Robust Pallet Detection for Automated Logistics Operations

Robert Varga and Sergiu Nedevschi
Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Keywords: Object Detection, Pallet Detection, Stereo Reconstruction.

Abstract: A pallet detection system is presented which is designed for automated forklifts for logistics operations. The system performs stereo reconstruction and pallets are detected using a sliding window approach. In this paper we propose a candidate generation method and we introduce feature descriptors for grayscale images that are tailored to the current task. The features are designed to be invariant to certain types of illumination changes and are called normalized pair differences because of the formula involved in their calculation. Experimental results validate our approach on extensive real world data.

1 INTRODUCTION

Automated Guided Vehicles perform (AGVs) logistics operations without human intervention. This requires the existence of a sensor capable of estimating the position of the pallet that needs to be loaded by the machine. This work focuses on developing a machine vision-based detection system for pallets.

Pallets are wooden supports designed to hold goods and are easily graspable by the forklift because of its pockets. Pallets are standardized and for our purposes they are handled from only one side. We desire a flexible detection module that can identify the relative position of the pallet from any image under various lighting conditions.

Stereo cameras offer a good solution for 3D sensing applications. The cost of such systems is lower compared to laser scanners. Also camera systems offer a full 3D view as opposed to 2D scan lines and the possibility of high level reasoning on data. The main drawback of such systems is the difficulty of working in poor and rapidly changing illumination conditions.

We have studied previous vision-based attempts at this problem and found that they are lacking because of the following reasons: they rely on features that do not possess good invariance properties; detection performance is poor in general and especially in dark regions; most systems are not thoroughly evaluated.

For the above mentioned reasons we propose improvements which constitute the main contributions of this work:

- Original candidate generation method that enables fast detection by quickly rejecting certain regions;
- The proposal of new grayscale features invariant to certain types of illumination changes.

The paper is organized as follows: Section 2 presents existing approaches and important contributions from the image processing literature: edge detection; feature vector extraction; classification. In Section 3 we describe our proposed system and give details about each processing step. Section 4 shows experimental results that validate our system. Section 5 concludes the paper.

2 RELATED WORK

2.1 Sensor Types

The specific topic of load handling is not a well-researched area. Approaches for autonomous load handling use different types of sensors to obtain an understanding about the environment. In (Weichert et al., 2013) the authors discuss the advantages of several sensors for this task. We will group these approaches into two main categories based on the sensors used: range sensors and vision-based sensors (monocular or stereo cameras). In the following we describe relevant approaches from each category.

Some available systems rely on laser scanner data. In most cases the sensor provides data along a 2D scanline. Using laser has the advantage over cameras that it is able to operate in complete darkness and it is not affected by lighting conditions.

In (Walter et al., 2010) a detection system is pre-

Varga, R. and Nedevschi, S.
Robust Pallet Detection for Automated Logistics Operations.
DOI: 10.5220/0005674704700477
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 470-477
ISBN: 978-989-758-175-5

sented for the autonomous manipulation of a robotic lift truck. The authors use closest edge detection applied on the sensor point cloud. SICK industries manufacture laser scanners for multiple purposes. A work from (Bostelman et al., 2006) presents a pallet detection method using such sensors. A solution is provided for unloading operations inside trucks. The walls of the trucks are detected by applying a Hough transform (Hough, 1962), (Duda and Hart, 1972). The paper (Katsoulas and Kosmopoulos, 2001) uses laser sensors to detect the positions of boxes of standard dimensions. Kinect sensors can be employed for distance estimation as in (Oh et al., 2013). However, they are not suitable for an industrial environment and they offer a small field of view.

A hybrid approach from (Baglivo et al., 2011) combines two types sensors: a laser scanner and a camera. A fusion is performed at object level between the detection from the color image and the points from the laser. Edge template matching with distance transform is applied on the color image. Both sensors must agree on the detection, ensuring robustness. The system requires the calibration of the two sensors. The authors have evaluated their system on 300 examples with results indicating a good localization precision. They have found difficulties due to lighting conditions in 5 cases.

Vision-based approaches employ multiple cues: in (Kim et al., 2001) line-based model matching is used; (Pages et al., 2011) performs colour-based segmentation; (Seelinger and Yoder, 2006) uses easily identifiable features (landmarks, fiducials); (Cucchiara et al., 2000) employ corner features, region growing and decision tree; in (Byun and Kim, 2008) least squares model fitting is applied. Most authors perform evaluation on a small dataset or in laboratory conditions. The work (Seelinger and Yoder, 2006) presents results on 100 operations with a success rate of 98%. Also, their approach requires the installation of landmarks on each pallet.

A paper from (Varga and Nedevschi, 2014) presents a detection approach relying on integral channel features. The authors evaluate their system on an impressive dataset containing 8000 test images. Other approaches include: (Nygårds et al., 2000), (Prasse et al., 2011), (Pradalier et al., 2008).

2.2 Detection Approaches

Sliding window object detection is one of the most commonly used approaches employed in the technical literature. Typical examples of particular detectors include face detectors (Viola and Jones, 2001), (Yang et al., 2002), pedestrian detectors (Dollár et al., 2012), (Benenson et al., 2014), (Dollár et al., 2014). The success of this general approach can be attributed to the fact that it uses a powerful classifier to discern between background and target object. Since the classifier is a cascade it eliminates zones without objects quickly.

Features for detection should capture structure, texture and color if possible. Some of the more important features that are relevant for this work are: any edge feature defined on the image gradient (Mikolajczyk et al., 2003); Histogram of Oriented Gradients (Dalal and Triggs, 2005) - developed originally for pedestrian detection; Haar features (Viola and Jones, 2001); integral channel features (Dollar et al., 2009); CENSUS features (Zabih and Woodfill, 1994); Local Binary Patterns and their histograms (Ojala et al., 1994), (Ojala et al., 1996).

Fast and accurate detection is possible with boosted classifiers (Schapire, 1990) and soft cascades (Bourdev and Brandt, 2005). This was first proposed by Viola & Jones for face detection in (Viola et al., 2005) but since has been adopted to pedestrian detection (Dollár et al., 2010). Many top performing methods on benchmarks utilize such classifiers for their speed.

3 PROPOSED APPROACH

Our proposed solution relies on exploiting two main sources of visual information: intensity images and stereo cameras. The intensity image provides information about 2D localization of the pallets. The stereo cameras are used to obtain the 3D position and orientation of the pallet relative to the cameras. We have found 3D-based detection less reliable because of poor reconstruction quality at pallet pockets.

Although our pallet detector is an application of the standard sliding window technique our system has to generate bounding boxes that are tight and precise. The requirements regarding exact localization are strict. Pallets need to be localized with a precision of 1 cm. This explains why experimenting and developing specific features are required. Also, the detection method should be highly accurate.

In the following we first present the processing steps required for detection. Stereo reconstruction is described at a glance. Next, we provide details about the candidate generation module. Afterwards, we introduce descriptive features proposed specifically for pallet detection. We have proposed several validation steps at the post processing stage for more robustness and enhanced localization.

3.1 Stereo Reconstruction

Reconstruction is performed with semi-global matching and CENSUS local descriptors. Our system makes use of the rSGM implementation (Spangenberg et al., 2014). CENSUS/LBP descriptors have been found to be a reliable local descriptor for many practical applications including those from the automotive industry. Semi-global matching (Hirschmuller, 2005) offers the advantage of having smooth disparity maps and it is fast enough for our purposes. The rSGM implementation is fast and runs on CPU. It includes optimizations with SSE instructions and it is a top performing method on stereo benchmarks.

3.2 Edge and Line Detection

For improving edge detection quality we rely on extracting normalized gradient values. This has been proposed and employed in calculating HOG (Dalal and Triggs, 2005) features and also in modern pedestrian detection algorithms (Dollár et al., 2010). Normalized gradient values are obtained by box-filtering the gradient magnitude and dividing the original gradient magnitude and other channels by the filtered values. This ensures successful edge detection even in dark regions.

In the following we provide the exact steps for calculating the normalized gradient maps. The gradient components along the x and y axes are obtained in a standard way by convolution with Sobel filters:

$$G_x = I * S_x \tag{1}$$

$$G_y = I * S_y \tag{2}$$

The gradient magnitude is defined as the L_1 norm of the two components:

$$M = |G_x| + |G_y| \tag{3}$$

The box filtered magnitude will act as a normalization factor:

$$\hat{M} = M * B \tag{4}$$

where B is a square box-filter of dimension w x w. Typical values for w are odd numbers from the interval $[5, 25]$. It is important to note that this filtering can be performed in $O(1)$ time per pixel for any filter size w. Filtering with a Gaussian would increase the computation with no significant benefit. The normalized magnitude and the normalized gradient components are obtained by dividing the original values with the box filtered gradient magnitude (pixel by pixel):

$$\overline{M} = M/(\hat{M} + \varepsilon) \tag{5}$$

$$\overline{G_x} = \lambda \cdot G_x/(\hat{M} + \varepsilon) \tag{6}$$

$$\overline{G_y} = \lambda \cdot G_y/(\hat{M} + \varepsilon) \tag{7}$$

All division and summation operations in the previous definitions are carried out element by element. The small constant $\varepsilon = 5e - 3$ avoids division by zero. The multiplier λ is required for converting the normalized values into the $[0, 255]$ interval.

Intuitively this operation produces strong responses where the relative change in intensity is large compared to the average intensity change in the neighboring region. This improves edge detection in poorly illuminated regions.

3.3 Candidate Generation

Considering all possible positions for sliding window detection results in a large number of possible candidates (see experimental results sections for typical numbers). It is not feasible to classify each possible candidate to see whether or not it is a pallet. This is why it is important to have a good candidate generation module. The main characteristics should be:

- high coverage - the module should not miss any real pallet positions (i.e. low number false negatives, high recall);
- fast to evaluate - can be executed instantly in comparison to following modules;
- high rejection rate - it should accept only a limited number of candidates to speed up, help and validate further processing steps.

Currently we are working with two main approaches for candidate generation. These improve the baseline approach which is just to take every possible rectangle at valid positions and scales. Edge-based candidate generation relies on edge detection while the other alternative uses stereo information. We provide details in the following.

3.3.1 Edge-based Candidate Generation

Since the frontal view of pallets is a rectangle the candidate generator should produce a list of candidate rectangles. For this we first employ the normalized gradient in the y direction as in eq. 7 to detect important horizontal lines called horizontal guide lines. A histogram that accumulates gradient values along each line is used to find local maxima. In other words we perform a projection along the horizontal direction. Since the structure of the image usually contains strong horizontal lines this step is robust and we can rely on the extracted guidelines later on.

Vertical lines are detected only between guideline pairs that respect the dimension constraints. These lines are detected where the sum of gradient along x

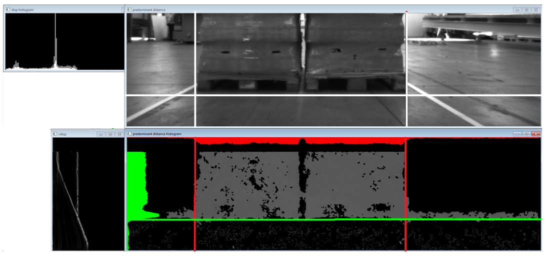

Figure 1: Stereo-based candidate generation; top-left - disparity histogram; top-right - original image with reduced region of interest marked; bottom-left - v-disparity map; bottom-right - disparity map with only the neighborhood of principal disparity highlighted, also the projections along the two axis are visualized and the new boundaries detected.

direction exceeds a certain percentage (10 %). The resulting candidate rectangles arise from combining vertical edges that fit the dimension constraints regarding width, height and aspect ratio.

3.3.2 Disparity-based Candidate Generation

We can limit the region of interest for processing by considering only the objects with fronto-parallel surfaces. The reason for this is that the axis of the stereo system is roughly perpendicular to the target pallets. Such objects appear as a line in the v-disparity and u-disparity map. Also, they lie on the disparity plane with high appearance frequency. We define the principal disparity as the disparity value that corresponds to the highest local maximum from the disparity histogram. The highest local maximum is considered because this corresponds to the obstacle in front of the camera. We call principal disparity plane the plane obtained by selecting only points that are close to the principal disparity. This is equivalent to highlighting only the objects that are closest from the visual scene.

Once the principal disparity value is determined the region of interest can be limited to the zone where such disparity values are frequent. We do this by starting from the extremities (left, right and bottom) and shrink the boundary of the original region of interest until the frequency of the preponderant disparity exceeds a limit (see Figure 1). Principal disparity also gives us information about the approximate and expected dimensions of the pallets in the image plane. This also reduces the number of possible candidates. We apply normal edge-based candidate generation on the reduced region of interest and apply the new constraints found regarding the size of the pallet.

3.4 Feature Extraction

The principal characteristic features of pallets are their structure. It is therefore important to have features that capture the structure of the pallet. Previous work used integral features defined on manual rectangular subregions, edge features, Hough transform and corner features. We have experimented with other features for two reasons: to capture the structure of the pallet in a concise way and to ensure a representation that is more invariant to illumination changes.

3.4.1 Proposed Grayscale Features - *Normalized Pair Differences*

Our goal was to introduce a grayscale feature that is sufficiently descriptive and also invariant to illumination changes. A simple way to model illumination change is multiplication by a constant value. Technically, this represents a gain change, but it is a good approximation. The features should be unaffected by this kind of operation. Weber's law states that "just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli" (Ross and Murray, 1996). Features therefore should be defined as ratios to capture relative change. This idea was employed before in other descriptors such as WLD (Chen et al., 2010), however here we propose a different form.

We use this principal to calculate our features. An option would be to normalize features by dividing with the mean of the surrounding region. However, we do not want the surrounding region to affect the descriptor of the pallet. Instead we want and invariant representation that will be the same for the same pallet. This observation leads to the necessity of defining features using only the intensity values inside the bounding box.

First, the bounding box is resized to a fixed size (5 x 20). This reduces the pallet to a smaller number of intensity values and also amounts to a low pass filtering. It is necessary to remove the regions corresponding to pallet pockets. These regions are not part of the object and bear no relevance to the detection task. Second, we take each possible pair of intensity values. The sample intensity values are denoted f_i and are obtained from the previous downsampling operation. See Figure 2 for illustration of the defined concepts.

We denote these features as normalized pair differences (**npd**). Feature values are calculated by considering all pairs, taking the difference and dividing by the first value from each pair. A sigmoid-type function is applied afterwards. Intensity features where the mask is 0 are not used:

$$D_k = tan^{-1}\left(\frac{f_i - f_j}{f_i + \varepsilon}\right) \qquad (8)$$

The role of the inverse tangent function is to limit

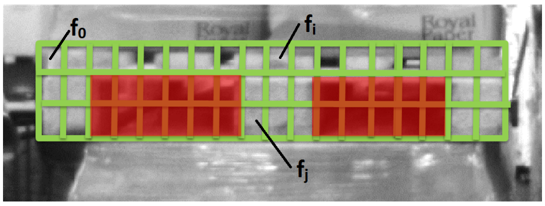

Figure 2: Feature grid of 3 x 16 overlaid on a pallet. After resizing each square cell will become one single intensity value. The cells from the red region are not used (mask = 0).

the range of the features, i.e. it is used as a sigmoid-type function. All possible pairs taken from valid positions form a signature that describes the pallet. Adding a small number $\varepsilon = 1e-2$ to the denominator avoids checking for zero division and simplifies the code for the algorithm. It is easy to see that if all intensity values are uniformly multiplied with a value α, signifying a change in illumination, the value of the descriptor does not change.

This signature will be compared by the classifier at detection time. The signature should remain roughly the same even after illumination changes. We use a rectangle grid of dimension 5 x 20. The dimension of this type of feature vector is 1350 (some pairs are missing from the $\binom{100}{2} = 4950$ because we exclude the zones from the pockets).

3.4.2 Edge Features

We also define edge features on rectangular areas near the pallet boundary in order to help in precise localization. The edge features are calculated on the normalized gradient channel. The descriptors are defined in equation 9 as normalized sums of the normalized gradient values calculated on rectangular areas depicted in Figure 3. The upper edges of the pockets are not used since they can be covered by plastic hanging from the palletized goods. The dimension of this type of feature vector is 9.

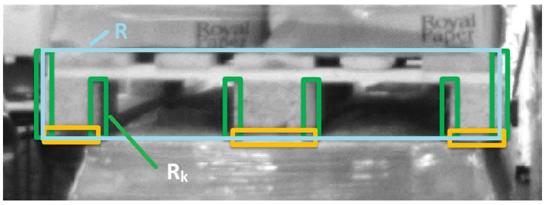

Figure 3: Support regions for calculating normalized gradient sums.

$$E_k = \frac{1}{area(R_k)} \sum_{(x,y)\in R_k} \overline{M}(x,y) \tag{9}$$

3.4.3 LBP Histogram

For texture features we use a histogram of local binary patterns. This has shown to be a reliable texture descriptor and it is also employed in many stereo matching systems. LBP descriptors also possess good illumination invariant properties since only the relative order of intensity values are important. The dimension of this type of feature vector is 256.

$$H_k = \frac{1}{area(R)} \sum_{(x,y)\in R} [mask(x,y) = 1][lbp(x,y) = k] \tag{10}$$

The last definition uses the standard Iverson bracket notation: $[expr]$ is 1 if the logical expression *expr* is true and 0 otherwise. The histogram is simply the count of each type of LBP feature that is in a valid position. The counts are normalized appropriately with the area of the bounding box R(see Figure 3). The area refers to only the zone from the rectangle that is not invalidated by the mask.

3.5 Classification and Detection

Boosted decision trees offer both high classification accuracy and fast prediction time. Since prediction is made by comparing individual features against threshold the time taken does not depend on the dimension of the feature vector. Since we know beforehand the number of desired pallets we can keep only the pallets with the highest confidence values.

The classifier is trained using the positive examples available from the manual annotations. Negative samples are generated automatically from each training image from regions that surely do not contain any pallets. Retraining the classifier with hard negatives has proven not to be helpful.

3.6 Refinement and Validation

We have found it best to enable the detector to return matches that are not precisely localized and then refine their position and scale. Bounding boxes that have good aspect ratio will have their scores improved by a multiplicative factor of 2. In cases where 2 pallets are required to be detected we boost the scores of each candidate pair that lies on the same y position and have approximately the same size.

The standard non-maximum suppression that is applied to every overlapping bounding box pair is slightly modified. In case of an overlap only the candidate with the higher score is retained. Two bounding boxes are considered to overlap if the overlap along the x axis is larger than 10 % and if the overlap along the y axis is larger than 0. A small overlap between detected bounding boxes along the x axis is possible when the pallets are far away and close to each other.

Since we have knowledge about the number of pallets that are required to be detected we can return only the most confident detections. Final pallet position is reconstructed from the plane fitted on the rectangular bounding box that is detected. This also provides us the orientation of the object.

4 EXPERIMENTAL RESULTS

All processing steps have been implemented in C++. The project uses OpenCV library for low-level image processing functions such as the bilateral filter, box filter, image reading/writing.

4.1 Feature Properties

We run tests to evaluate the invariance properties of the features we use. A sequence containing 317 measurements is recorded of a static pallet with varying exposure time. The change in exposure time modifies the appearance of the pallet from barely visible to saturated white. Descriptors are extracted from the same region. We evaluate the mean and the maximum of the standard deviations of each component. Also, the Euclidean distance is calculated between each descriptor pair and the mean and the maximum is found. We divide by the feature dimension for a fair comparison. All feature values are normalized to be in the range [-1, 1]. Table 1 shows the results, entries are ordered from top to bottom from least invariant to most invariant (we show only values for differences). The **npd** features have similar properties as the **lbp** histogram but they are more descriptive and structure information is maintained. These features change less under the tested conditions compared to the intensity and edge features.

Table 1: Measuring exposure invariance properties of different descriptor types.

Feature	dim.	mean diff.	max diff.
intensity	53	3.78e-02	1.09e-01
edge	53	2.44e-02	4.23e-02
npd	1327	2.74e-03	6.11e-03
lbp	256	2.93e-04	8.43e-04

4.2 Pallet Detection Accuracy

For evaluation purposes we use the same dataset and the same criteria as the work from (Varga and Nedevschi, 2014). The dataset was acquired from a real warehouse and was manually labeled. The detector is trained on a subset of the whole dataset. This part does not overlap with the test set on which we perform all evaluation. Two test sets are available: test set 1 which is somewhat similar to the training set having been acquired in the same recording session, also this contains the most annotated pallets; and test set 2 originating from a separate recording session. The second test set is more challenging and contains mostly difficult cases. The composition of the sets is as follows: training set contains 467 images and 891 labeled pallets (there can be zero or more than one pallet in each image); test set 1 contains 7122 images and 9047 labeled pallets; test set 2 contains 224 images and 356 labeled pallets. The final model installed in the system on the AGV was trained on all the available data.

The values of some of the parameters are given in the following. Region of interest dimensions: 400 x 1440; Bilateral filter sigma in the coordinate space $\sigma_x = 5$; Gradient box filter dimension $w = 15$; Gradient multiplier $\lambda = 40$; Horizontal edge detection non-maximum suppression neighborhood size $h = 3$; Vertical edge detection non-maximum suppression neighborhood size $v = 3$.

Since all scores depend on determining whether or not two rectangles overlap sufficiently we state precisely what we consider as an overlap. Usually for object detection intersection over union (PASCAL VOC criteria) is used to determine overlap. Here, we define the absolute positioning error along the x axis E_x as the difference between the union and overlap of the intervals along the x axis of the two rectangles. E_y, The absolute positioning error along the y axis is defined analogously. We consider an overlap a **precise** match if $E_x \leq 15$ and $E_y \leq 15$; and a **normal** match if $E_x \leq 50$ and $E_y \leq 50$. Our overlap measures are more strict than the relative overlap of the pascal VOC measure because of the system requirements. E_x is approximately equals twice the positioning error in pixels. A precise position amounts to an error of 7.5 pixels ≈ 1.5 cm using our hardware setup.

Candidate generation algorithms are evaluated by checking if every bounding box defined in the ground truth is provided by the module. The percentage of recalled bounding boxes is defined as the coverage. A box is recalled if it overlaps sufficiently with the ground-truth box. We have considered an absolute overlap when the absolute positioning error is less than 15 px along both axis. The results with different methods on the training dataset is presented in Table 2. Even though we do not achieve full coverage, rectangles near the ground truth are obtained. Using post processing and corrections the localization precision of the detection can be improved.

Table 2: Comparison of different candidate generation schemes. The approach from the last row offers an acceptable coverage while drastically reducing the number of candidates generated per image. The numbers in the parentheses indicate the step size in horizontal and vertical direction and the filter size (where applicable).

Method	**Coverage**	**Avg. nr. candidates**
All(5,5)	100 %	1370k
Grid(7,7)	99.40 %	508k
Edge(5,3)	99.52 %	374k
Normalized gradient (3,3,15)	98.81 %	35k

We now turn to evaluating pallet detection accuracy. Table 3 shows the detection accuracy on the two test sets using different configurations. The effect of adding new feature types is evaluated. We present test results using a boosted classifier with 100 and 1000 weak learners respectively. The number of negatives signifies per image is set in accordance with the power of the classifier. The training set can contain more than 1 million examples. If we weigh the error on positive instances more by ω times we can obtain a more precise localization. The npd-linear feature performs worse on the harder test set 2. Clear improvements can be seen with the new features and each additional feature improves the detection accuracy. Missed detections arise when the images are too dark, when the pallets are not fully visible or when false detections appear due to glare from the plastic covering the palletized goods.

The typical running times for the processing modules are: rectification and disparity map generation 60 ms; candidate generation 20 ms; feature extraction 800 ms; classification 300 ms. All these operations are performed on the region of interest of size 400 x 1440 = 0.576 Mpixels. Training the classifier with approximately 1 million examples and the feature vector of dimension 1591 takes a couple of hours.

5 CONCLUSIONS

The purpose of this work was to present a pallet detection method. We have improved on existing results by designing and implementing a better candidate generation module and providing better features. Detection accuracy was evaluated on a large test set and compared to an existing approach. Our system performed much better in every category.

Table 3: Detection accuracy in percentages for multiple model configurations; evaluation on both test sets; normal localization and precise localization is considered. For comparison we include the integral features from (Varga and Nedevschi, 2014) (code is provided by the authors).

	test set 1		**test set 2**	
Features	**normal**	**precise**	**normal**	**precise**
100 weak learners + 100 negatives/image				
integral ftrs.	79.0	64.2	-	-
npd	80.6	65.1	80.9	40.1
npd+edge+lbp	97.1	90.2	87.7	46.0
npd+edge+lbp + $\omega = 10$	97.7	92.6	87.7	70.5
1000 weak learners + 1000 negatives/image				
integral ftrs.	92.0	75.4	77.0	38.0
npd+edge+lbp	100	94.9	93.5	65.7
npd+edge+lbp + $\omega = 2$	98.9	95.4	91.9	68.8

We have learned that normalized gradient values enable a more robust edge detection and permit us to generate a small set of candidates. More descriptive features result in higher detection accuracy.

Future work will involve optimizing the execution time of the feature extraction module because it currently dominates the pipeline. Increasing the localization precision with post-processing steps is also of interest.

ACKNOWLEDGEMENTS

PAN-Robots is funded by the European Commission under the 7th Framework Programme Grant Agreement no. 314193. The partners of the consortium thank the European Commission for supporting the work of this project. We would like to thank our project and Elettric80 for the opportunity to perform tests at the warehouse in Viano, Italy.

REFERENCES

Baglivo, L., Biasi, N., Biral, F., Bellomo, N., Bertolazzi, E., Lio, M. D., and Cecco, M. D. (2011). Autonomous pallet localization and picking for industrial forklifts: a robust range and look method. *Measurement Science and Technology*, 22(8):085502.

Benenson, R., Omran, M., Hosang, J., and Schiele, B. (2014). Ten years of pedestrian detection, what have we learned? In *ECCV-CVRSUAD*. IEEE.

Bostelman, R., Hong, T., and Chang, T. (2006). Visualization of pallets. In *SPIE Optics East*.

Bourdev, L. and Brandt, J. (2005). Robust object detection via soft cascade. In *CVPR*, pages II: 236–243.

Byun, S. and Kim, M. (2008). Real-time positioning and orienting of pallets based on monocular vision. In *ICTAI (2)*, pages 505–508. IEEE Computer Society.

Chen, J., Shan, S., He, C., Zhao, G., Pietikäinen, M., Chen, X., and Gao, W. (2010). Wld: A robust local image descriptor. *IEEE Trans. Pattern Anal. Mach. Intell*, 32(9):1705–1720.

Cucchiara, R., Piccardi, M., and Prati, A. (2000). Focus based feature extraction for pallets recognition. In *BMVC*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, pages I: 886–893.

Dollár, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast feature pyramids for object detection. *PAMI*.

Dollár, P., Belongie, S., and Perona, P. (2010). The fastest pedestrian detector in the west. In *BMVC*, pages 1–11. British Machine Vision Association.

Dollar, P., Tu, Z. W., 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. *IEEE Trans. Pattern Anal. Mach. Intell*, 34(4):743–761.

Duda, R. and Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *CACM*, 15:11–15.

Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR*, pages II: 807–814.

Hough, P. V. C. (1962). A method and means for recognizing complex patterns. U.S. Patent No. 3,069,654.

Katsoulas, D. and Kosmopoulos, D. I. (2001). An efficient depalletizing system based on 2d range imagery. In *IEEE International Conference on Robotics and Automation, 2001. Proceedings 2001 ICRA.*, volume 1, pages 305–312. IEEE.

Kim, W., Helmick, D., and Kelly, A. (2001). Model based object pose refinement for terrestrial and space autonomy. In *International Symposium on Artificial Intelligence, Robotics, and Automation in Space, Montreal, Quebec, Canada*.

Mikolajczyk, K., Zisserman, A., and Schmid, C. (2003). Shape recognition with edge-based features. In *BMVC*.

Nygårds, J., Högström, T., and Wernersson, Å. (2000). Docking to pallets with feedback from a sheet-of-light range camera. In *IROS*, pages 1853–1859. IEEE.

Oh, J.-Y., Choi, H.-S., Jung, S.-H., Kim, H.-S., and Shin, H.-Y. (2013). An experimental study of pallet recognition system using kinect camera. In *Advanced Science and Technology Letters Vol.42 (Mobile and Wireless 2013)*, pages 167–170.

Ojala, T., Pietikainen, M., and Harwood, D. (1994). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *ICPR*, pages A:582–585.

Ojala, T., Pietikainen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59.

Pages, J., Armangue, X., Salvi, J., Freixenet, J., and Marti, J. (2011). Computer vision system for autonomous forklift vehicles in industrial environments. *The 9th. Mediterranean Conference on Control and Automation*.

Pradalier, C., Tews, A., and Roberts, J. M. (2008). Vision-based operations of a large industrial vehicle: Autonomous hot metal carrier. *J. Field Robotics*, 25(4-5):243–267.

Prasse, C., Skibinski, S., Weichert, F., Stenzel, J., Müller, H., and Hompel, M. T. (2011). Concept of automated load detection for de-palletizing using depth images and RFID data. *International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 249–254.

Ross, H. and Murray, D. J. (1996). *E.H.Weber on the tactile senses 2nd ed.* Hove: Erlbaum (UK) Taylor and Francis.

Schapire, R. (1990). The strength of weak learnability. *MACHLEARN: Machine Learning*, 5.

Seelinger, M. J. and Yoder, J.-D. (2006). Automatic visual guidance of a forklift engaging a pallet. *Robotics and Autonomous Systems*, 54(12):1026–1038.

Spangenberg, R., Langner, T., Adfeldt, S., and Rojas, R. (2014). Large scale semi-global matching on the CPU. In *Intelligent Vehicles Symposium*, pages 195–201. IEEE.

Varga, R. and Nedevschi, S. (2014). Vision-based automatic load handling for automated guided vehicles. In *Intelligent Computer Communication and Processing*, pages 239–245. IEEE.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proc. CVPR*, 1:511–518.

Viola, P. A., Platt, J. C., and Zhang, C. (2005). Multiple instance boosting for object detection. In *NIPS*.

Walter, M. R., Karaman, S., Frazzoli, E., and Teller, S. J. (2010). Closed-loop pallet manipulation in unstructured environments. In *IROS*, pages 5119–5126. IEEE.

Weichert, F., Skibinski, S., Stenzel, J., Prasse, C., Kamagaew, A., Rudak, B., and ten Hompel, M. (2013). Automated detection of euro pallet loads by interpreting PMD camera depth images. *Logistics Research*, 6(2-3):99–118.

Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell*, 24(1):34–58.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages B:151–158.