Evaluation of an Intelligent Collision Warning System for Forklift Truck Drivers in Industry

Armin Lang(✉)

Chair of Materials Handling, Material Flow, Logistics, Technical University of Munich, Munich, Germany
lang@fml.mw.tum.de

Abstract. The number of collisions caused by lift trucks in the area of intralogistics is still increasing despite the availability of collision avoidance systems. Commercially available products for collision avoidance mounted on forklifts often issue false alarms activated by minimum distances during daily work. Consequently, the drivers sooner or later turn those systems off. A collision warning system based on computer-vision methods combined with a time-of-flight camera delivering 2D and 3D data can overcome inflationary warnings in warehouse situations. The 3D data delivered is used to identify objects by clustering as well as to get information about the movement of objects in forklift's path. Machine-learning algorithms use the 2D data mainly to detect people in the path. Distinguishing people from non-human objects makes it possible to establish a two-level warning system able to warn earlier if humans are endangered than in collision situations in which no humans are in sight. This system's general functionality has already been proven in lab tests. To transfer the academic results to application in an industrial environment, the same test procedure has been executed during daily work in a warehouse at a company in the production sector. In this paper, the authors aim to list the differences and commonalities between the academic and industrial runs.

Keywords: Forklift safety · Warning system · Computer vision
People detection · Machine learning · Evaluation

1 Introduction

Occupational health and safety are especially important in those fields where employees cross paths with mobile machines. This is happens often in the warehouses of small and medium-sized enterprises, where industrial trucks or similar vehicles follow the same routes as walking employees do. Since forklifts are repeatedly involved in accidents, products have been developed to try to prevent accidents with humans as well as with storage equipment. They all share distance as the criterion for deciding whether or not a collision is impending. That leads to many warnings in a lift-truck driver's day-to-day routine, because forklifts often travel within small distances of storage equipment such as racks. Due to this problem, methods of computer vision will be investigated in the "PräVISION" project for use in a collision

V. G. Duffy (Ed.): DHM 2018, LNCS 10917, pp. 610–622, 2018.
https://doi.org/10.1007/978-3-319-91397-1_50

warning system for industrial trucks. The main aim is to simultaneously reduce the incidence of false warnings and retain warnings of true collisions. To accomplish this, distance will no longer be the sole indicator of an impending collision but also the movement of objects in the forklift's path. This makes it possible for the collision warning to be issued as soon as a collision is inevitable by calculating the time-to-collision for each object. In addition, a distinction between human and non-human objects enables a two-level warning system. If a non-human object is endangered, a minimum-time-to-collision threshold is used. The threshold is increased by a security factor if a human is involved. Further information about the system and computer-vision methods employed can be found in the paper by *Lang* and *Günthner* [1]. This paper also reveals the reliability of person detection in academic test scenarios.

The system's general functionality has already been proven, but its industrial application has not yet been examined. That's why we wanted to record the daily routine of forklift trucks in a warehouse at a production company in Germany. By evaluating these recordings, we analyze the ability to transfer the results of previous academic tests into industrial application.

2 State of the Art

Several assistant systems are available that try to prevent accidents caused by forklifts. Those systems use technologies such as radar, ultrasonic waves, radio, or laser to capture the environment [2–5]. They all decide whether or not a collision is impending based on a minimum distance being fallen below. When that happens, a sound alerts the driver to stop. Because lift trucks work in tight areas, those systems issue warns very often, resulting in their being turned off. New technologies in the sector of depth cameras and computer vision allow the development of more intelligent warning systems. The most important methods required for such a computer-vision-based collision warning system will be introduced below.

2.1 3D Cameras

The combination of ordinary color images with depth information is no novelty. This has be done for years, but the color-image data always had to be calibrated onto the depth data. The calibration can be very complex and expensive if special calibration patterns are required. Newer 3D cameras such as stereo or time-of-flight cameras no longer require this. They're already calibrated at delivery. Stereo cameras supply an RGB image and a matching depth image. Sometimes, the camera doesn't calculate the depth image itself, but it can be done easily with a suitable framework. For time-of-flight cameras, the depth image can always be obtained directly from the camera. Depth information is calculated by measuring the time between the emission and the reception of infrared light waves. Obviously, the camera also takes infrared images through the infrared sensor, which is almost independent from environmental illumination. This type of image always matches the depth image. In addition, some cameras such as the "Microsoft Kinect One®" also supply color images, but unlike the infrared ones these

have to be calibrated when using the Kinect. Color-image calibration is unnecessary for industrial time-of-flight cameras.

2.2 Movement Calculation with Computer Vision

An important part of predicting impending collisions is predicting the movement of objects in a forklift's path. That can be done either by clustering objects within the 3D camera's field of view and tracing their movements or by using the methods of optical flow. The latter exploit distinctive features in a 2D picture to calculate the movement of pixels in consecutive images [6]. By adding the depth data, the so-called scene flow can be calculated. Scene flow represents the real movement of all pixels within the camera's field of view in world coordinates [7].

2.3 Object Classification and Localization with Computer Vision

There are a few methods for classifying objects in 2D or 3D data. The past was shaped by defining the geometrical patterns of different kinds of objects to be able to detect those objects. More recent approaches use machine learning algorithms, which are able to identify objects without given geometrical patterns. Those algorithms can be trained by feeding them suitable training data. In regard to computer vision, this data contains pictures with and without the objects to be classified.

The theory of machine learning had already been postulated in the early 20th century. One of the first algorithms, the support vector machine (SVM) postulated in 1963, first became applicable in 1992 [8, 9]. This algorithm belongs to the category of deformable-part models and is able to classify data with predefined features. In combination with the sliding-windows method, localization of objects is also possible. The most common feature used to detect people is the "histogram of oriented gradients [10]."

The latest developments in hardware have also made neural networks suitable for object classification and localization. Those also have to be trained first, but the features don't have to be defined. Neural networks find the important features by themselves.

In their paper "Pedestrian Detection: An Evaluation of the State of the Art," *Dollar et al.* introduced a method for evaluating people-detection algorithms. This became a standard reference in the field of person detection. The "miss rate" (mr) and "false positives per image" (fppi) are the target values. The miss rate is the percentage of unrecognized persons; the fppi indicator describes how many false detections are made per image [11]. Those two indicators depend on the threshold used to accept a detection proposal as a human or not. By using various thresholds, a chart as seen in Fig. 1 can be created.

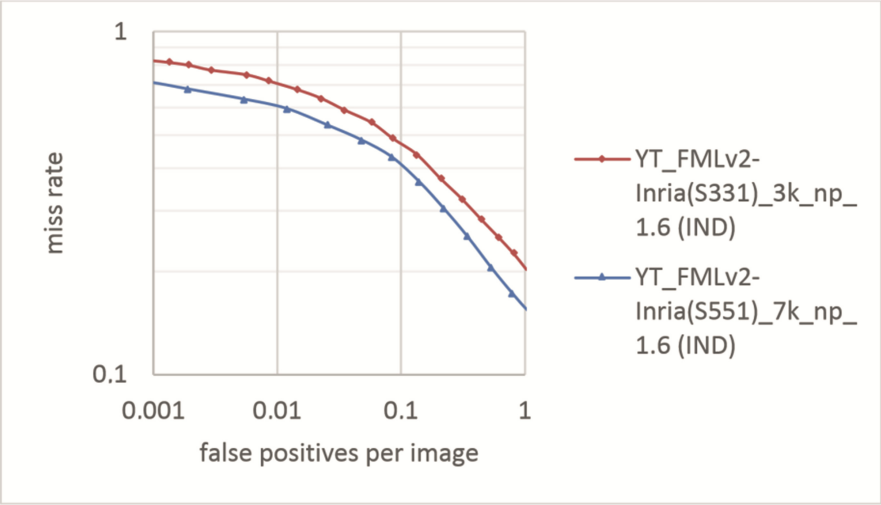

Fig. 1. Example chart of the evaluation of two people detectors according to Dollar et al.

Each marker on a line represents the miss rate and the false positives per image at a specific threshold. The nearer the line is to the origin of coordinates, the better the detector used. Consequently, detector 'YT_FMLv2-Inria(S551)_8k_np_1.6 (IND) is better than the other one, because the former's line is nearer to the origin of coordinates.

3 Software Framework

The developed system uses only open source libraries and its own methods. The main library is OpenCV 3. This library contains a huge number of image processing methods and, among other things, implementations of various machine learning algorithms, too [12].

Figure 2 shows an overview about the developed methodology. The first step is to fetch 2D and 3D images from a camera. Those images will be preprocessed and segmented. Preprocessing includes all of the methods that are necessary to prepare the images for the following collision and person detection. These include, for example, scaling, color conversion, or filling empty pixels. The last is needed mainly for the depth image, because due to the sensor technology there's no depth value for all pixels.

After some preprocessing has been done, irrelevant parts of the images will be removed in the "segmentation" module. The most important part for clustering reasons is removal of the ground floor in the depth image, because otherwise all pixels would be connected through the ground floor. The sequences of preprocessing and segmentation methods vary depending on the following algorithm and performance.

Collision detection uses a scene-flow method [7] to calculate the relative movement of objects in the forklift's path. This data is used to get each pixel's time-to-collision with a virtual box surrounding the forklift.

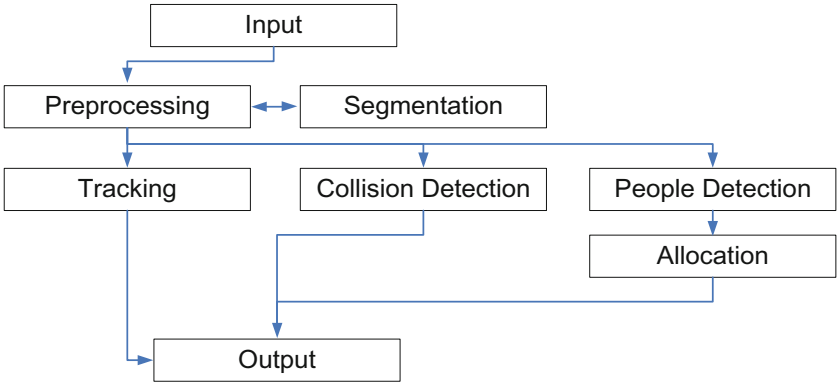

Fig. 2. Methodology of the collision warning system

Machine learning algorithms are used for person detection. The intent is to use both shallow learning methods like the support vector machine and deep learning networks. The SVM is implemented because it needs very little computing power, so it can be used in real-time. Deep learning is better at detection, but execution can't be done in real-time on a mid-sized computer. So detection based on deep learning is being reviewed for future application when graphic-card performance has increased.

When a person is found, a history is created or updated in the "allocation" module. If at least one person has been found before, various tracking methods [13–16] are provided to relocate him or her in the current frame. Tracking should increase detection performance, because no person detector can detect everyone.

The last step is output to the driver. The information from the tracking, collision, and person detection modules is collected and evaluated. The number of pixels falling below either the maximum time-to-collision for humans or the maximum time-to-collision for objects will be counted. The pixel count will be normalized according to the pixels' real distance in meters. A warning will be issued if a certain amount is exceeded, whereby different sounds are used for endangered humans or objects.

4 Approach and Frame Conditions of the Industrial Evaluation

The developed system can in principal be used on most kinds of forklifts, but some up-front work has to be done before making the recordings. Frame conditions in the industrial test area, such as the types of lift trucks available, first have to be determined. Data recording should be prepared based on previous recordings in the laboratory. Having done the preparation, the system is started-up on-site and run over several days. Afterwards, we extracted interesting video scenes from the dataset, because evaluating the whole dataset entails considerable difficulty. Persons have to be tagged manually on each frame to control the results of person detection.

The following evaluation focuses on person-detection performance. General collision detection has already been validated, so it's not repeated.

The same components are used in the industrial evaluation as were in the academic one. The only difference is the forklift used and some mechanical additions for mounting the system on the lift truck.

4.1 System Setting

The following equipment was used for the demonstrator:

- Forklift: STILL FMX-14 (type: high-reach truck)
- Notebook: Tuxedo XC1707 (i7-6820HK, 16 GB RAM, NVIDIA GTX 1070)
- Camera: Microsoft Kinect One (mount height: 3.4 m to bottom)
- Power converter: DC/AC 400 W 48 V/220 V

The developed demonstration system is generally applicable on just about every kind of industrial truck or other vehicle, but the required equipment (camera, notebook) has to be mounted and powered. A converter was needed to power the notebook and the camera. A box containing notebook, power converter, and camera was constructed for mounting the system on the forklift (see Fig. 3).

Fig. 3. Mounting the system on the forklift

4.2 Test Area

The records for the industrial trial were taken over three days during the daily routine of a lift-truck driver in the warehouse of a white goods manufacturer. The forklift on

which the system was mounted operated within two areas (see Fig. 4). The lift truck loads its goods in the first area (1), where all incoming goods are temporarily stored on pallets. Different kinds of vehicles such as hand pallet trucks, roll containers, electric stacker trucks, and reach trucks as well as pedestrians cross ways there. The destination is the mobile pallet racks (mpr) (2). The other mobile pallet racks shown in Fig. 4 are supplied only rarely. The maximum speed allowed is at 8 km/h for every kind of vehicle.

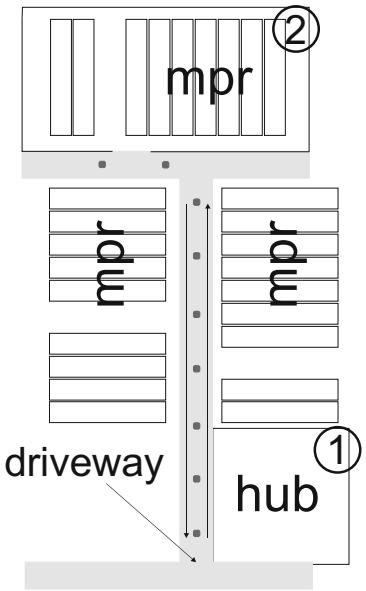

Fig. 4. Schema of the storage area

5 Statistical Aspects of the Data

The recorded data is analyzed below to enable the results to be compared with other person-detection evaluations.

Three shifts were recorded representing over 2 million frames per image type. Forty-four sequences containing persons totaling 17,385 frames overall in the whole set were chosen for evaluation. In these scenes, all of the persons standing or walking were tagged manually (see Table 1). 14,933 tags arose this way for the infrared images. In 84% of the selected frames, the camera angle was at 30°, in 7% at 35°, and in 9% at 40°, whereby the number of degrees is for the angle between the horizon and the camera axis. The reason for the difference in the distribution is that the bigger the angle the less area, and so the fewer persons are within the camera's field of view.

Table 1. Summary of tag data recorded in the industrial environment

Image type	Frames	Frames with tags	Frames with multiple tags	Tags	30% occluded tags	50% occluded tags	70% occluded tags
Color	17,385	10,763	2,671	14,349	2,761	1,823	1,014
IR	17,385	11,186	2,805	14,933	2,798	1,830	1,027

There are fewer tags for the calibrated color images because the depth sensor only captures pixels of objects that are at most 8 m away, so some persons are visible in the infrared images but not in the depth image. That's also true for the calibrated color image, because it's generated from the color and depth image. About 64% of the infrared frames contain at least one person, 16% of all infrared frames contain more than one tag. In nearly 19% of the tags, about 30% of the person is occluded, for example by boxes or shelves. 12% of the tags are almost half occluded and 7% are occluded more than 70%. Persons are consequently fully visible in only about 60% of the tags.

The position distribution of the tags can be seen in Table 2. The tags are divided into the maximum distance from the top respectively from the left of the image. A tag whose center is 40% of the image width (512 pixels) to the left and 40% of the image height (424 pixels) to the top of the image, counts as "Left 50% | Top 50%". The horizontal distribution is quite normal, but the vertical one is very imbalanced to the top. The reason for this is that humans are most likely far away from the forklift during daily work. They are thus found more often in the images' upper regions.

Table 2. Distribution of tag positions

	Left 0–30%	Left 30–50%	Left 50–70%	Left 70–100%	Σ
Top 0–30%	18.15%	15.15%	14.14%	12.03%	59.47%
Top 30–50%	8.49%	7.09%	6.62%	5.63%	27.83%
Top 50–70%	2.94%	2.45%	2.29%	1.95%	9.63%
Top 70–100%	0.93%	0.78%	0.73%	0.62%	3.06%
Σ	30.52%	25.47%	23.78%	20.23%	200%

The size distribution of the tags in the images can be obtained from Table 3. The real dimensions of the tagged persons were also analyzed (see Fig. 5). The sizes were determined automatically by creating clusters of tags in the depth image and transforming the data into the world coordinate system. Some false clusters caused the data show not just the dimensions of whole persons, but also the dimensions of some parts of them. That's especially important for the height, because some clusters only reach from the ground to the hip. In addition, the floor was removed for clustering purposes. This is done by removing all pixels which are closer than 0.2 m to the ground floor. Consequently, 0.2 m has to be added to all height values. The highest peaks can be found at widths from 0.4 m to 0.6 m and at heights from 1.4 m to 1.6 m (raw data, 0.2 m has to be added).

Table 3. Tag size distribution in pixels

	Minimum	Maximum	Average	Median	Deviation
Width [px]	8.00	266.00	65.84	57.00	34.21
Height [px]	16.00	306.00	127.87	117.00	57.93

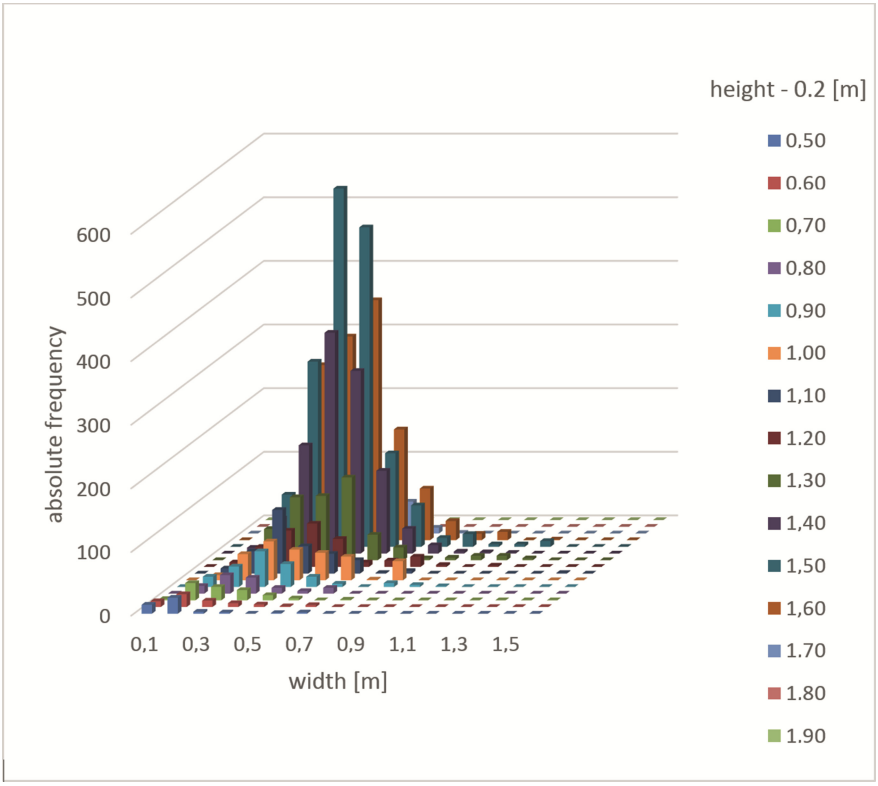

Fig. 5. Real-person dimensions of the tags

6 Evaluation of Person Detection

The industrial run was intended to validate the hypothesis that person detection trained with data from an academic environment and with reference to pedestrian databases also works in an industrial environment. 114 different support machines have already been trained and evaluated on test data that was recorded in the experimental laboratory of the Chair for Materials Handling, Material Flow, Logistics. The different SVMs' training data varied. Videos from the laboratory or warehouse videos from Youtube were taken as negative data—meaning pictures lacking humans. For positive data containing only humans, we used pictures from reference pedestrian databases such as INRIA [10] or the 'Monocular Pedestrian Detection' database [17] as well as artificially generated

pictures of humans. The evaluation was done according to *Dollar et al.* as described in Sect. 2.3.

For better presentation, only the results of the three best-detecting support vector machines evaluated in the academic test data are shown in Fig. 6.

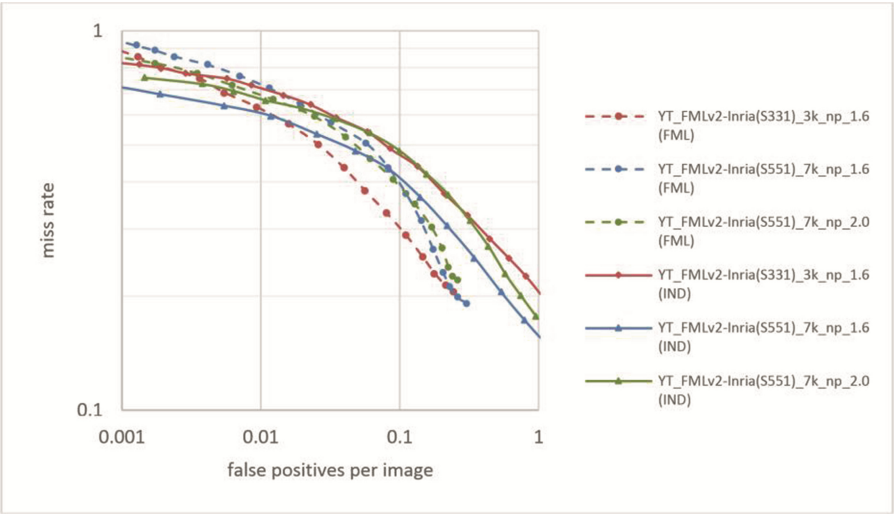

Fig. 6. Evaluation results for academic (FML) and industrial (IND) test data, whereby calibrated color images were used for person detection

Table 4. Comparison of the best (academic environment) three support vector machines

Image type	SVM	Avg. miss rate (FML)	Avg. miss rate (IND)	Relative change
Color	YT_FMLv2-Inria(S551)_7k_np_1.4	0.286194	0.521047	+82.06%
Color	YT_FMLv2-Inria_7k_np_1.8	0.306845	0.487052	+58.73%
Color	YT_FMLv2-Inria_3k_np_2.0	0.335521	0.525198	+56.53%
Infrared	YT_FMLv2-Inria(S551)_7k_np_1.8	0.187238	0.295612	+57.88%
Infrared	YT_FMLv2-Inria_7k_np_2.0	0.189408	0.341561	+80.33%
Infrared	YT_FMLv2-Inria_3k_np_1.4	0.191005	0.303023	+58.65%

The figure shows – in this case for calibrated color images – that detection is worse when the trained support vector machines have to detect people in an industrial environment (solid lines). There is also a discrepancy in the infrared image with regard to detection performance. It is not always worse, but the miss rate's variance is significantly high.

The two scenarios' average miss rates are compared in Table 4. The average miss rate is calculated by integrating the miss rate over the range of 0.01 to 0.1 false positives per image. The miss rate approaches 1 (100%) at fppi less than 0.01, so that no one will be detected any longer. At fppi greater than 0.1, classification generally ceases to come under discussion, because there are too many wrong detections [11].

Some SVMs performed better in the industrial evaluation than in application at the laboratory. The maximum increase is 11.71% more detections when using color images and 23.32% when infrared images are used (see Table 5). On average, the miss rate was about a quarter higher in the industrial test data than in the academic test data, hence about 25% fewer people were detected.

Table 5. Evaluation over all SVMs

Image type	Average miss-rate change	Standard deviation	Minimum MR change	Maximum MR change
Color	24.25%	22.33%	−11.71%	129.78%
Infrared	26.18%	18.42%	−23.32%	98.94%

As shown in Table 6, the best SVM in the industrial evaluation is still 59% (respectively 54%) worse than the best SVM in the academic evaluation.

Table 6. Comparison of the absolute best (lowest) average miss rates

Image type	Best avg. academic miss rate	Best avg. industrial miss rate	MR(IND)/MR(FML)
Color	0.286194	0.455677	159%
IR	0.187238	0.288959	154%

7 Discussion

The resulting assumption is that training is very application sensitive and thus has to be adapted to the use case. The SVM has never been trained with positive data from the videos taken in the laboratory so as not to complicate the results. Unlike the positives, the negative data was always trained with videos from the laboratory due to the lack of negative data from the warehouses. Only some SVMs were trained with a mix of the previously mentioned Youtube videos and videos from the laboratory. An example of this is the "YT_FMLv2-Inria_7k_np_2.0" SVM, which was trained with a mix of both negative data sources and positives from the INRIA data set.

So not adding negative data from the application area to the training set could be one reason for the decrease in the detection rate. Another reason could probably be the different quota of occluded tags. In the industrial data set, at least 30% of the person was

occluded in about 38% of the tags as against 17% in academic recordings. In addition, there are many more people in non-upright postures such as bowing in the industrial set.

Person detection with the SVM applied works best if persons are standing and are fully visible, because those positions were trained before. The combination of more occluded and non-upright persons therefore could be the reason for the decrease in the detection rate. The tag types were not differentiated enough during evaluation at this point, so this can't be validated as a reason yet.

8 Summary

A new collision-warning system for lift trucks is currently being developed in the "PräVISION" project. The system uses a time-of-flight camera to capture information about the environment. Computer-vision algorithms use this data to predict impending collisions, whereby collisions with humans and objects are distinguished. The distinction is made using machine learning algorithms for detecting people in different image types. This system has already been evaluated in academic settings. Industrial evaluation of the support vector machine used for people detection showed that the results of academic tests cannot be transferred to industrial application. The detection rate decreased about 25% on average. Either the lack of training data from the application area or the divergent distribution of the tags is responsible for the decrease. Consequently, the next step will be to find out why detection is significantly worse when trained SVMs are applied in an industrial environment.

Acknowledgement. The "PräVISION" project is funded by the German Social Accident Insurance (DGUV). The project partners involved are:

– BIBA Bremen
– SICK AG
– STILL GmbH
– Berufsgenossenschaft Handel und Warendistribution (BGHW)

References

1. Lang, A., Günthner, W.A.: Evaluation of the usage of support vector machines for people detection for a collision warning system on a forklift. In: Nah, F.F.-H., Tan, C.-H. (eds.) HCIBGO 2017. LNCS, vol. 10293, pp. 322–337. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58481-2_25
2. Acure: Blaxtair. https://blaxtair.com. Accessed 20 Feb 2018
3. ELOKON GmbH: ELOprotect; ELOshield; ELOback^2. http://www.elokon.com. Accessed 20 Feb 2018
4. tbm hightech control GmbH: RRW-207/3D; RAM-107; RRW-107plus. http://www.tbm.biz. Accessed 20 Feb 2018
5. U-Tech GmbH: U-Tech. http://www.u-tech-gmbh.de. Accessed 20 Feb 2018
6. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)

7. Jaimez, M., Souiai, M., Gonzalez-Jimenez, J., Cremers, D.: A primal-dual framework for real-time dense RGB-D scene flow. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2015)
8. Vapnik, V., Lerner, A.: Pattern recognition using generalized portrait method. Autom. Remote Control **24**, 774–780 (1963)
9. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT 1992, New York, NY, USA, pp. 144–152 (1992)
10. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_33
11. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
12. OpenCV. https://opencv.org/. Accessed 20 Feb 2018
13. Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1090–1097 (2014)
14. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: automatic detection of tracking failures. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 2756–2759. IEEE (2010)
15. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2012)
16. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: TV-L1 optical flow estimation. Image Proc. On Line **3**, 137–150 (2013)
17. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)