Improved Autonomous Load Handling with Stereo Cameras

VARGA Robert, Arthur Costea, NEDEVSCHI Sergiu
Technical University of Cluj Napoca
{robert.varga, arthur.costea, sergiu.nedevschi}@cs.utcluj.ro

***Abstract*—We present newly added modules for our autonomous load handling system. The stereo camera system provides information about the scene in front of the automated forklift. A new alternative module for pallet detection is described. Several processing modules for unloading operations are also presented. Our system is evaluated by means of detection rate and by performing field tests. The tests show that it is capable of providing a sufficiently accurate position of the pallets in order to perform loading and unloading operations in multiple scenarios.**

I. Introduction

In this paper we present stereo-vision-based solutions for autonomous load handling operations. We expand on previous work by adding new modules for several unloading tasks. Autonomous Guided Vehicles (AGVs) must perform pallet loading and unloading operations without manual intervention. Since the localization of the AGVs inside the warehouse is limited in precision, slight missalignments might appear when approaching the pallet.

The role of the vision system is to identify the pallet or the empty space designated for unloading in order to help correct the approach of the AGV. A stereo camera system can estimate the relative distance of objects from the common field of view. The system must be able to provide the position of the pallet with an accuracy of 1 cm and the orientation of the pallet with 1 degree accuracy.

The work presented here is part of the PAN-Robots project [1] whose aim is to create an automated logistics environment. Installation and maintenance of such an environment is costly and time consuming. Thus, one of the main goals of the project is to ensure this with low installation time and costs.

The theoretical and practical contributions of this paper include:

- the introduction of a new alternative pallet detection module with aggregate features;
- the extension of our previous detection module with new features and classifier types;
- development of solutions for unloading operations;
- introduction of validation modules;
- qualitative test results from executing real operations.

II. Previous work

Approaches for autonomous load handling use different types of sensors for obtaining an understanding about the environment. We will group these approaches into two main categories: vision-based (monocular or stereo cameras) and 2D or 3D time-of-flight Laser Range Finder (LRF). We start by describing the approaches from the second category.

[2] presents a method for autonomous manipulation of a priori unknown palletized cargo with a robotic lift truck. The sensor involved in detection is a horizontal LIDAR. Operating on the noisy points from the sensor an algorithm is applied to perform closest edge detection.

Sky-Trax System offers a solution to detect the presence of pallets on the forklift. The system uses an ultrasonic sensor that uses sound waves to detect objects in its range. It has a broad sensing area, is compact, durable, accurate, and inexpensive. Pepplerl-Fuchs also offers ultrasonic sensors for solving the same task.

SICK industries manufacture laser scanners for multiple purposes. A paper from the National Institute of Standards and Technology [3] presents a pallet detection method based on LADAR (laser detection and ranging) using SICK S3000 laser scanners. This has the advantage over cameras that it is able to operate in complete darkness and invariant lighting. They also tackle unloading operation for trucks using the same sensor. Hough transform [4], [5] is applied to detect lines that represent walls and other boundaries from the gathered laser points.

The system designed in [6] combines two sensors, a laser scanner and a camera, to localize the pallet given some prior knowledge with large uncertainty. The pallet is detected from the color image acquired from the camera and the points from the laser. The vision part uses edge template matching and distance transform. Both sources must agree on the detection in order for it to be considered valid. This approach suffers from the disadvantages: the calibration between the laser scanner and camera; edge information is not reliable; laser scanner only offers information along scan lines. 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.

The paper from [7] describes a vision-based system functioning outdoors consisting of an autonomous hot metal carrier. The system from [8] uses easily identifiable features (landmarks, fiducials) of the form of concentric circles for easy registration. This however, requires the labeling of all pallets with such features. The success rate they obtained from 100 operations is 98%. The work [9] makes use of corner features, region growing and decision trees. Least squares line fitting and a single camera is employed in [10]. Other approaches

978-1-4673-8200-7/15/$31.00 ©2015 IEEE

include: [11] line-based model matching is used; [12] Colour-based segmentation; [13] sheet-of-light range camera.

The detection module of our system relies on previous advances in pedestrian detection such as Haar features and fast boosted classifiers [14]; the usage of integral features [15]; HOG features [16]. Recent advances and surveys are presented in [17].

Our previous work from [x] presented the system as a whole including the requirements, software and hardware architecture and initial tests. In this paper we describe the newly added modules and improvements compared to the original system.

III. Proposed improvements

Several new modules have been added to our system and many have been updated and improved.

A. Detection with aggregate features

A key element in a reliable system is to have alternatives for each important module. We have decided to try a different approach for pallet detection. Pallet detection is a subtask of object detection. Relevant advances have been made in the field of pedestrian detection with the sliding window approach. A pictorial overview of the representative aggregate channel features ACF method is shown in Figure 1.

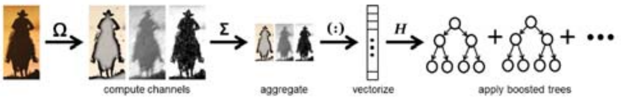

Fig. 1. Processing steps for detection based on aggregate channel features - figure from [18]

In the following we describe a similar solution adapted for the detection of pallets. In our experiments the pallets have a fixed aspect ratio of 5 and height varying between 20 and 100 pixels. The region of interest for pallet detection has a size of 264 x 972 pixels. We use a single detection window having a size of 24 x 120 pixels. This window includes also some context (20 %) and corresponds to a pallet of 20 x 100 pixel size.

For multiscale detection, we resize the image multiple times and use the same detection window. We use a scale range of 0.2 to 1.0 with scaling factor of 1.15. This way we are able to detect pallets with heights between 20 and 100 pixels using 12 scales. At each scale we use a step of 2 pixels for sliding. In order to classify the content of the sliding window we use a boosting classifier based on aggregated channel features (ACF) [18].

Eight image channels are computed from the input image for generating classification features: grayscale, gradient magnitude and oriented gradient magnitudes at six orientations (see Figure 2). Instead of the LUV color channels we use the gray level intensity. In [18] the channels were partitioned into 4 x 4 pixel aggregates. In our case we use smaller 2 x 2 aggregates due to the smaller resolution of the pallet model. This way the 24 x 120 pixel size sliding window is represented by 8 x 12 x 60 aggregates. The aggregated channels obtained by computing an average for each pixel aggregate and the classification features become simple pixel lookups. For multiscale detection the channels have to be recomputed for each individual scale. Using an efficient CPU implementation, feature computation for all 12 scales can be obtained at over 50 FPS.

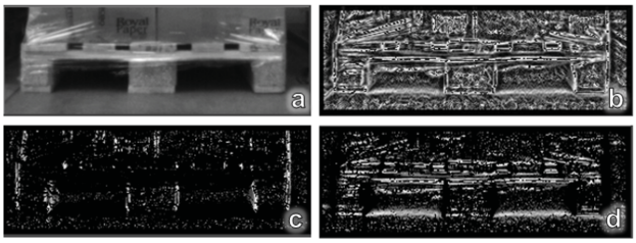

Fig. 2. Sample pallet and visualization of certain features - a) grayscale; b) gradient magnitude; c) gradient magnitude at orientation 0 degrees; d) gradient magnitude at orientation 90 degrees.

For classification we train a boosting based classifier using Adaboost with 2048 two level decision trees. We use the training protocol with 3 bootstrapping rounds described in [18]. This approach achieved outstanding detection accuracy and precision rates on multiple challenging benchmarks and handled well also difficult occlusion cases. The disadvantage of the approach is that the localization is not very precise. Usually multiple detections are obtained around an object and the non-maximum suppression does not always retain the window that is the best fit for the object.

B. Deeper trees

State-of-the-art pedestrian detection systems employ decision trees that are limited to height 2. It has been shown that this is a sweet spot that ensures the most accurate detection. Here we test whether or not such an observation holds for pallet detection. Aside from the fact that we detect different types of objects our feature vector also has other characteristics.

We operate with more descriptive features called normalized pair differences (**npd**). These features were proposed by us in a different work specifically for the task of pallet detection. They represent intensity pair differences normalized in such a way as to ensure certain illumination invariance properties.

Decision trees with limited height/depth are usually a compromise. It is often the case that leaf nodes contain a mixture of positive and negative instances and thus the tree must misclassify a certain percentage of the instances. By enabling a larger depth one can expect to have a more powerful classifier because of the increased possibility of branching further.

C. Unloading operations for racks

Unloading operations require the target position for the place of one or two pallets. We define the **unloading cuboid** as the cuboid in 3D space which represents the empty area in front of the AGV bounded by obstacles on the left and right and the floor on the bottom. Our aim is to detect this unloading cuboid. An error is signaled if for some reason the dimension of the cuboid is smaller than the dimension of the pallets that are to be unloaded.

Rack storage is comprised of poles holding up shelves at different heights. We start by detecting the structure of the rack. This can be obtained from the disparity map. The poles and the support for the pallets must be at roughly the same distance from camera because the approach of the AGV is almost perpendicular to the rack. We extract the disparity corresponding to the main fronto-parallel object from the scene. We call this the **principal disparity** d^*.

The principal disparity is taken to be the largest local maximum from the disparity histogram above a certain minimum threshold value.

$$d^* = \arg \max_{d > d_{min}} \{h(d), d \in N(d)\} \tag{1}$$

where h is the disparity histogram. This definition takes into account the following considerations. It is a local maximum because it must appear frequently in the image. It is the largest local maximum because we want to consider the closest fronto-parallel object. We must ensure that the disparity is higher than a limit because we want to eliminate cases where the background walls have a larger appearance frequency. This limit is calculated from the maximal admissible distance to the rack (around 3 meters).

Once the principal disparity is determined we filter the disparity image and retain only the disparity values that are close to the principal disparity and set to zero the rest of the map. We then construct vertical and horizontal projections of the non-zero elements.

$$D^*(x,y) = abs(D(x,y) - d^*) < 5 \tag{2}$$

$$D^*_x = \sum_y D^*(x,y) \tag{3}$$

$$D^*_y = \sum_x D^*(x,y) \tag{4}$$

These projections provide us the necessary information to delimit the free zone in front of the camera. Columns and rows with high projection values in D^*_x and D^*_y respectively mean the presence of the rack. We start from a point located in the middle of the region of interest of coordinates (x_0, y_0) and travel in three directions to find the left, right and bottom limit to the open space available.

The coordinates of the left limit in pixel coordinates is x_{left} and it is first value x left to x_0 for which the vertical projection $D^*_y(x)$ is above $max(h) * 0.8$. The value of x_{right} is determined in the same manner in the other direction. For y_{bottom} we use the horizontal projection $D^*_x(y)$.

The points (x_{left}, y_{bottom}) and (x_{right}, y_{bottom}) can be reconstructed using the principal disparity d^* to obtain the real world coordinates of the corners of the unloading cuboid: $(X_{left}, Y_{left}, Z_{left})$ and $(X_{right}, Y_{right}, Z_{right})$. The pallets must be placed inside the cuboid on the level of Y_{left}. The value of Y_{right} should be very close to Y_{left} otherwise we signal an error. The pallets must be placed next to the closest pole. This is the left pole if $|X_{left}| < |X_{right}|$ otherwise the right pole. All the presented operations are illustrated in Figure 3 for a sample case.

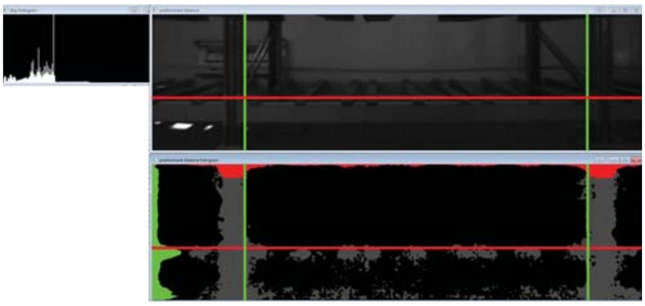

Fig. 3. Unloading operation for rack - visualization of processing steps. Top: disparity histogram and the limits of the unloading cuboid drawn on the input image; bottom: disparity image with principal disparity highlighted, the projections and the limits of the unloading cuboid are overlayed.

D. Block storage - ground level

It is required for scenarios where the unloading operation takes place in a block storage on ground level to determine the target position of the unloaded pallets based on markings on the ground. We have developed a module that tackles this specific task. The main tool employed in this process is the Hough transform [4], [5] for finding the important lines in the input image.

Since stereo reconstruction is not very reliable on the floor region both images need to be processed in order to have stereo information. We start by performing a standard Hough transform for line detection. The input image is a linear combination of the edge image and the original intensity image. The importance given to the edge image is 0.95 while for the intensity image it is 0.05. The reason for this is that we want to detect bright lines only. The bin size for angles is set to 1 degree. We locate the local maximum in the Hough accumulator using a neighborhood of 5. The left line from the floor marking is obtained by finding the line with the largest angle $\in [50, 70]$ degrees. We find the middle horizontal line through the region of interest. This line must also lie close to the middle of the region of interest. The position from the top of the region of interest must be $[90, 150]$. The right line from the floor marking is obtained by finding the line with the largest angle $\in [290, 310]$ degrees.

After having the three main lines for the floor marking we can extract the intersections. Name the intersection between the left line and the middle line point A_l for the left image and the other intersection as point B_l (see Figure 4). For the right image we define A_r and B_r analogously. We can reconstruct the 3D position of these points by using as the disparity values the differences $A_l - A_r$ and $B_l - B_r$. Alternatively, we can operate with the disparity values from the stereo matching algorithm.

The position of the pallets can be found by knowing their dimensions and placing them at a certain distance from each floor marking. Note, that currently we are assuming that the AGV is roughly facing the floor markings and that the middle line is approximately horizontal.

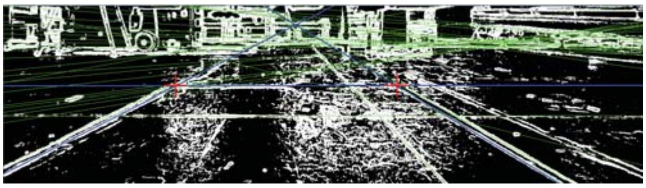

Fig. 4. Unloading operation ground level - edge map and the 50 most important lines obtained via Hough transform in green. In blue are the lines of the actual floor markings. Red crosses indicate the intersection points. The edge image is noisy because of light reflections from the floor.

E. Block storage - stacking

For this case we use most of the ideas from the rack case. However, the unloading cuboid must be on top of the block storage. In this case we pick x_{left} and x_{right} differently. Define x_{left} as first value x left to x_0 for which the vertical projection $D^*_y(x)$ is below $max(h) * 0.05$. This ensures that pallets are placed on top of the block storage. Figure 5 shows a sample scenario for this type of operation.

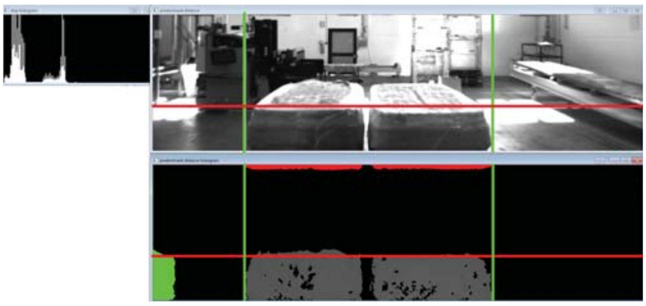

Fig. 5. Unloading operation for block storage - visualization of processing steps. Top: disparity histogram; limits of the unloading cuboid drawn on the input image; bottom: disparity image with principal disparity highlighted, the projections and the limits of the unloading cuboid.

IV. Experimental results

First, we define the overlap criteria that determines when to consider a detection a correct match for a ground truth bounding box. Let the rectangle A represent the predicted pallet bounding box (rectangle) and let B represent the ground truth hand labeled bounding box (rectangle). Let D represent the union between A and B obtained by taking the union of the intervals covered along x and y axis of the rectangles A and B (see Figure 6). Let C be the intersection of the rectangles A and B. We then define the absolute positioning error along x E_x and along y E_y:

$$E_x = D.width - C.width \tag{5}$$

$$E_y = D.height - C.height \tag{6}$$

If there is no overlap then $E_x = A.width + B.width$ and the error term attains the maximum value. In case of a perfect overlap the error is 0. We now define two separate criteria for considering matches. We define a **precise match** exists between A and B if $E_x \leq 15$ and $E_y \leq 15$. We define a **normal match** exists between A and B if $E_x \leq 50$ and $E_y \leq 50$.

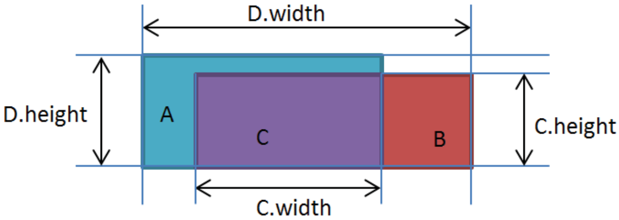

Fig. 6. Illustration of definitions from the overlap criteria

We evaluate the detection accuracy of different approaches for pallet detection. All classifier models are trained on the Viano2 training set and evaluated on two different test sets: test set Viano2 which is somewhat similar to the training set having been acquired in the same recording session, also this contains the highest number of annotated pallets; and test set Viano3-5 originating from several separate recording sessions. The second test set is more challenging and contains mostly difficult cases including over/under-exposed images; heavy glare; light artifacts. The composition of the sets is as follows: the training set contains 467 images and 891 labeled pallets (there can be zero or more than one pallet in each image); the test set Viano2 contains 7124 images and 9047 labeled pallets; test set Viano3-5 contains 467 images and 891 labeled pallets 224 and 356 labeled pallets.

We compare the newly introduced pallet detection module with aggregate features to the previous approach. The results with different detection module configurations are presented in Table I. The previous method that uses boosted decision trees and integral features is shown as baseline for comparison. We also evaluate the influence of increasing the depth of the decision tree and our newly introduced features normalized pair differences (**npd**). According to the results increasing the depth not only improves overall performance, it also has a positive effect on the detection accuracy for precise matches. This is essential for a more precise pallet localization. Increasing the depth only slightly increases training time and the execution speed for prediction.

	Viano2		**Viano3-5**	
Configuration	**normal**	**precise**	**normal**	**precise**
100 weak learners + 100 negatives				
old features - depth 2	79.0 %	64.2 %	-	-
npd - depth 2	95.4 %	91.4 %	87.6 %	55.9 %
npd - depth 3	97.7 %	92.0 %	88.2 %	68.8 %
npd - depth 4	98.3 %	93.7 %	90.5 %	72.2 %
npd - depth 5	98.3 %	94.7 %	93.8 %	78.1 %
1000 weak learners + 1000 negatives				
old features - depth 2	92.0 %	75.4 %	77.0 %	37.9 %
npd - depth 2	100 %	94.7 %	93.8 %	57.3 %
npd - depth 5	98.9 %	94.9 %	97.5 %	64.9 %
2048 weak learners + 3 bootstrap rounds				
aggregate features - depth 2	99.4 %	46.3 %	85.4 %	25.8 %

TABLE I. Detection accuracy on test sets Viano2 and Viano3-5

The newly added modules were tested in field during the most recent test session named Viano6. Several operations were performed using feedback from our system to correct the AGV position and fork placements. Qualitative results are presented in Table II where we count the number of operations performed and the ratio of the successful operations. All tests were successful, the single issue was that some parameters for unloading cuboid detection needed to be tuned for the current scenarios.

operation type	nr. op.	nr. successful op.	percent
loading - rack	101	101	100 %
loading - block	23	23	100 %
unloading - rack	82	82	100 %
unloading - block	15	15	100 %
total	221	221	100 %

TABLE II. VIANO6 FIELD TEST RESULTS

One of the main problems still present is the need for a robust auto-exposure module. Illumination can change from very dark - when there is a load on the forks - to very bright - when light is reflected from the floor or when light comes from behind the pallets. Another issue is the existence of minor reconstruction errors from the stereo module that show up as points placed very close to the camera. We have eliminated these by considering only a limited range of disparities.

V. CONCLUSION

We have presented improvements and extensions for a stereo-camera sensor system that is responsible for autonomous load handling. The newly added functions tackle unloading operations using scene understanding obtained from the stereo disparity map.

Tests show that increasing the depth of the decision trees is beneficial. It results in an increased detection rate and also an improvement in localization. Field tests reveal that the system is functional and the data sent to the AGV is precise enough to perform loading and unloading operations automatically.

Of course increasing robustness and execution time is an always present goal. For future work we plan to include a glare removal function. Glare is present almost always because of reflective plastic that covers the palletized goods. A position validation module based on location priors is under development. Also, a more sophisticated verification based on pallet position configuration is planned.

ACKNOWLEDGMENT

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 partners and especially Elettric80 for the opportunity to perform tests with their AGV at the warehouse in Viano, Italy.

REFERENCES

[1] “Pan-robots plug and navigate robots for smart factories,” http://www.pan-robots.eu/.

[2] M. R. Walter, S. Karaman, E. Frazzoli, and S. J. Teller, “Closed-loop pallet manipulation in unstructured environments.” IEEE, 2010.

[3] R. Bostelman, T. Hong, and T. Chang, “Visualization of pallets,” in *SPIE Optics East*, 2006.

[4] P. V. C. Hough, “A method and means for recognizing complex patterns,” 1962, u.S. Patent No. 3,069,654.

[5] R. Duda and P. E. Hart, “Use of the hough transformation to detect lines and curves in pictures,” *CACM*, vol. 15, pp. 11–15, 1972.

[6] L. Baglivo, N. Biasi, F. Biral, N. Bellomo, E. Bertolazzi, M. D. Lio, and M. D. Cecco, “Autonomous pallet localization and picking for industrial forklifts: a robust range and look method,” *Measurement Science and Technology*, vol. 22, no. 8, p. 085502.

[7] C. Pradalier, A. Tews, and J. M. Roberts, “Vision-based operations of a large industrial vehicle: Autonomous hot metal carrier,” *J. Field Robotics*, vol. 25, no. 4-5, 2008.

[8] M. J. Seelinger and J.-D. Yoder, “Automatic visual guidance of a forklift engaging a pallet,” *Robotics and Autonomous Systems*, vol. 54, no. 12, 2006.

[9] R. Cucchiara, M. Piccardi, and A. Prati, “Focus based feature extraction for pallets recognition,” in *BMVC*, 2000.

[10] S. Byun and M. Kim, “Real-time positioning and orienting of pallets based on monocular vision.” IEEE Computer Society, 2008.

[11] W. Kim, D. Helmick, and A. Kelly, “Model based object pose refinement for terrestrial and space autonomy,” 2001.

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

[13] J. Nygårds, T. Högström, and Å. Wernersson, “Docking to pallets with feedback from a sheet-of-light range camera.” IEEE, 2000.

[14] P. A. Viola, J. C. Platt, and C. Zhang, “Multiple instance boosting for object detection,” in *NIPS*, 2005.

[15] P. Dollar, Z. W. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *BMVC*, 2009.

[16] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005, pp. I: 886–893.

[17] R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Ten years of pedestrian detection, what have we learned?” in *ECCV-CVRSUAD*. IEEE, 2014.

[18] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *PAMI*, 2014.