not belonging to the class of interest (cf. Fig. 9). To do so, we use the popular Hough Forest framework [10], and sample dense feature patches from both the NIR and the color images, each contributing in a generalized Hough voting procedure to infer bounding box detections with the correct position, size and aspect ratio.

**Hardware Setup.** Our camera setup extends the configuration presented in [1], which uses a stereo pair of customized NIR cameras equipped with a combination of band-pass filter and flash unit. The purpose of the configuration is to detect and locate reflective markers and it provides images in which reflectors appear as characteristic high-intensity regions in front of the dark, non-reflective background (cf. Fig. 2b, 2d).

Even though a low-resolution color camera was present in the setup, its images were not used during processing. Our extended setup features a 1-megapixel color camera fixed in the center of the sensor unit. All cameras are synchronized and use the same image resolution. Fig. 2 shows the camera setup and the images obtained from it.

### *A. Training*

Apart from common features used in human appearance models, such as gradient histograms [2], we also want our detector to learn the specific properties of the reflective safety garments in use. These mainly include the characteristic color and the appearance of the reflectors as observed in the color and the NIR images. Furthermore, we do not limit our model to upright standing people and pay particular attention on covering a large variety of different body positions, including standing, sitting, lying, kneeling and crouching. As our approach targets real-world applications, it is also a central requirement that our detector can be efficiently trained and applied to different environments and scenarios. However, there exists no publicly available dataset featuring the proposed combination of NIR and color images. As it is not feasible to collect and label training material from every potential application environment, we propose an efficient pipeline that allows us to learn the appearance model from image sequences recorded in a dedicated environment in the form of a photo studio. By doing so, we strongly reduce the annotation effort in the process of preparing the training images.

**Data Collection, Annotation, and Enhancement.** With our camera setup statically placed in a photo studio, we collected a rich set of foreground training images (cf. Fig. 3a–b) containing a single person wearing a reflective safety vest. A colored background allows efficient background subtraction and extraction of bounding box annotations. We further define the object center to be the centroid of the reflective markers on the vest, which has shown to be the more stable reference point than the bounding box center. Individually colored clothing further allows to extract a pixel-level segmentation into different body parts. Fig. 3c illustrates the set of annotations automatically extracted from a positive training image. A second set of images containing various background scenes is recorded in different industrial indoor and outdoor facilities.

*Color Mapping.* As we want our model to incorporate the characteristic color of the safety vests, we also compute features from the color image. To ensure that these features do not include the characteristics specifically introduced for the automated extraction process, namely the green background and the individually colored body parts (cf. Fig. 3b), we propose to apply a color mapping function to the input images. The mapping $I \mapsto I'$ is performed in the HSV domain and selectively shifts the hue-, saturation and value components according to,

$$I'_{h,s,v}(\mathbf{x}) = I_{h,s,v}(\mathbf{x}) + b_{h,s,v}(l(\mathbf{x})) \tag{1}$$

where $l(\mathbf{x}) \in \{l_0, l_1, ..., l_N\}$ are the body part labels at location $\mathbf{x}$, and $b_{h,s,v}(l)$ are additive correction factors for the individual labels and HSV channels, learned from a small validation dataset. As shown in Fig. 3d, the transformation desaturates the background and clothing parts except the safety vest and normalizes colors of the clothes to the dark tones most frequently observed at industrial work sites.

*Feature Patch Sampling.* Training material for the Hough Forests consists of a large collection of local fixed-sized images patches (cf. Fig. 3f). Positive training images are resized to a reference scale and foreground patches are sampled from random locations within the silhouette of a person. Negative patches are randomly sampled from the background images. Every image patch $\mathcal{P}_i = (\mathcal{I}_i, c_i, \mathbf{d}_i, S_i, bb_i)$ holds an appearance descriptor $\mathcal{I}_i$, consisting of multiple feature channels calculated from both the NIR and color images (cf. Fig. 3e). A binary class label $c_i \in \{0, 1\}$ indicates whether the patch was sampled from a foreground or background training image. Foreground patches further hold an offset vector $\mathbf{d}_i$ that stores the location of the patch with respect to the object center. Finally, the patches hold a local pixel-wise segmentation mask $S_i$ with respect to the body part labels $l$, as well as the bounding box $bb_i$ of the training image from which the patch was extracted.

**Hough Forest Training.** Training our multi-band Hough Forests follows the supervised procedure proposed by Gall and Lempitsky [10]. An ensemble of randomized decision trees is trained in a recursive manner on the set of foreground and background patches. At each node the patches are split into two subsets by applying binary tests that compare the intensity difference of two pixels within a selected feature channel to a threshold value. A large number of randomized tests is evaluated at each node, and the test which minimizes the uncertainty inside the two subnodes is selected. In [10], two uncertainty measures are suggested for a given set of patches $A$: the class label uncertainty $U_{\text{class}}(A) = |A| \cdot E(\{c_i\})$ and the offset uncertainty $U_{\text{offset}}(A) = \sum_{i:c_i \neq 0} (\mathbf{d}_i - \bar{\mathbf{d}})^2$, where $E$ is the standard entropy function in terms of the class labels $c_i$, and $\bar{\mathbf{d}}$ the mean offset vector of the set. While the first measure separates foreground from background patches, the second aims at grouping patches sampled from a similar spatial location with respect to the object center. However, in cases of increased articulation of the target class, as it is the case for non-upright body positions, we have found that minimizing offset uncertainty does not necessarily group patches extracted from similar body parts. We therefore propose an additional uncertainty function $U_{\text{segmentation}}(A)$ defined as the sum of entropies calculated on a per-pixel level with respect to the annotated body part segments:

$$U_{\text{segmentation}}(A) = |A| \sum_{(x,y)} E\big(\{S_i(x, y)\}\big). \tag{2}$$

For each split, one of the three uncertainty measures is chosen with equal probability. Leaf nodes are created if either the number of patches in a set is below a threshold or if a

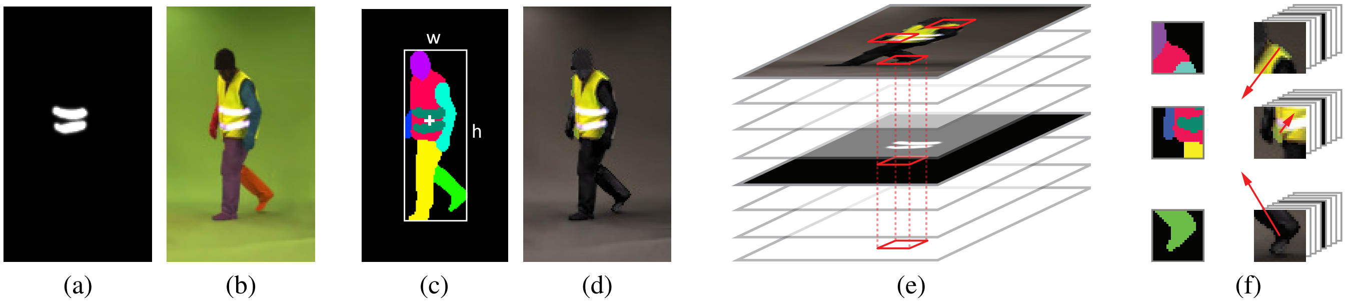


Fig. 3: Proposed automated training pipeline: (a–b) Raw NIR and color images captured in a photo studio. (c) Automatically extracted annotations, including bounding box, object centre, and body part labels. (d) Enhanced training image after applying body part wise tone mapping. (e) Stack of feature channels computed from both the color and the infrared image. (f) Local foreground training patches $\mathcal{P}_i$ sampled from the human silhouette with the according pixel-wise segmentation masks. The offset vectors (red) indicate the relative location of the patch with respect to the object center.

maximum depth is reached, and each leaf stores references to the set of patches $\mathcal{P}_i$ it contains.

### *B. Detection*

Given a synchronized input image triplet (two NIR and one RGB image), we aim to detect the center location of humans wearing a reflective safety vest and to estimate a bounding box with the correct size and aspect ratio. To do so, we densely sample feature patches from image regions surrounding detected reflective markers and apply the learned appearance model in a generalized Hough voting procedure

**Reflector Detection and ROI Generation.** Regions showing reflective markers are extracted from the pair of NIR images. In favorable circumstances, i.e. in absence of any additional NIR light source such as the sun, images resemble the example shown in Fig. 2b, and extracting reflectors is achieved by simple thresholding. However, we employ the more robust approach proposed in [1], in which each NIR image is compared to a reference image captured in short succession without flash. We then obtain a characteristic depth of the centroid of each reflector using stereo triangulation. Finally, each detected reflector defines a square ROI whose size is chosen to delimit the image region in which a hypothesized person is fully enclosed, regardless of the body posture.

**Generalized Hough Voting.** The image content delimited by the ROIs is resized to the reference scale used in training, and a set of feature channels are computed from both the NIR and color images. Patches are then densely sampled from the feature channels and are propagated down each tree of the learned Hough forest model, until a leaf node is reached. The set of training patches stored in the respective leaf node of each tree then casts votes for the location and scale of persons in a generalized Hough voting scheme as described in [10]. The votes are accumulated in a 3-dimensional Hough space, represented by a set of stacked 2D Hough images where each layer of the stack corresponds to a specific scale. The third scale dimension is included even though the depth of the reflectors is known, which allows to discriminate persons at different distances that appear spatially close in the image plane. The Hough space is subsequently smoothed with a Gaussian adapted to each scale layer. Non-maximum suppression is applied to extract object hypotheses, comprising the 2D center location in the image plane, the characteristic scale and a score. In contrast to [10], we do not include the bounding box aspect ratio as a fourth parameter in the Hough space, but instead estimate it in a second step with the portion of supporting votes of a given hypothesis.

## IV. Experimental Evaluation

We report a detailed experimental evaluation of our human detector on image sequences recorded in a real-world industrial environment, and discuss performance results with respect to multiple aspects. We adhere to the experimental protocols proposed by the Caltech pedestrian detection benchmark [4]. However, we do not apply the normalization of the bounding box aspect ratio as proposed there, becaue its estimation is a central task of our approach. Detector performance is reported in precision-recall curves, obtained by varying the threshold on the hypothesis score. Detections are considered matched to an annotation, if the intersection-over-union (IoU) overlap of the detected and annotated bounding box is larger than 0.5.

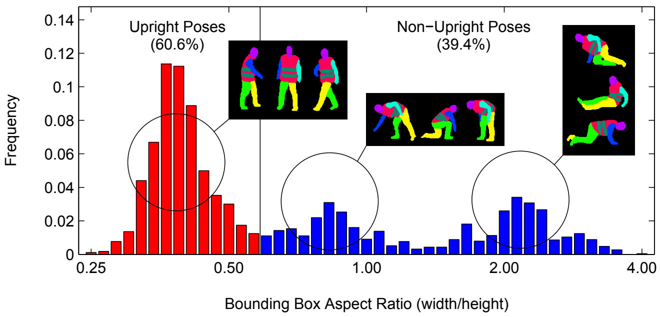


Fig. 4: Distribution of person annotations in the test sequences with respect to the bounding box aspect ratio.

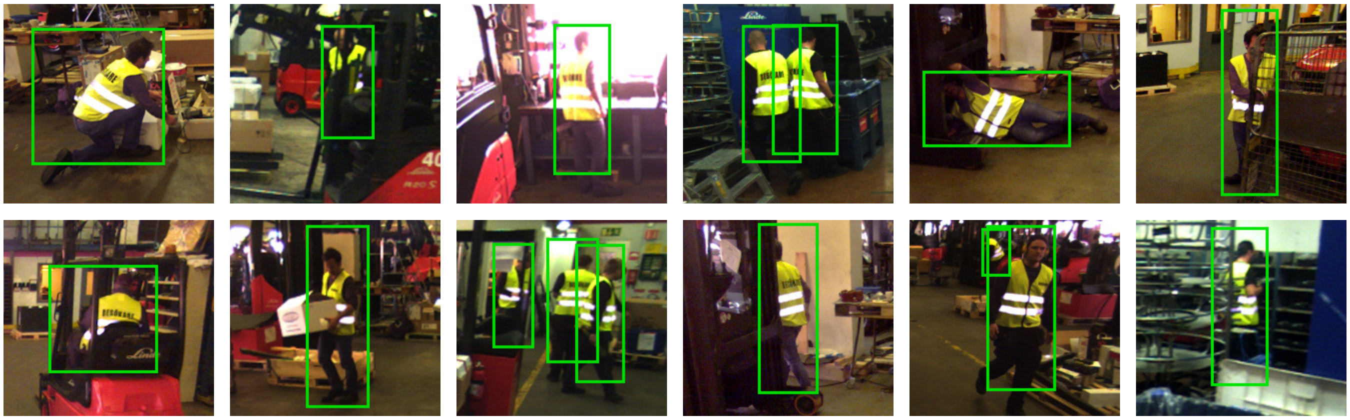


Fig. 5: Examples of successful detector responses with correctly estimated bounding box aspect ratios.

**Training.** Multi-band Hough Forests consisting of 4 trees were trained on a total of 70’000 foreground and 30’000 background patches of size 16 × 16 pixels. Every tree is trained on 50% of the patches only. At each node, 1000 binary tests are evaluated to find the best split. Foreground patches are sampled from a total of 5000 training images (cf Fig. 3) while background patches are extracted from a set of 100 images collected at several industrial work sites, including the environment in which the detector is evaluated. To learn the tone-mapping function (cf. Eq. 1) we use a manually annotated validation set of 30 frames.

**Features.** Feature channels are computed from the color image and from one of the pair of NIR images. Features include raw intensities, absolute values of the first and second x- and y- derivatives, and 8 HOG-like [2] channels. For the color image, raw intensities are represented in the *Lab* color mode. As in [10], all feature channels are further filtered by min and max filters with a window of size 5 × 5 pixels, resulting in a total of 56 channels.

**Test Data.** Evaluation is performed on 4 different test sequences recorded in a realistic production scenario at an industrial worksite, with the sensor unit mounted on a forklift truck. The environment is characterized by a high degree of background clutter and changing illumination conditions. All persons wear the same type of safety garment that was also used during training. A total of 6000 frames with 3500 manual bounding box annotations are evaluated. Fig. 4 shows the distribution of annotations with respect to the bounding box aspect ratio. Annotations with an aspect ratio larger than 0.6 are considered non-upright persons.

**Baseline.** As baseline serves the purely NIR input based method from [1]. As this system has no notion of body position, we fix the aspect ratio of the detected bounding boxes to 0.4, corresponding to the most frequently observed upright position in the test data according to Fig. 4. To illustrate the complexity of the task, we further apply two state-of-the art pedestrian detectors out-of-the box, namely a deformable part model (DPM) [9] and an implementation of the popular HOG detector [12]. Both detectors were trained on the INRIA Person dataset which does not specifically contain persons wearing reflective safety vests.

**Overall Performance.** Fig. 8a shows the performance of the baseline detectors and our best Multi-band Hough Forest based detector. For upright persons, the baseline approach [1] yields 86% of precision and recall at the equal error rate (EER), purely by observing the reflective patterns on the safety vests. However, due to the fixed upright bounding boxes it is bound to fail for non-upright persons. Our Multi-band Hough Forest detector increases the overall performance at the EER from 58% to 74%, but the additional estimation of the aspect ratio results in a slight performance reduction on upright persons (86% to 83%). The curves further illustrate that the HOG detector does not cope well with the different appearance of the persons between the INRIA Person dataset and the industrial scenario.

**Hough forest training.** Fig. 8b depicts detector performance for different combinations of supervision criteria in the Hough forest training procedure. Class and offset supervision corresponds to the combination suggested by Gall and Lempitsky [10]. As can be seen, the addition of our proposed

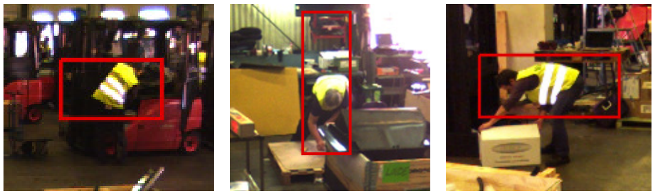


Fig. 6: Examples of detections with inaccurate estimation of the bounding box aspect ratio.

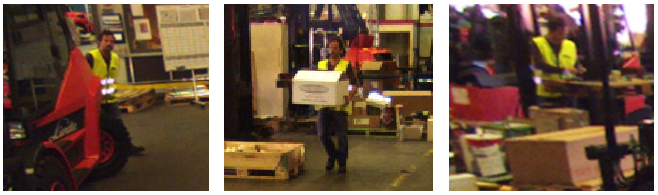
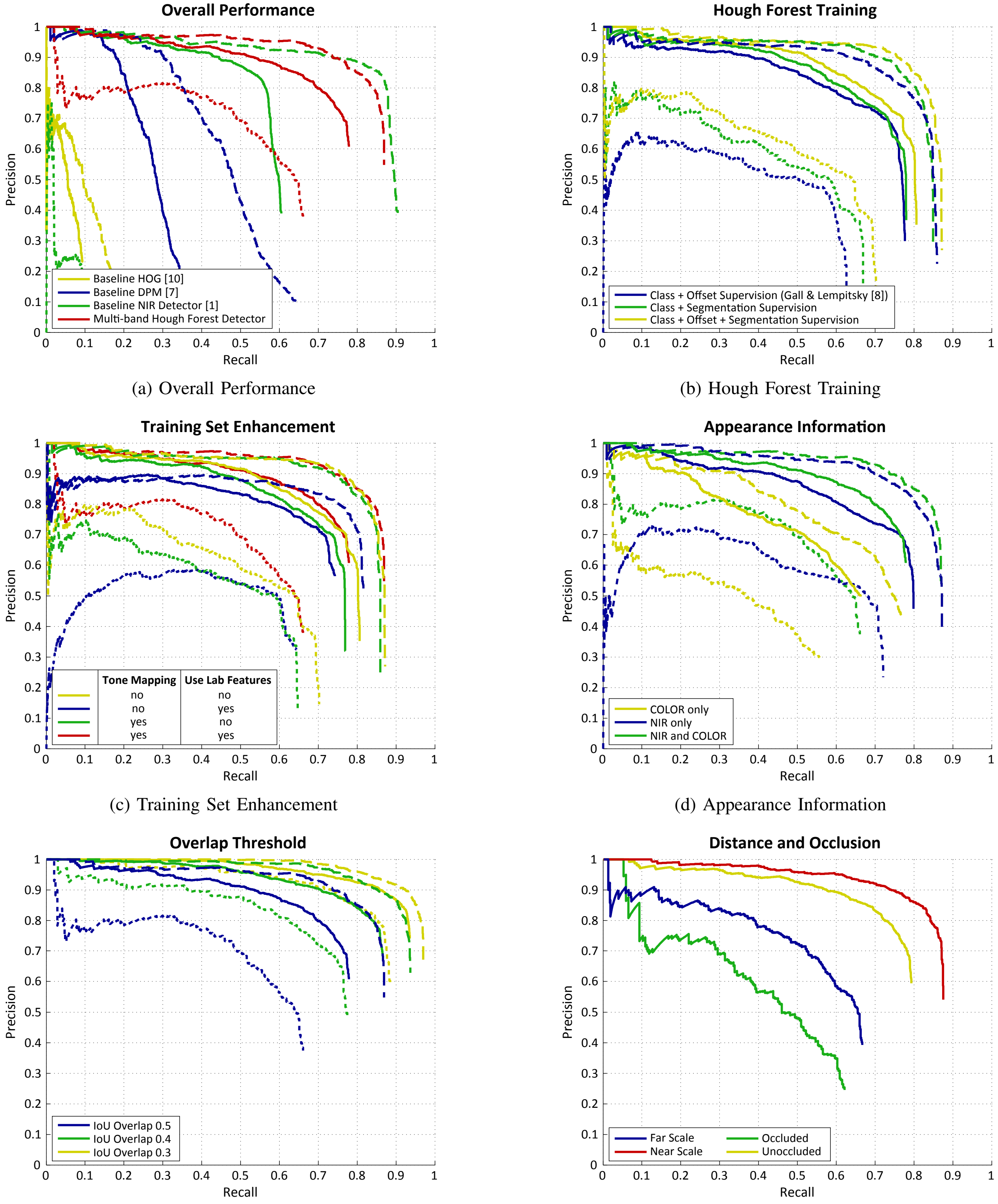


Fig. 7: Examples of missed detections due to weak detector response after occlusion of the reflective stripes.



(a) Overall Performance

(b) Hough Forest Training

(c) Training Set Enhancement

(d) Appearance Information

(e) Overlap Threshold

(f) Distance and Occlusion

Fig. 8: Quantitative detector performance in precision versus recall. Solid lines show overall performance, dashed lines performance on upright, and dotted lines on non-upright persons. (a) Performance of the baseline approach [1], two state-of-the-art pedestrian detectors (DPM [9] and HOG [12]), and the best detector obtained using our approach. (b) Influence of the different supervision criteria in the Hough forest training procedure. (c) Influence of the training set enhancement (d) Relative contribution of the color and NIR image (e) Detector performance for different overlap criteria. (f) Distance and occlusion statistics.

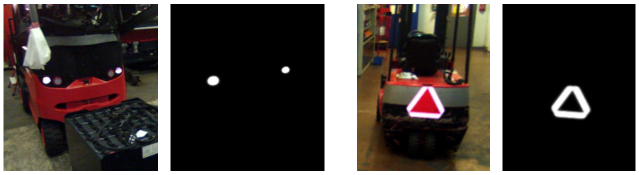


Fig. 9: Examples of reflective objects that generate regions of interest but yield weak detector response, and therefore are correctly eliminated in the process.

segmentation uncertainty criterion based on the annotated body part segments consistently improves performance.

**Training Set Enhancement.** Fig. 8c shows that best performance is obtained if the characteristic yellow color of the safety vest can be learned from the *Lab* channels of the color image and incorporated in the appearance model. However, to successfully achieve this, the proposed tone-mapping (Sec. III-A) has to be carried out before training.

**Appearance Information.** Fig. 8d shows the performance of detectors that use subsets of all available feature channels. Features computed from the NIR image are shown to contain most of the discriminative information for successful detection. Using the RGB image only shows comparably weak performance. This can be explained by the fact that the latter is much more affected by background clutter and illumination changes. It is further shown that our proposed fusion of features from the NIR and color images yields the best performance and outperforms pure NIR vision, especially in detecting non-upright persons.

**Upright versus non-upright poses.** The results show a discrepancy between performance on upright and non-upright persons. This is explained by the larger variability of possible articulations in non-upright positions, which makes the detection task much more challenging. However, Fig. 8e illustrates that by setting the overlap threshold to a less restrictive value (0.3), a much higher detection score is reached also for non-upright persons. This shows that localizing persons actually works well, only the estimation of the bounding box aspect ratio is more difficult in the non-upright case.

**Distance and Occlusion.** Fig. 8f indicates a significant performance drop for far-scale with respect to near-scale occurences, which is due to the decreasing spatial image resolution. The curves further show that our approach has a high sensitivity to occlusion. This is due to the fact that the method needs reflectors to be visible to the camera to initiate the detection process. However, the effect of the problem can be limited by using garments with more reflectors.

**Qualitative Results.** Fig. 5 shows various successful detector responses for challenging situations, while several examples of missed detections are illustrated in Fig. 7. Fig. 6 depicts examples where the object center has been correctly identified, but the bounding box aspect ratio was poorly estimated. Finally, Fig. 9 lists examples of other reflective objects present in the test area. Such secondary reflectors also generate regions of interest, but then typically yield weak detector response and are therefore filtered out.

## V. Conclusion

We presented an approach for detecting human workforce wearing reflective safety clothing from industrial machinery. By combining NIR and RGB color vision, our approach clearly outperforms a previous method based on NIR input only [1]. The proposed fusion of two spectral bands in a Multi-band Hough Forest, allows us to learn an appearance model that comprises both the properties of the safety garments in terms of reflectivity and color, as well as the typical human appearance. The obtained detector deals with a variety of body postures which is important for safety-critical industrial applications. We further illustrated the importance of our work by showing that a direct application of existing human detectors ([9], [12]) to our industrial dataset does not lead to a satisfying performance.

We further proposed a procedure to efficiently generate high-quality training data with automatically annotated body part labels and we showed that the inclusion of those body part labels into the training of Hough Forests significantly increases detection performance. This automated pipeline is an important aspect for practical applicability in industrial scenarios since it considerably reduces the training effort.

Future work includes using the proposed detection pipeline within a tracking framework, and inferring further body pose information from the specific IR and RGB sensor data.

## References

[1] R. Mosberger, H. Andreasson, and A. J. Lilienthal, “A customized vision system for tracking humans wearing reflective safety clothing from industrial vehicles and machinery,” *Sensors*, vol. 14, no. 10, pp. 17 952–17 980, 2014.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, June 2005, pp. 886–893.

[3] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detection: Survey and experiments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, pp. 743–761, 2012.

[5] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems.” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 7, pp. 1239–58, 2010.

[6] T. Heimonen and J. Heikkilä, “A human detection framework for heavy machinery,” in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2010, pp. 416–419.

[7] J. S. Dickens, M. A. van Wyk, and G. J. J., “Pedestrian detection for underground mine vehicles using thermal images,” in *Proceedings of IEEE Africon Conference*, 2011.

[8] P. V. K. Borges, R. Zlot, and A. Tews, “Integrating off-board cameras and vehicle on-board localization for pedestrian safety,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 720–730, June 2013.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] J. Gall and V. Lempitsky, “Class-specific hough forests for object detection,” in *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2009.

[11] B. Leibe, A. Leonardis, and B. Schiele, “Combined object categorization and segmentation with an implicit shape model,” in *In ECCV workshop on statistical learning in computer vision*, 2004, pp. 17–32.

[12] P. Sudowe and B. Leibe, “Efficient use of geometric constraints for sliding-window object detection in video,” in *International Conference on Computer Vision Systems (ICVS’11)*, 2011.