Improving the performance of a SVM+HOG classifier for detection and tracking of wagon components by using geometric constraints

Camilo Lélis A. Gonçalves*, Ronaldo F. Zampolo*, Fabrício José B. Barros*,
Ana Claudia S. Gomes†, Eduardo C. de Carvalho†, Bruno Victor M. Ferreira†,
Rafael L. Rocha†, Rodrigo C. Rodrigues‡, Giovanni Augusto F. Dias‡, and Diego A. Freitas‡
*Universidade Federal do Pará, Instituto de Tecnologia
†Instituto SENAI de Inovação em Tecnologias Minerais ISI/SENAI
‡Estrada de Ferro dos Carajás, Vale S. A.
Correspondig author: camilo.goncalves@ieee.org, zampolo, fbarros@ufpa.br,
claudia.isi, eduardo.isi, bruno.isi@senaipa.org.br, rafael89.rocha@gmail.com, Giovanni.Dias@vale.com

***Abstract*—The inspection of train and railway components that can cause derailment plays a key role in rail maintenance. To improve productivity and safety, service providers look for automatic and reliable inspection solutions. Although automatic inspection based on computer vision is a standard concept, such an application challenges development community due to the environmental and logistic factors to be considered. Previous publications presented automatic classifiers to evaluate integrity and placement of wagon components. Although the high classification accuracy reported, ineffective object detection affected the general performance. Our object detector/tracker consists of a descriptor based on the histogram of oriented gradients, a support vector machine classifier, and a set of geometric constraints, which takes in account the ideal trajectory path of the wagon's components of interest and the distances between them. We detail training and validation procedures, together with the metrics used to assess the performance of the system. Presented results compare two other techniques with our approach, which exhibits a fair trade-off between reliability and computational complexity for the application of wagon component detection.**

I. Introduction

Computer Vision has been largely used in many production control and inspection systems in various fields, like Civil Engineering [1]–[3], Automotive Industry [4]–[6], Manufacturing [7]–[9] and Agriculture [10]–[12]. In the rail industry, the inspection of train and railway components that can cause derailment is an important maintenance issue. Thus, the design of automatic and reliable inspector systems is crucial to improve productivity and safety. Although automatic inspection by computer vision is already a standard concept, the task in this case is not trivial, due to inherent conflicting aspects: railway inspection facilities are subject to vibration, dust, and changing climate conditions; while the decision pipeline requires reliable stages of component estimation, detection, and classification. Considering the great variety of components of interest and the adverse environmental conditions, the development of effective inspector systems is, indeed, challenging. Extensive research is being done in regard to the inspection of train and railroad tracks [13]–[15]. In previous works, Rocha et al. [16]–[18], in a partnership with Vale S.A., developed and refined an approach, based on Convolutional Neural Networks, specific for inspection of the *pad*, one of the components in a wagon wheelset.

This work is inspired by the achievements and difficulties reported in [18]. Therein, the authors presented a pad evaluation system of two stages: detection and classification. Their detection approach is based on the Histogram of Oriented Gradients (HOG) [19], while the classification relies on Convolution Neural Networks (CNN) [20]. Although classification results were encouraging for the case of manual pad segmentation, when automatic segmentation is considered, the performance of the whole processing chain decays significantly as a direct consequence of an inefficient detector.

Thus, we propose an improved method to detect wheelset components of a wagon train. A HOG-based object detector is used, but its predictions are refined using a set of geometric constraints tailored to the characteristics and motion pattern of the wagon elements.

The remainder of this paper is organised as follows. The next section addresses the proposed detection and tracking strategies used to identify *axle boxes*, the elements of interest in our study. The section also describes the structural characteristics of the input images of the system. Section III details the training dataset used in the experiments, the methodology applied to design the geometric constraints that are the base of our component identification strategy, and the metrics used to assess the performance. In Section IV our results are presented and discussed. Section V concludes the paper, emphasising our contributions.

II. Detection and tracking

Figure 1 shows a shot taken from the typical viewpoint of the video data used in our experiments. Outlined in red, on can see the so-called *truck* (or *bogie*), which is a framework that carries the *wheelset*, a modular sub-assembly of *wheels* and *axles*. As show in Figure 2, in a truck, three main components

Fig. 1. Sample image of a train wagon with one of its trucks (outlined in red) and two wheels bolts (outlined in green).

Fig. 2. A closer view of the truck: the axle boxes in green, the pads in red, and the coil springs of the suspension in blue.

can be identified: the *coil spring* of the *suspension* (in blue), the *pads* (in red) and the *axle boxes* (in green).

The proposal consists in the joint actions of an object detector and a set of geometric constraints. Next we describe the arrangement of components in the side view of a train wagon, the particular structure of the *axle box*, and the theory behind the adopted tracking strategy.

A. HOG-based detector.

An object detector is used to identify any new elements of interest that appear in the monitored video frames (see Section II-B). Such new elements are added to the list of detected elements, which are tracked in the subsequent frames as long as they remain on sight.

In this work, the object detector is based on HOG descriptors [19]. For a given image patch, we compute a vector of HOG descriptors, which will be used as input features for a SVM [21] classifier that will find out whether the patch represents our object of interest. We can detect the occurrences of the object at various locations in an arbitrary image by applying the *sliding-window* approach. In addition to the high availability of implementations of the algorithm and its prevalent use, other reasons behind this choice are the fast training, ease of use, and low hardware requirements when compared with another methodologies (e.g. Haar-like features [22], and deep learning-based [23]–[29]).

B. The axle box as reference element

The axle box (Figure 4a) is chosen as the key-component to our object detection strategy. The first reason for this choice is that the characteristics of axle box bolts are easily identifiable by the object detector. Moreover, such an element is always present in the wagon truck structure, what makes it a good and reliable reference point.

The fact that axle boxes always come in pairs can be used to aid in identifying the train wagon by a simple matching strategy. By considering just the side view, one train wagon contains one pair of trucks, which, in turn, has one pair of axle boxes each. So, each pair of elements can be grouped in order to determine the wagon they belong to.

Another reason is that the positions of each pair of axle boxes can be used as a reference to locate other elements of interest in the truck. For example, by assuming geometric regularity in the arrangement of wagon elements, expected positions and dimensions of *suspension* and *pads* can be easily determined (Figure 2).

Finally, grouping the pairs of axle boxes and identifying the trucks and wagons to which they belong can aid in the prevention of false positive detection cases. Furthermore, this information can help tracking the elements of interest over video frames. This property is used as base for the tracking methodology presented in this paper, and will be further elaborated in the next sections.

C. Axle boxes trajectory profile

In our experimental set-up, as the video camera remains fixed and the movement of the wagon occurs parallel to the image plane, one can infer that each *axle box* will occupy some predefined spatial positions in the image, describing a *fixed linear trajectory path*, what we refer as *trajectory profile*.

Let the vector

$$\mathbf{b}_i = (b_x, b_y, b_w, b_h) \tag{1}$$

represents the position and dimensions of the i-th box being currently tracked, where b_x and b_y are the horizontal and vertical coordinates of the box centre, and b_w and b_h denote the box width and height, respectively.

Thus, the trajectory path can be approximated by a line equation

$$\alpha b_x + \beta b_y + \theta = 0 \tag{2}$$

where α, β and θ are the line parameters.

The assumption of a *trajectory path* enables a fine-grained detection and tracking of new *axle boxes*. Any new element detected whose distance between its centre coordinates and the trajectory line is above some predefined threshold will be ignored and, consequently, will not be tracked, decreasing the risk of tracking errors in subsequent frames.

D. Distance between axle boxes

Analysing Figure 2, one can note a pattern in the distance between axle boxes, which can be used to determine if two axle boxes belongs to the same truck.

First, aiming to have a scaling-invariant figure, the ratio R is computed as

$$R = \frac{d_{len}}{o_{dim}} \tag{3}$$

where o_{dim} is the value of one of the dimensions [1] of the box detected and d_{len} is the length being analysed (distance between boxes).

By using R, the distance between two axle boxes can be classified in:

- *Intra-truck* – between two consecutive axle boxes of the same truck;
- *Inter-truck* – between axle boxes of different trucks in the same wagon;
- *Inter-wagon* – between axle boxes of different wagons.

Figure 3a shows an example of a distance between axle boxes of the same wagon, but in different trucks, while the Figure 3b depicts another example with distances between boxes of different trucks in different wagons and between boxes of the same truck, respectively.

Due to the rigid arrangement of the axel boxes in the structure of the wagon, false positive detection can be reduced by checking whether the distance between the new and the last detected boxes can be classified in one of the mentioned categories. If not, the new element will be ignored. The classification procedure also takes into account the previous classes assigned to detected boxes to reduce error rates. For instance, in Figure 3a, it is clear that the next axle box (not visible yet) will belong to the class *intra-truck*, while in the Figure 3b the *box 3* may be either of class *inter-truck* or *inter-wagon*, depending on its distance with relation to *box 2*. Such strategy helps preventing high misdetection and tracking error rates.

E. Axle box tracking

A *tracking-by-detection* approach was adopted to handle the tracking of the axle boxes through video frames. Let us define the list $\boldsymbol{b}_F = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, ..., \mathbf{b}_M)$ of the M axle boxes (Equation 1) being currently tracked by the system in frame F, and the list $\boldsymbol{d}_{F+1} = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, ..., \mathbf{d}_N)$ of N axle boxes in frame $F+1$ found by the object detector ($M \geqslant N$). The correspondence of the boxes in $\boldsymbol{b}_F$ and the ones in $\boldsymbol{d}_{F+1}$ is determined by the box pairs with highest *intersection over union* (IoU) index (Algorithm 1). Boxes in $\boldsymbol{d}_{F+1}$ not matched to any of the boxes in $\boldsymbol{b}_F$, but in compliance with the geometric constraints presented in Sections II-C and II-D will be added to the list of tracked boxes $\boldsymbol{b}_{F+1}$. The boxes in $\boldsymbol{b}_F$ with no corresponding box in $\boldsymbol{d}_{F+1}$ will be handled by the method described in the next section.

F. Position variance and occlusion handling

The performance of the object detector can deteriorate due to changes in illumination exposure, camera vibration, object occlusion (as a consequence of dust, employees passing, etc.) and presence of motion blur, just to name a few factors. Such deterioration leads to variations in the tracked positions or even detection failure of elements already tracked. The fixed structure of the wagon and the assumption of the *trajectory profile* (Section II-C) enable position correction and occlusion handling for tracked axle boxes applying simple strategies.

[1]Which dimension of the bounding box, height or width, is unimportant, as all boxes are squares.

Algorithm 1 Algorithm to make the correspondence of the boxes of two lists.

```
function CORRESPONDENCE(b_F, d_{F+1})
    matches ← [ ]
    for all b_m in b_F do
        candidates ← [ ]
        for all d_n in d_{F+1} do
            if d_n intersects b_m then
                append d_n to candidates
            end if
        end for
        best_match ← null
        best_value ← 0
        for all c in candidates do
            overlap ← jaccard(c, b_m)
            if overlap > best_value then
                best_value ← overlap
                best_match ← c
            end if
        end for
        append (b_m, best_match) to matches
    end for
    return matches
end function
```

To address the position variance problem, consider $\boldsymbol{k}_F$ as the list of centre coordinates of N *axle boxes* detected in the actual frame F (thus a subset of $\boldsymbol{b}_F$) that matches some of the boxes in $\boldsymbol{d}_{F+1}$.

The displacement vector $\boldsymbol{\Delta k}_F^i$ that will update the position of the i-nth box of $\boldsymbol{k}_F$ is computed as:

$$\boldsymbol{\Delta k}_F^i = (\boldsymbol{d}_{F+1}^i - \boldsymbol{k}_F^i) + \sigma(\boldsymbol{d}_{F+1}^i - \boldsymbol{\theta}_{F+1}^i) \tag{4}$$

where $\boldsymbol{\theta}_{F+1}^i$ is the intersection point of the trajectory profile (Section II-C) and its perpendicular that contains the coordinates $\boldsymbol{d}_{F+1}^i$, and σ is a correction factor. Then the updated coordinates $\boldsymbol{k}_{F+1}^i$ are computed as

$$\boldsymbol{k}_{F+1}^i = \boldsymbol{\Delta k}_F^i + \boldsymbol{k}_F^i \tag{5}$$

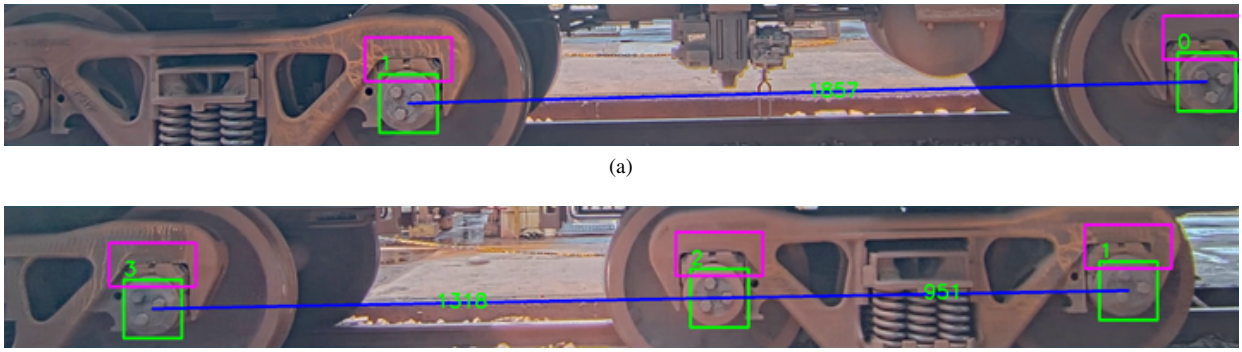
The function in (5) smooths the real trajectories draw by the *axle boxes*, alleviating noisy position estimations.

The update $\boldsymbol{\Delta u}_F$ of the set $\boldsymbol{u}_F$ of the $M - N$ boxes that do not have a corresponding detection in frame $F+1$ (caused by occlusion or misdetection) is addressed by determining the average of the displacement vectors $\boldsymbol{\Delta k}_F^i$

$$\boldsymbol{\Delta u}_F = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Delta k}_F^i \tag{6}$$

Hence, the update of $\boldsymbol{u}_F$ is given by

$$\boldsymbol{u}_{F+1}^i = \boldsymbol{\Delta u}_F + \boldsymbol{u}_F^i. \tag{7}$$



(a)

(b)

Fig. 3. Examples of distance lengths between wheels bolts. top-down, left-right order: (3a) inter trucks; (3b) inter wagons and inter bolts of the same truck.

III. Experimental set-up

This section explains implementation details of the proposed system and the routines designed to evaluate its performance. Such routines consist in:

1) Building the dataset that will be used to train/test the object detector;
2) Designing the geometric constraints of the tracking system;
3) Computing performance metrics.

A. Dataset creation

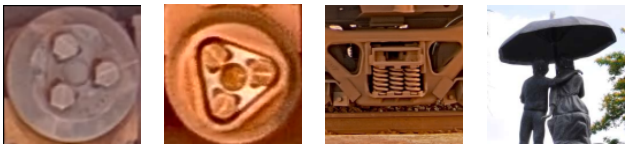
To create the dataset, a large number of photos were taken depicting the side view of wagons, as show in Figure 1.

For the creation of the set of positive examples, patches were sampled from these images as show in Figures 4a and 4b. One can note that the region comprises all the bolts of one *axle box*. In all, approximately 65 positive patches were extracted from different images of wagons. To build a robust detector, these samples were further processed using techniques of *Data Augmentation*: each sample was rotated 10 times, with steps of 36 degrees, and gamma corrected with a factor ranging from 0.35 to 1.4 with steps of 0.15, resulting in approximately 50,000 positive samples.

For the negative sample set, non-positive patches from wagon photos were sampled at multiple scales (example in Figure 4c). Additionally, other patches were randomly sampled from high-resolution images publicly available in the internet. These images were carefully selected to ensure that none of them have bolts-like structures (example in Figure 4d). Approximately 10,000 negative samples were obtained and augmented by using the same gamma adjusting technique, resulting in nearly 80,000 negative examples.

All sampled patches of the dataset were resized to 68×68 pixels. From the whole dataset, 70% of images were used to train the detector and the remaining 30% for testing.

Our database were created using sample images acquired during the operation of the railway, with the permission of our partner Vale S.A.. Due to internal policy of the industrial partner, the data collected is a private intellectual property.



(a) (b) (c) (d)

Fig. 4. Examples of images of the dataset: (4a) and (4b) positive samples of wheel bolts; (4c) negative sample from a wagon image; and (4d) negative sample from a random image.

B. Geometric constraints

To select the geometric constraints that better fit the tracking system, a short video containing typical moving wagons was selected. Every axle boxe boundaries present in each video frame were manually annotated. The annotations were used as the reference data to the routines that will be presented in the following.

The first routine is responsible for computing the ideal path trajectory of every axle box across frames. While the reference data supplies sufficient information to determine the trajectory profile described in Section II-C, the objects found by the trained object detector are used to compute the mean variance of their centre coordinates from the ideal trajectory. The mean variance is then used to determine the distance threshold.

Figure 5 depicts the *region of interest* (red outline) overlaid by the *path profile* line (green line), the *distance threshold* region (blue outline) and the detected *axle boxes* (dark-green rectangles) with their centres (red dots). Ideally, the distance threshold must be small enough to exclude any possible false-positive detection, like the upper-left detection of the image, but large enough to accommodate the position variation of the detected boxes.

The second routine is responsible for determining the distances between axle boxes. Figure 3 shows the visual feedback of such a script. The reference data are used to determine R classification ranges as described in Section II-D.

C. Performance assessment

The performance assessment of the proposed system is two-fold. First, we use the *stratified cross-validation* [30]

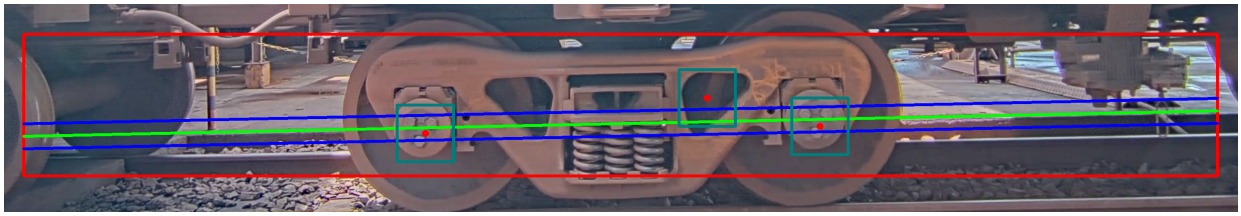


Fig. 5. Visual feedback of the path profiling script: the light-green line represents the ideal wheels bolts trajectory path; the blue lines denote the estimated distance threshold; dark-green rectangles and red dots represent candidate bolt regions and their centres, respectively.

TABLE I
METRIC COMPUTED FOR EACH STAGE OF THE VALIDATION. EACH STAGE REPRESENTS THE TRAIN/TEST SPLIT OF THE DATASET.

Stage	Accuracy	Precision	Recall	F-Score
Trainning				
1	0.9992	1.0000	0.9938	0.9969
2	0.9991	1.0000	0.9931	0.9966
3	0.9990	1.0000	0.9925	0.9962
4	0.9990	0.9994	0.9925	0.9959
5	0.9994	1.0000	0.9950	0.9975
Test				
1	0.9987	1.0000	0.9900	0.9950
2	0.9994	1.0000	0.9950	0.9975
3	0.9990	0.9975	0.9950	0.9962
4	0.9997	1.0000	0.9975	0.9987
5	0.9981	1.0000	0.9850	0.9924

methodology to evaluate result consistency of the proposed object detector. Then, we compare the technique described so far with other two similar systems:

- System 1 – uses the raw HOG object detection technique for detection/tracking, as described in Section II-E, without any geometric constraints;
- System 2 – applies HOG detection at each 10 frames to update the estimates of the object tracker based on the *kernelized correlation filters* technique [31].

The choice of such systems is justified as their use is consolidated in solving common *Computer Vision* problems [32]–[36].The same dataset were used to compute the performance metrics of all systems.

IV. RESULTS

A. Evaluation of the HOG-based detector

As stated in Section III-C, *stratified cross-validation* was used to evaluate the performance of the HOG-based detector. Five folds were used.

Table I shows the performance metrics of each training/testing split. Analysing the performance metrics, we can verify the detector is consistent and performs well (all rates above 0.9).

Although the *geometric constraints* are the base for this research, note that the object detector plays an important role in this methodology, and the quality of its predictions impacts upon the overall performance of the system.

B. Geometric constraints evaluation.

To check the effectiveness of the geometric constraints imposed, the path profile and distances classification ranges were adjusted using the routines described in section III-B.

In our experiments, the classification ranges of the R ratio used were:

- Intra-truck – [6.0, 7.0];
- Inter-truck – [12.0, 14.0];
- Inter-wagon – [9.0, 10.0].

In this work, an annotated bounding box and an estimated bounding box *match* when they have an IoU of, at least, 50%. A match is considered an *true positive* (TP) detection. If an annotated box does not have a corresponding matching estimated box, the annotation is considered a *false negative* (FN) detection, and if an estimated box does not have a corresponding matching annotated box, the estimated box is considered a *false positive* (FP) detection.

Performance metrics were computed for the proposed technique and the two systems mentioned in section III-C. The results are shown in Table II. Analyzing these results, it can be inferred that, although system 1 has a greater number of TP detections and system 2 has a smaller number of FP detections when compared with the proposed technique, our approach achieves the better in-balance between true and false detection rates than the other systems. For example our system deals better with FP detections compared with system 1.

Our methodology also outperforms system 2 FN rate by a factor of $\frac{1}{3}$. This is possibly due to the fact that, differently from *object detectors*, *object trackers* work in the "instance level" of the objects. In other words, each tracker will be responsible only for the tracking of the object which it was initialised. The detection of new objects will be handled by a new object detector, that will run at some predefined interval of frames (10 in this experiment). Although less computationally expensive with relation to our approach and system 1, this methodology still fails to detect and track a great number of new elements, leading to an increased rate of false negatives.

V. CONCLUSION

In this paper, we have proposed a simple but effective approach to detect and track wheelset components of train wagons using computer vision techniques. The focus was on a component of the wheelset called *axle box*. Our method considered the rigid structure of the wheelset, its motion pattern and the usual disposition the axle box assumes in each video

TABLE II
COMPARISON OF THE PERFORMANCE METRICS COMPUTED FOR THE TWO REFERENCE SYSTEMS AND OUR PROPOSED TECHNIQUE.

	System 1	System 2	Proposed Technique
TP	1597	1433	1593
FP	22	0	4
FN	76	240	80
Precision	0.9864	1.0000	0.9975
Recall	0.9546	0.8565	0.9522
Miss Rate	0.0454	0.1435	0.0478
Total o elements		1673	
Total of Frames		650	

frame. The detection of components uses descriptors based on the *histogram of oriented gradients* and *support vector machines*. Images of axle boxes (positive samples) and random images (negative samples) formed the dataset. Techniques of *data augumentation* were applied to expand the database to cover more variations of axle boxes. The tracking of axle boxes adopted a tracking-by-detection approach. We considered an update of object position when a positive detection in the current frame exhibits intersection over union above 50 % with an object in the frame immediately before. Aiming to reduce the false positive rate, and cope with false negative detection and object occlusion, we proposed two geometric constraints.

The first constraint was based on the fact that the camera point-of-view is fixed and the component movement describes an almost linear trajectory. We defined a *trajectory profile* of the axle boxes and a corresponding distance tolerance. Any detection which the distance of its centre coordinates to the trajectory profile is above the distance tolerance will be discarded as a false positive detection.

The second constraint considered that the normalised distances between axle boxes are predefined according to theirs positions in the train wagon. Those distances can be classified in *inter-truck*, *intra-truck* and *inter-wagon*. If the distance of a new detection to its neighbour object cannot be classified into one of these groups, the detection is considered a false positive and is ignored.

For the performance evaluation of our HOG-based detector, we designed a specific dataset and applied the *stratified cross-validation* methodology with five folds. Our results shows the quality of our classifier's estimates, an essential factor for the overall tracking performance.

The performance evaluation of our geometric constraints was made by computing performance metrics of the tracking of axle boxes in a sample video. Those metrics were compared with the metrics of two widely used tracking methodologies in literature. The results show the promising effectiveness of our strategy.

Note that the detection, tracking and constraint validation stages are independent to each other, and other methodologies can be applied to each one. The next step of our research is to introduce modern strategies, like a *deep learning*-based object detectors [23]–[29] or tracker [37], in one or more of the stages in order to improve computational efficiency and precision.

REFERENCES

[1] A. M. A. Talab, Z. Huang, F. Xi, and L. HaiMing, "Detection crack in image using Otsu method and multiple filtering in image processing techniques," *Optik*, vol. 127, no. 3, pp. 1030–1033, feb 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402615012164

[2] S. Iyer and S. K. Sinha, "A robust approach for automatic detection and segmentation of cracks in underground pipeline images," *Image and Vision Computing*, vol. 23, no. 10, pp. 921–933, sep 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885605000764

[3] S. Alam, A. Loukili, F. Grondin, and E. Rozière, "Use of the digital image correlation and acoustic emission technique to study the effect of structural size on cracking of reinforced concrete," *Engineering Fracture Mechanics*, vol. 143, pp. 17–31, jul 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0013794415003045

[4] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes, and R. Arroyo, "DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 2014, pp. 240–245.

[5] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An Unsupervised Approach for Inferring Driver Behavior From Naturalistic Driving Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3325–3336, 2015.

[6] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, "Driver cell phone usage detection on Strategic Highway Research Program (SHRP2) face view videos," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 35–43.

[7] Y. Lu, X. Xu, and J. Xu, "Development of a Hybrid Manufacturing Cloud," *Journal of Manufacturing Systems*, vol. 33, no. 4, pp. 551–566, oct 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027861251400048X

[8] L. Wang, M. Törngren, and M. Onori, "Current status and advancement of cyber-physical systems in manufacturing," *Journal of Manufacturing Systems*, vol. 37, pp. 517–527, oct 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612515000400

[9] D. Wu, D. W. Rosen, L. Wang, and D. Schaefer, "Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation," *Computer-Aided Design*, vol. 59, pp. 1–14, feb 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010448514001560

[10] K. Kapach, E. Barnea, R. Mairon, Y. Edan, and O. Shahar, "Computer vision for fruit harvesting robots—State of the art and challenges ahead," *International Journal of Computational Vision and Robotics*, vol. 3, pp. 4–34, 2012.

[11] H. Zareiforoush, S. Minaei, M. R. Alizadeh, and A. Banakar, "Potential Applications of Computer Vision in Quality Inspection of Rice: A Review," *Food Engineering Reviews*, vol. 7, no. 3, pp. 321–345, 2015. [Online]. Available: https://doi.org/10.1007/s12393-014-9101-z

[12] J. G. Arnal Barbedo, "Digital image processing techniques for detecting, quantifying and classifying plant diseases," *SpringerPlus*, vol. 2, no. 1, p. 660, 2013. [Online]. Available: https://doi.org/10.1186/2193-1801-2-660

[13] E. Resendiz, J. M. Hart, and N. Ahuja, "Automated Visual Inspection of Railroad Tracks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 751–760, 2013.

[14] I. Tang and T. P. Breckon, "Automatic Road Environment Classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 476–484, 2011.

[15] M. Singh, S. Singh, J. Jaiswal, and J. Hempshall, "Autonomous Rail Track Inspection using Vision Based System," in *2006 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*, 2006, pp. 56–59.

[16] R. L. Rocha, A. C. Q. Siravenha, A. C. S. Gomes, G. L. Serejo, A. F. B. Silva, L. M. Rodrigues, J. Braga, G. Dias, S. R. Carvalho, and C. R. B. de Souza, "A Deep-learning-based Approach for Automated Wagon Component Inspection," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18. New York, NY, USA: ACM, 2018, pp. 276–283. [Online]. Available: http://doi.acm.org/10.1145/3167132.3167157

[17] E. Fernandes, R. L. Rocha, B. Ferreira, E. Carvalho, A. C. Siravenha, A. C. S. Gomes, S. Carvalho, and C. R. B. de Souza, "An Ensemble of Convolutional Neural Networks for Unbalanced Datasets: A case Study with Wagon Component Inspection," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–6.

[18] R. Rocha, A. Siravenha, A. C. S. Gomes, G. L. Serejo, A. Silva, L. Mousinho Rodrigues, J. Braga, G. Dias, and S. Carvalho, "Avaliação de técnicas de Deep Learning aplicadas à identificação de peças defeituosas em vagões de trem," in *Workshop of Industry Applications (WIA) in the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17)*, E. Clua and F. L. C. Pádua, Eds., Niterói, RJ, Brazil, 2017. [Online]. Available: http://sibgrapi2017.ic.uff.br/e-proceedings/assets/papers/WIA/WIA6.pdf

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893.

[20] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object Recognition with Gradient-Based Learning," in *Shape, Contour and Grouping in Computer Vision*. London, UK, UK: Springer-Verlag, 1999, pp. 319—-. [Online]. Available: http://dl.acm.org/citation.cfm?id=646469.691875

[21] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511– I–518.

[23] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[24] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.0, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[25] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *CoRR*, vol. abs/1506.0, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[26] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *CoRR*, vol. abs/1506.0, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *CoRR*, vol. abs/1612.0, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[28] ——, "YOLOv3: An Incremental Improvement," *CoRR*, vol. abs/1804.0, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *CoRR*, vol. abs/1512.0, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[30] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[31] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *CoRR*, vol. abs/1404.7, 2014. [Online]. Available: http://arxiv.org/abs/1404.7584

[32] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian Detection using Infrared images and Histograms of Oriented Gradients," in *2006 IEEE Intelligent Vehicles Symposium*, 2006, pp. 206–212.

[33] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora, "Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–5.

[34] R. Xu, S. Y. Nikouei, Y. Chen, A. Polunchenko, S. Song, C. Deng, and T. R. Faughnan, "Real-Time Human Objects Tracking for Smart Surveillance at the Edge," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[35] H. Cheng, L. Lin, Z. Zheng, Y. Guan, and Z. Liu, "An autonomous vision-based target tracking system for rotorcraft unmanned aerial vehicles," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1732–1738.

[36] J. Greenhalgh and M. Mirmehdi, "Real-Time Detection and Recognition of Road Traffic Signs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1498–1506, 2012.

[37] D. Held, S. Thrun, and S. Savarese, "Learning to Track at 100 FPS with Deep Regression Networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 749–765.