2017 IEEE Winter Conference on Applications of Computer Vision

Telecom Inventory management via object recognition and localisation on Google Street View Images

Ramya Hebbalaguppe, Gaurav Garg, Ehtesham Hassan, Hiranmay Ghosh, and Ankit Verma
TCS Research, New Delhi, India
contact author - ramya.hebbalaguppe@tcs.com

Abstract

We present a novel method to update assets for telecommunication infrastructure using google street view (GSV) images. The problem is formulated as a object recognition task, followed by use of triangulation to estimate the object coordinates from sensor plane coordinates; To this end, we have explored different state-of-the-art object recognition techniques both from feature engineering and using deep learning namely HOG descriptors with SVM, Deformable parts model (DPM), and Deep learning (DL) using faster RCNNs. While HOG+SVM has proved to be robust human detector, DPM which is based on probabilistic graphical models and DL which is a non-linear classifier have proved their versatility in different types of object recognition problems. Asset recognition from the street view images however pose unique challenge as they could be installed on the ground in various poses, orientations and with occlusions, objects camouflaged in the background and in some cases inter class variation is small. We present comparative performance of these techniques for specific use-case involving telecom equipment for highest precision and recall. The blocks of proposed pipeline are detailed and compared to traditional inventory management methods.

1. Introduction

It took nearly two centuries to move from a mechanical string phone, invented in 1667, to the electronic phone invented in 1876. Subsequently, a century for the whole world to be connected through a copper network. But then, it took just 5 decades more for the world to become wireless. And now, within a matter of just two decades, we have moved to IPTV and the disruptive VOIP from the plain old telephone service using fiber. The changes in technology and the increasing expectations of customers have put Telecom companies under tremendous pressure to overhaul their networks, and corresponding equipments swiftly. Telecom companies also need to keep their costs under control to keep shareholders happy. This requires the need to have accurate knowledge of their on the ground and under-ground assets with their corresponding location to upgrade from copper to fiber network[1]. Asset inventory management, hence, is a challenging problem for many organizations which provide infrastructural services such as telecommunication, power utilities, transport amongst others. Our solution aims to reduce the survey requirement by identifying over-ground inventory details from GSV imagery using leading edge algorithms that utilize computer vision techniques based on object recognition/localisation. It helps to get the survey done at desktop/tablet with great ease circumventing manual labour.

The key contributions of the paper are twofold:

1 A novel pipeline for inventory management using GSV images[2] collection is presented. We have focussed on four types of equipments namely cabinets, manholes, joint boxes and poles (Refer Figure 2). We propose a method to estimate the real-world position of such objects using GPS co-ordinates associated with street view collection. The pipeline plays critical role in prioritizing the maintenance or replacement of assets.

2 We deal with specific types of object that pose serious challenges such as recognition of assets installed on the ground is extremely difficult with lots of clutter and significantly different point of views. Many of telecom assets also bear significant resemblance with the surroundings. This means our dataset is harder than the likes of the standard PASCAL VOC dataset [7] as will be shown later in this paper.

Telecom and Utilities companies maintain their outside plant inventory details in Geo-spatial system in the intelligent vector data format [4]. These details are not always accurate as they were originally recorded on paper and during migration to geospatial systems. Industry partner reported

[1] http://www.greatachievements.org/?id=3625

[2] Our industrial partner obtained a license agreement with Google to use Street View images.

978-1-5090-4822-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WACV.2017.86

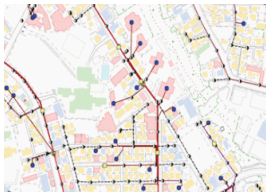
a  b  c 

Figure 1. Traditional methods used for survey to update Telecom assets. (a) field personnel conducting surveys (b) Collection of survey identifying the Telecom equipment (c) Telecom records are updated to correct the GPS locations of the newly placed equipment/correction of old equipment locations.

error rate of $30-40\%$, to reduce the error rates, we propose a pipeline to reduce the error rates in asset database. The data updated by field surveys are time consuming, resource intensive and prove costly. While several Telecom companies have started to leverage GIS based asset management system capabilities, a significant percentage of asset data in such systems is incomplete or inaccurate [2]. Creating a comprehensive telecom inventory on the condition of all assets is challenging due to the high-volume of data that must be collected. The Telecom companies incur high cost due to manual labour for surveys. Inaccurate and incomplete data leads to: (a) delays in the planning phase if new equipment is added (b) need for additional surveys before plan finalization (c) delays can lead problems such as customer order fulfilment. Figure 1 illustrates surveying method for updating the inventory prior to remote sensing technologies. Since the introduction of Google Street View (GSV), a part of Google Maps, vehicles equipped with roof-mounted mobile cameras have methodically captured street-level images of entire cities. The image acquisition from GSV API is detailed in [1] [9]. GSV creates an excellent opportunity for developing computer vision algorithms to leverage these assets for detecting, classifying, and localizing them.

Organisation of the paper is as follows: Section 2 presents the prior work in asset detection and inventory management. Section 3 details the pipeline we propose for asset detection using computer vision techniques. In Section 3.2, the popular shallow and deep features are discussed. Subsection 3.3 describes estimating object coordinates using triangulation. Then the data correction tool is updated as summarised in Section 3.4. Section 4 presents details about the dataset, results – Precision-recall, insights from confusion matrix, object localisation, and sample detections. Finally we conclude with summary of our contributions and list pointers for future work.

2. Related Work

Use of satellite images for maintaining records of natural resources such as lakes, forest cover, ice cover and for studying the environment changes has been well researched topic in remote sensing domain [20]. Guler et al. manage railway assets that include signs, signals and equipment boxes, amongst others [12]. They utilize a moving platform for survey routes along which assets of interest are placed. While the positioning equipment provides geographical information, the imaging sensor provides data related to assets. This method is labour intensive and requires equipments for positioning and moving platform. Hodges et al. [15] propose a system and method for the detection and management of network assets which have unique IP addresses for assets like computers, printers on the remotely connected network. These assets differ from our data, as they are network based in nature and hence their IP address can be used to track and update the inventory.

Warsop and Singh [21] present a detailed survey on computer vision based asset recognition and inventory management. However, these techniques involve the usage of handcrafted shallow features unlike the method using deep learning based on faster RCNN [18]. Deep learninng based detectors usually are easy to train on generic categories of assets, much more robust unlike [21] where one has to handcraft features again for particular object class. The literature mainly uses feature-based scale-invariant method for this task; Scale-Invariant Feature Transform (SIFT [17]) has been one of the most preferred feature representation in this category. Hu and Tsai [16] proposed an intelligent traffic sign detection and recognition system from videos which uses shape, color, position and texture attributes for sign localization and recognition in rule based formulation. The traffic signs installations is in upright position for clear visibility unlike telecom assets which can be installed on the ground. Also, they do not use GSV images unlike us. Balali et al. [1] use images from GSV API to leverage a sliding window based scanning to detect potential candidates for traffic signs. Each candidate window is recognized using multi-class SVM by extracting HoG[5] and color histogram combined by linear concatenation. In real-life situation where we have large number of images to process with assets appearing in varying sizes, this approach in not suitable as they do not address the localisation problem. We show in our work, a simple HoG fails miserably in our asset data even as we train using hard negative mining, at different scales to ensure robustness. The descriptors such as

Figure 2. Illustration of assets of interest to our industry partner. These assets are mainly the street furnitures - (a) pole, (b) jointbox occluded by a pole, (c) manholes, and (d) the green colored cabinets with jointboxes in the same image. A single image can contain multiple object categories and also could occlude each other.

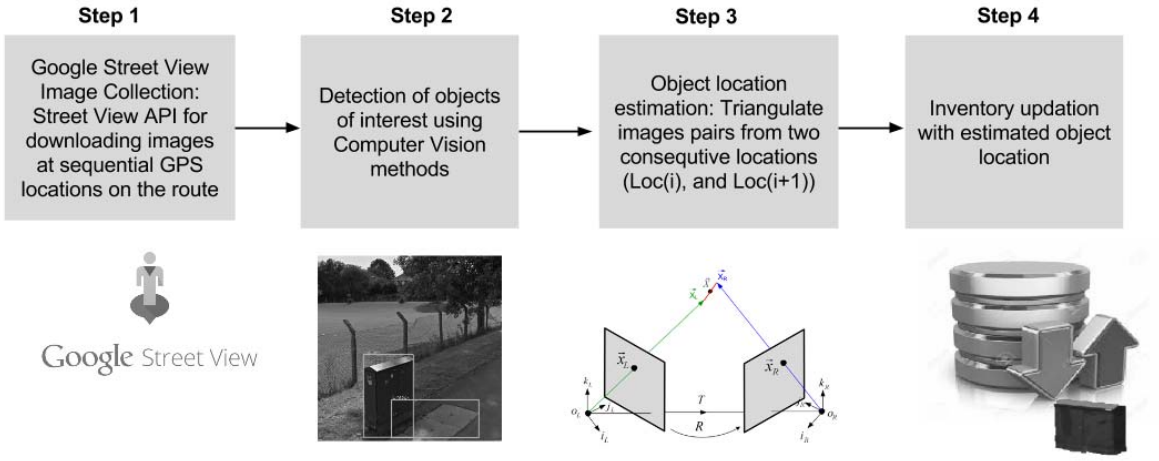


Figure 3. Proposed Method to update assets using Computer Vision for object detection and updating the inventory remotely without the need for laborious surveys. The asset recognition and updation pipeline comprises of 4 steps. Step 1 involves acquisition of images from Google Street View (GSV). Refer Figure 4 for more details. Step 2, involves, identification of assets using deep our proposed method. Step 3, is the estimation of coordinates of the asset using triangulation, step 4, updating the estimated GPS coordinates from step 3 into record management system

SURF [3], SYBA [6], and TreeBASIS [10] are hand crafted features werent as powerful enough to provide matches between two distinct object instances from the same visual object category. Mechanism for capturing object part deformation relative to each other and some image patches were missing. Faster RCNN which exploits on deep features directly extracted from the image space making them robust against illumination changes, view-point variations and blur offer accurate and real-time performance for updating the asset. It is proven that deep learning based recognition outperforms the shallow descriptors.

GSV images are collection of digital images captured by car mounted camera, and contributed by individuals that are stitched in street-view application to create panoramic appearances of the corresponding surrounding. In this work, we are focussing on street view images shared publicly by Google[3] to extract the required information as satellite images have its limitation in terms of availability; and the resolution may not be sufficient for recognition of small scale objects. While street-view images provide human head view of the scene; recognizing objects which are on the ground is a difficult. As shown, recognition becomes more challenging if the texture/color of the surrounding region around the object also share similarity in appearance (Figure 9 illustrates that some of our objects are occluded and object classes look very similar). While many promising image based asset management systems are available, a principled approach for inventory management of business objects which applies state-of-the-art vision and machine learning techniques including deep features with Faster RCNN, and image triangulation based asset localisation has not been explored.

[3]https://www.google.co.in/maps/streetview/

Figure 4. On each route where GSV images are available, there are predefined viewpoints marked as blue circles above to create a panoramic spherical view of the locality. These images downloaded at each location are stored with unique identifiers describing the field of view and geo-location tag.

3. Proposed Method

Our industry partner specified following assets of interest which are basically the street furnitures: cabinets, poles, manholes, and jointboxes. Figure 2 shows the representative samples of the objects to be recognised on GSV. Figure 3 shows the pipeline for vision based inventory management used in this scenario. The process has four blocks that are explained in the following subsections.

3.1. Image Aquisition

In Figure 3, Step-1 involves the data collection for the purpose of inventory management. Using the street view API, we can download images corresponding to given GPS co-ordinates [4]. Figure 4 describes the setting how street images are collected for creating the GSV panorama.

These images can be accessed by Google Maps static API through an HTTP request. With the commercial agreement with our industrial partner, we have access to GSV image collection of resolution 2048×2048 pixels. Subfigures (a) and (b) in the figure 4 show a specified buffer zone where the download tool will navigate through each viewpoints at a given location. We assume that images corresponding to two consecutive GPS locations in the collection have finite degree of scene overlap from different camera angles which could be varying at different locations as described in the Figure 5.

3.2. Object detection Methodology

In the Step-2, we refer to the real problem of identifying the telecom assets in the collection of street view image. We are focussing on Joint Boxes, Manholes, cabinets and poles. These assets are civil inventories and have no IPs associated with them for remote tracking. Joint Boxes and Manholes are underground structures housing cable joints and/or equipments which can be accessed via a surface cover. Joint boxes and manholes can be situated in the footpath, verge or carriageway. Special cover variations exist for paved precincts and pedestrianized areas. The former two, cabinets and manholes are over ground assets. We formulate the task as detection and recognition problem, where we predict the bounding box containing the asset of interest and the corresponding asset type. Each bounding box should be output with an associated real-valued confidence of the detection which we use to plot the precision-recall curve which can be used for selection of the threshold for final labelling. The obvious choice for automation of asset detection are through object recognition methods which are typically governed by shallow or the hand crafted features and deep learning/feature learning. These methods can be trained to perform detection of assets of interest in the captured data. In this work, we have experimented with three robust object detection and recognition algorithms namely, HOG with linear SVM [5], Deformable Parts Model with linear SVM [8], and Faster RCNN [18] that have achieved state-of-the-art performance. The application specific parameter selection in these algorithms for the present context is discussed in Section 4.1.

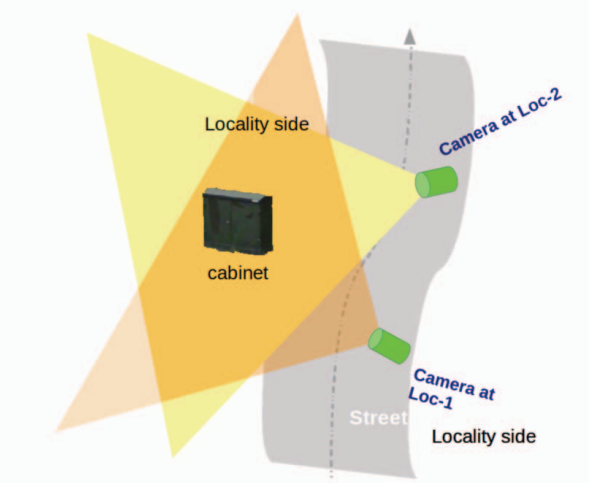


Figure 5. Scene overlap between street view images from different locations. Illustration shows Cabinet is captured from both camera location 1 and 2 with significant overlap. This is a realistic assumption as GSV images make use of 9 omni directional cameras.

[4]The parameters to be supplied for Google Street API are (a) Location (lat/lon value), (b) Size (image size), (c) Heading (0-360○), (d) Horizontal Field of View, (e) Pitch, and (f) up/down angle of the camera relative to the GSV vehicle

3.2.1 HOG detector with Linear SVM

The HoG approach [5] presents rigid template based approach for object recognition at various scales. The detector filters the image using fixed size template window by scanning in overlapping fashion at all positions and scales. Each template window is represented by distribution of orientations weighted by gradient magnitude. The approach models each asset category as a foreground object. Therefore, we learn four separate detectors one for each asset type.

3.2.2 Deformable Parts based model with Linear SVM

The Deformable Parts Model (DPM) has recently emerged as a very useful and popular tool for tackling the intra-category diversity problem in object detection. Given that our objects are taken from different viewpoints with variations in color, pose and illumination - the choice of using DPM is obivious.

The DPM has following salient points in comparison with the HOG: Latent discriminative learning and the idea of multiple components with objects (subcategories). The idea behind deformable parts is to represent an object model using a lower-resolution root template, and a set of spatially flexible high-resolution part templates. Each part captures local appearance properties of an object, and the deformations are characterized by links connecting them. Latent discriminative learning involves an iterative procedure that alternates the parameter estimation step between the known variables (e.g., bounding box location of instances) and the unknown i.e., latent variables (e.g., object part locations, instance-component membership). Finally, the idea of subcategories is to segregate object instances into disjoint groups each with a simple (possibly semantically interpretable) theme e.g., frontal vs profile view, or sitting vs standard person, etc, and then learning a model per object/asset type.

3.2.3 Faster RCNN (RCNN with RPN)

In the recent computer vision research, deep learning based methods have made significant stride surpassing earlier techniques which were learned using handcrafted features. R-CNN proposed by Girshick et al. [11] is a state-of-the-art visual object detection system that combines bottom-up region proposals with deep feature representations learned by a convolutional neural network. We are using Faster RCNN which is a combination of Region Proposal Networks (RPN) and Fast RCNN [18]. An RPN is a fully convolutional network that predicts both object bounding boxes and objectness scores (confidence scores) at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast RCNN for detection. RPNs are designed to efficiently predict region proposals with a wide range of scales and aspect ratios. This approach directly solves a $(n+1)$-category problem i.e. labelling all proposals detected by RPN in n pre-defined classes, and the background is modelled as separate class.

3.3. Estimating object coordinates from camera co-ordinates

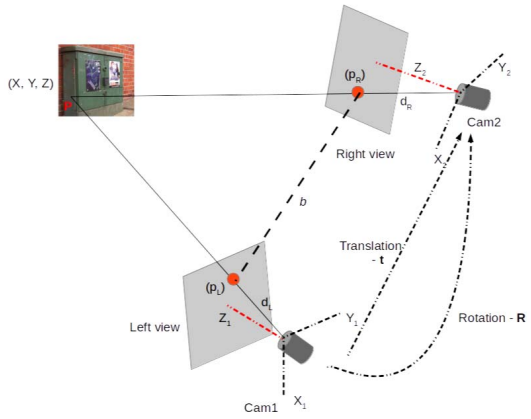


Figure 6. Distance estimation using stereo-vision

The street view images annotated with asset labels are further processed for asset location identification using triangulation referred in the Step-3. GSV image collections have associated GPS co-ordinates with each captured image. However, the real world position of objects will be different from the GPS coordinates of camera coordinates. Assuming the condition of scene overlap as shown in the figure 5, we apply triangulation on pair of images from two consecutive locations on the street/locality [14]. We get the estimate of real-world location of the asset object which can be used for verifying and updating the inventory records as described in the Step-4.

Figure 6 shows the experimental setting in this case. Cam_1 and Cam_2 are the street-view camera position at two different locations. Consider a point **P** on the object described by (X, Y, Z) in the word co-ordinate system. Point **P** is captured at position p_L in the image plane of the Cam_1, and at position p_r in the image plane of Cam_2. For estimating the depth from two image views, there are two main problems.

- Identifying the point correspondeces: We assume the available images have rectified beforehand to remove projective distortions. We assume the general setting where camera's intrinsic and extrinsic parameters are not known. We begin with identifying the regions of images which are similar in both the views. For points p_l and p_r in different views, there are two rays in 3D space connecting with cameras center of projection at

different locations. A simple approach to estimate the distance of point **P** is to find the 3D point $\mathbf{P}_a$ that lies closest to 3D rays corresponding to the matching feature locations $\{\mathbf{p}\}$ from bothe views. We propose SIFT based image matching [17] for computing this set. There are other methods available in the literature, nevertheless, SIFT based matching applies set of local feature vectors computed on pixel gradients. The approach has shown superior performance in many recent applications due to scale and illumination invariance and also returns partially matched regions.

- Points p_l and p_r on different image planes are related as

$$d_r \hat{p}_r = \mathbf{R}\,(d_l \hat{p}_l) + \mathbf{t} \tag{1}$$

Here $\hat{p}_l = \mathbf{K}^{-1} p_l$ and $\hat{p}_r = \mathbf{K}^{-1} p_r$ are ray direction vectors connecting **P** to projection centers of Cam_1 and Cam_2. **K** represents the camera caliberation parameter, and **R** and **t** are the rotation matrix and translation vector between two camera positions. The simplification of equation 1 gives the following condition which is defined as *epipolar constraint* in the literature.

$$\hat{p}_r^T\ \mathbf{E}\ \hat{p}_l = 0 \tag{2}$$

E is defined as the *essential matrix* computed as cross product of **t** and **R**. Equation 2 can be rewritten as

$$p_r{}^T\ \mathbf{F}\ p_l = 0 \tag{3}$$

Here **F** is defined as the fundamental matrix. For a set of n matching points in **p**, we have n homogenous equations such as for i^{th} match is as follows

$$\begin{bmatrix} p_{li} & q_{li} & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} p_{ri} \\ q_{ri} \\ 1 \end{bmatrix} = 0 \tag{4}$$

- Equation 4 can be solved using normalized 8-point algorithm [13]. The factorization of **F** as equation 5 returns the epipole vector **e** and corresponding homography **H**.

$$F = [\mathbf{e}]_\times \mathbf{H} \tag{5}$$

The factorization is not unique and can be obtained by different methods including direct approach, and SVD []. With the valid homography **H**, we can compute projection matrices P_o and P_o' as

$$P_o = [\mathbf{I} \mid \mathbf{0}],\ \text{and}\ P_o' = [\mathbf{e} \mid \mathbf{H}] \tag{6}$$

Matrix P_o and P_o' deine the relationship between real-world point **P** and its projections on images planes of Cam_1 and Cam_2 as

$$p_l\ =\ P_o\mathbf{P},\ p_r\ =\ P_o'\mathbf{P} \tag{7}$$

The above equation can be solved to estimate the location of **P**. Using the retrieved co-ordinates, and GPS values of camera positions, we can estimate the actual object position with reasonable accuracy.

3.4. Data Correction

The asset location identified in the street-view will now be transferred to data correction tool. A marker gets placed at the inventory location in street-view similar to GSV pegs. The marker is expected to detail the data correction process. The correction tool is updated with generic details such as latitude, longitude, asset type, street name, and so on.

4. Validation Results

For evaluation, we use rules mentioned in PASCAL VOC criteria [7] for object recognition to determine the overlap criteria between the ground truth object and algorithm detected object is considered. Also, precision-recall curves for 3 object detection algorithms investigated are reported (HoG detector with linear SVM, deformable parts based models with Linear SVM, and Faster RCNN with RPN for generating object proposals). PR curves are chosen as $N_{neg} \gg N_{pos}$, where N_{neg}, N_{pos} are number of negatives and positives in real scenarios. We also report F_1 score that summarizes the performance of the precision-recall curves in a single number. Table 1, shows P and R values at highest F_1 score. We also present the confusion matrix heat map that suggests the number of confusing classes.

4.1. Dataset and Experimental Setup

The validation dataset used for experiements comprises of approx. 3000 GSV images. For the four object classes that we are focussing in this paper, we had 900 positive examples. The experiments were performed in 3-fold cross validation setting and the average results are reported in the paper. These images were manually annotated for performance computation. The annotations for bounding box followed the PASCAL VOC format using the xml files to train a detector. We have used HOG with linear SVM implementation available in VLFeat open-source library [19]. For all asset types, we have used common parameter setting as described below.

- Neighbourhood for histogram computation: $hogCellSize$ - 8×8 pixels.
- With the $hogCellSize$ fixed as above, the detector searches for the foreground object in the testing image at multiple scales: the image is resized by 15 different scale factors spread between 2^{-1} to 2^3 on the logarithmic scale.
- Linear SVM parameters: λ = 0.00333333, C = 1 with hinge loss function and maximum number of iterations

- 3000.
- Optimization using Stochastic Dual Coordinate Ascent (SDCA) method.

For DPM detector, the default parameters are used except for setting number of components = 6 in the case of DPM detector. The implementation[5] made available by the authors of [8] were used for experiments.

We have chosen RCNN with RPN [11] [18] for our object detection. We show that this algorithm when used with ZF model produces outstanding results despite the complexity of the dataset as we have discussed in the Section 4.

4.2. Precision-Recall (PR) Curve

Figure 7 shows the PR curve for the detection performance from different techniques. Also the results have summarized in the Table 1. Its obvious that that faster RCNN provides best precision, recall and also F_1 score, the factors used to evaluate the performance of a detector. This is in comparison to HoG and DPM detectors. Note the steep decline in case of HoG and DPM unlike the Faster RCNN detector.

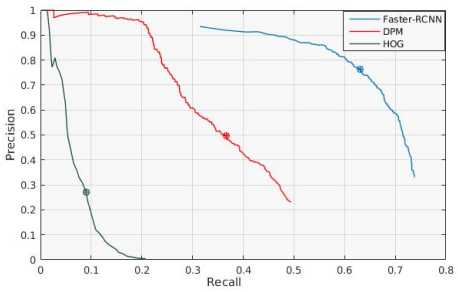


Figure 7. PR curve for proposed method Vs. the other state of the art detectors for object detection. The optimal PR values corresponding to maximum F1 score are highlighted.

Table 1. Precision and Recall values at highest F_1 Score

Method	Precision	Recall	F_1 Score
HoG with Linear SVM	0.2711	0.0896	0.1347
DPM	0.4983	0.3664	0.4233
Faster RCNN	0.7633	0.6312	0.6910

4.3. Confusion Matrix

The figure 8 shows the confusion matrix for Faster RCNN descriptor when the maximal F_1 score is achieved. The confusion matrix heat map plot depicts that the Cabinets, Jointbox and manholes have been efficiently detected as we observe along the diagonal. Also, the cabinets and

[5]https://github.com/rbgirshick/voc-dpm

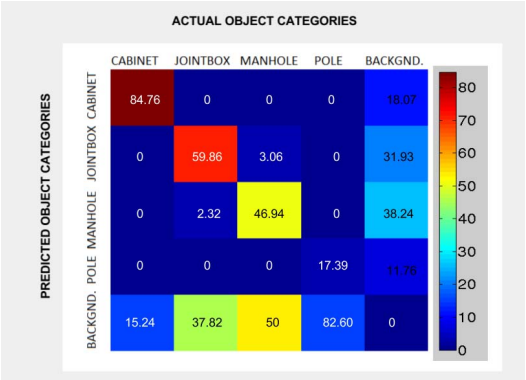


Figure 8. Confusion Matrix Heat Map: This figure depicts that Cabinets,jointboxes and manholes have been efficiently detected as we observe along the diagonal.

jointbox are efficiently detected as obvious from the color bar. In the last row, we refer to the detection's which were missed out as background by the proposal detection by RPN. Similarly, the last column refers to the removal of detected proposals which included some false positives i.e. backgrounds, by correct labelling by RCNN. These values are remarkably lower than the corresponding entry in the dominant diagonal of the matrix (we are considering only top four values) which establish the efficacy of Faster-RCNN algorithm in present application scenario. While Joint-box and manholes have also been detected reasonably well, some cases show confusion for some manholes were categorized as jointboxes and some manholes categorized as joinboxes as both of which are underground assets and lie on the footpath and visually very similar as seen the third row of Figure 9.

Fig 9 illustrates sample detections using HoG, DPM, and Faster RCNN. While the best performing method is faster RCNN, It can be noted that performance of DPM detector is superior to HoG. Nevertheless, the faster RCNN also figures out the side view of cabinet that is missed out during ground truth creation.

4.4. Object Localization Performance

We evaluated the proposed asset localization pipeline shown in Figure 3 for 10 image-pairs collected from our office promises, the image pairs comprised of manholes, cabinets and joint-boxes as the object of interest. Each image-pair captured the same scene from two different two different views. The RCNN based detector trained on the data described in the Section 4.1 returned us 100% accuracy. We achieved an average of $\approx 70\%$ accuracy in-terms of the actual distance measurement of the objects from the left-view camera position. (The accuracy of localisation = calculated distance via triangulation/ Measured distance or the Ground Truth distance *100)

The approach discussed in the Section 3.3 is based on

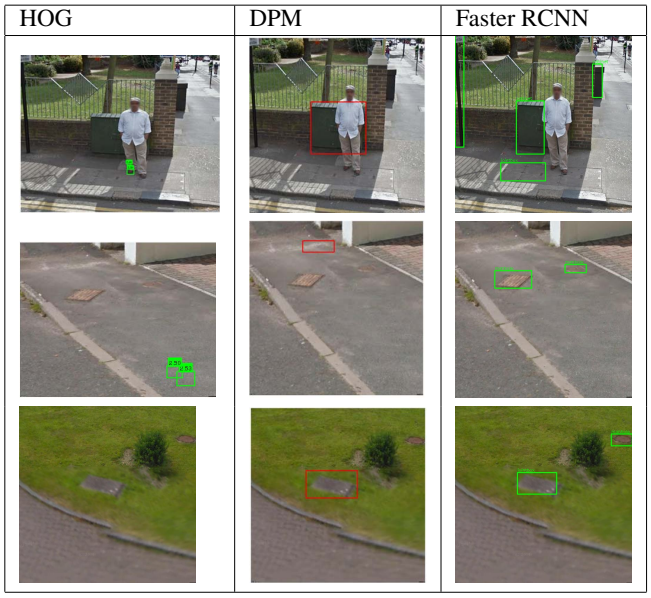


Figure 9. Illustration of sample detections of cabinets using HOG (shown in first column), DPM (shown in second column) and faster RCNN (shown in third column) detectors. First row depicts cabinet detection, second for manhole and third for both jointbox and manhole.

the standard projective reconstruction theory, however the accuracy of this method mainly depends on the point correspondence establishment and the selection of valid homography. While investigating the average performance of localization algorithm, we observed issues with SIFT based feature matching because of the outdoor setting returning only few local key-points. The selection of valid homography is based on empirical analysis of the $dist(\mathbf{H}\hat{p}_l, \hat{p}_r) \leq \epsilon$. The accuracy of $\mathbf{H}$ can be improved by applying the background knowledge of the environment, and with accurate measurement of $dist$ which should be invariant to the object positioning in background or foreground.

5. Discussion

We have evaluated our proposed pipeline for object localization using 3D for a set of image pairs. The results have discussed in the Section 4.4. We have achieved an average of 70% accuracy in localizing the object in terms of the actual distance with respect to the reference. While our approach for triangulation is based on conventional projective geometry, the accuracy of this method depends on feature matching. In the present case we have used SIFT, nevertheless, both the images for a scene have limited overlap resulting in only few matching points. We restricted the limited overlap constraint on the experimental image pairs as we also observed in the real-life scenario where google street view images also less overlap for a scene from two view-points. The conventional SFM based methods can significantly improve this performance, however, would require sequence of frames capturing the scene. These results have verified by our industry partner as well have been endorsed. We believe that good localisation accuracy is enough for localisation of inventory in most cases as the proposed pipeline could recover the missing inventories as humans missed assets when a block level test run was done.

It can be noted that localization may fail in situations when we don't have another view of same object or the scenario where images for an object have very less overlap, and if the object is partially/fully occluded in real world scenarios. Faster RCNN detector can be trained with few images as long as the training images cover diverse view points, various degrees of occlusions, different illuminations as in the real testing data. The model is cross validated for unbiased recognition and generalization. Our application is for a demo purpose where training is done on 3000 images with diverse views of objects. We are planning to deploy a much more robust model with more data to further improve the accuracy of the pipeline.

6. Conclusions and Future Work

An automatic asset management system using multi object recognition and localisation using Google street images has been presented with an use-case of telecom industry using GSV images. Our method utilizes computer vision and machine learning strategies to (a) discover the missing assets through asset recognition and localisation pipeline which was hitherto done through traditional survey process (b) update new assets. The pipeline is not only intended to reduce the use of manpower, but also reduced the overall survey costs. We have investigated both shallow and deep learning-based computer vision methods that give promising results for detection and triangulation for object localisation in real-world. Faster RCNN was found to give the highest precision/recall as we show in the results. While our method is yet to be deployed by our industry partner, the prototype functions well at a block level when tested on the unseen dataset at various daylight conditions and occlusions. At a solution level, we demonstrate the potential of leveraging GSV images as an economically viable solution for creating up-to-date inventories of street furnitures. The solution can be implemented for any of the following industry: power and utilities, telecom, transport and logistics etc.The generic implementation across multiple use-cases will add revenue for other business units for our industry partner. The data updation and correction tool is currently a concept that is tested as blocks of the pipeline and needs to be implemented by the industry partner.

References

[1] V. Balali, E. Depwe, and M. Golparvar-Fard. Multi-class traffic sign detection and classification using google street view images. In *Transportation Research Board 94th Annual Meeting, Transportation Research Board, Washington, DC*, 2015.

[2] A. Barbosa, J. Fernandes, and L. David. Key issues for sustainable urban stormwater management. *Water research*, 46(20):6787–6798, 2012.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[4] P. Bolstad. *GIS Fundamentals: A First Text on Geographic Information Systems.* Eider Press, 2005.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] A. Desai, D.-J. Lee, and D. Ventura. An efficient feature descriptor based on synthetic basis functions and uniqueness matching strategy. *Computer Vision and Image Understanding*, 142:37–49, 2016.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[9] A. Flores and S. Belongie. Removing pedestrians from google street view images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 53–58. IEEE, 2010.

[10] S. Fowers, A. Desai, D.-J. Lee, D. Ventura, and J. Archibald. Treebasis feature descriptor and its hardware implementation. *International Journal of Reconfigurable Computing*, 2014:12, 2014.

[11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[12] H. Guler, M. Akad, and M. Ergun. Railway asset management system in turkey: A gis application. In *FIG working week*, pages 22–27, 2004.

[13] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, Jun 1997.

[14] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[15] D. Hodges, R. Wright, G. Hagan, and J. Lockhart. Systems and methods for the detection and management of network assets, July 10 2007. US Patent 7,243,147.

[16] Z. Hu and Y. Tsai. Generalized image recognition algorithm for sign inventory. *Journal of Computing in Civil Engineering*, 25(2):149–158, 2011.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[19] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[20] A. C. Vibrans, R. E. McRoberts, P. Moser, and A. L. Nicoletti. Using satellite image-based maps and ground inventory data to estimate the area of the remaining atlantic forest in the brazilian state of santa catarina. *Remote Sensing of Environment*, 130:87–95, 2013.

[21] T. Warsop and S. Singh. A survey of object recognition methods for automatic asset detection in high-definition video. In *Cybernetic Intelligent Systems (CIS), 2010 IEEE 9th International Conference on*, pages 1–6, Sept 2010.