EasyChair Preprint

№ 9589

Determining Image Scale in Real-World Units Using Natural Objects Present in Image

Saurabh Singh and Rhea S Shrivastava

January 18, 2023

Chapter 3

Determining Image Scale In Real-World Units Using Natural Objects Present In Image

Saurabh Singh[1*], Rhea S Shrivastava[1]

[1]Amity School of Engineering and Technology, Amity University, India

**Corresponding Author: saurabh.iiet@gmail.com*

ABSTRACT

The photograph is a 2D representation of intensity values in columns and rows that is stored in a digital computer. Defining the physical size of article in actual world units (e.g., mm, cm) is a challenging task. Some images have a real-world unit scale that helps the individuals to assess the size of the objects present in the image. This scale is introduced at the acquisition time of the picture by the metal ware infrastructure. Inopportunely, the scale is not present in a lot of visual descriptions (images), and it poses a problem to define the genuine size of the objects captured in an illustration. Thus, determination of size of objects in an image became the main goal of this research proposal.

The reason that this segment of study is yet to be explored to its fullest was the incentive behind the work. In this research endeavour, our proposition is to find image scale (size in real world unit per pixel) using the common size of the objects existing in the visual description. These regular objects could be people, cars, bikes, signposts etc., depending upon the location, traffic and time when the image was taken. The dataset which is used for the investigation is attained from KITTI and the RGB-D images of the dataset have been taken into account. It is then wielded in derivation of the mathematical function to co-relate depth of the concerned entity with object in real world unit per pixel. In this way, the desired outcome is achieved.

3.1 INTRODUCTION

An image comprises rows and columns and is stored in a digital computer in a digitized arrangement. The picture elements are an important aspect of a photograph as it opens the path for numerous dimensions of research. The pixels help in understanding the image and also provide a way to interpret it. By the means of existing tools and image refining techniques, an image can be enhanced, filtered, cropped, stretched, sharpened etc., and be studied in various forms.

An image is attained by a camcorder in the proximity of sunlight or any varied source of illumination, followed by pre-processing techniques which varies relaying upon the individual or the problem

account. The illustration is of miscellaneous kinds. The types of images that this account has closely dealt with are monocular and stereo images.

Fig 3.1: Image

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Figure 3.2: Image Representation

Monocular image incorporates a single image whereas stereo images deal with two images. The contradiction between these 2 is that the monocular vision ascertains articles which are categorized like cars, pedestrians whereas stereo vision detects even those objects which are yet to be categorized. It takes every object which appears in the field of vision into consideration, unlike mono-vision. When mono-vision is considered, it heavily relies on what it has been skilled or experienced on. These entities are grouped into the predefined classes which it was trained on. Whenever a new object pops in its field of vision, it simply ignores it.

This approach leads to many mistakes, some of which have been fatal. A common example of this mistake is the cars manufactured with mono-vision system. On the other hand, stereo vision overcomes this shortcoming very easily. It uses the inherent 3D capabilities and views all the entities that are in the field of vision. It has a broad field of sight as well and takes appropriate measures and avoids casualty. Therefore, stereo images are commonly used and commercially preferred as well. The table given below elaborates the points more clearly.

TABLE 3.1: Mono Camera System vs. Stereo Camera System.[18]

Comparison Parameter	**Mono Camera System**	**Stereo Camera System**
Number of image sensors, lenses and assembly	1	2
Physical size of the system	Small (6” x 4” x 1”)	Two small assemblies separated by ~25-30cm distance
Frame Rate	30 to 60 frame/second	30 frames/second
Image Processing Requirements	Medium	High
Reliability of detecting obstacles and emergency braking decisions	Medium	High
System is Reliable for	Object detection (lanes, pedestrians, traffic signs)	Object detection “AND” calculate distance to object
System Cost	1X	1.5X
Software and Algorithm Complexity	High	Medium

The next important aspect is the size of the objects. It determines the real world characteristics of the entities and in turn assists the individuals to understand the image well. The height of the article, expanse from the camcorder and size in physical world unit all helps in understanding the image.

The size assessment aspect has been majorly used in [3, 16, 17, 29, 31, 35] which is of utmost importance for this work. It has helped in understanding the appropriate time for harvest of fruits and vegetables [3, 16], for precise calorie intake [17, 29], for dietary estimation [31] and food serving size assessment [35]. It is being widely used in various categories and has ever since proven helpful.

The novelty of the work is determination of scale of an image. Size estimation is still not widely explored and through this work it will be highlighted and experimentally scrutinized for achieving the motive of the work. Therefore, determining size estimation of articles in an image will surely help in producing the scale in relation to the physical world which will assist in understanding ambiguous descriptions with more confidence.

3.2 LITERATURE REVIEW

According to Zhao et al. [1] Depth information is predominant for autonomous systems in order to perceive the atmosphere and to proximate the phase. Conventional depth assessment methods are based on feature correspondence from several perspectives. Meanwhile, depth maps predicted by

research are rare thus inferring depth information from an image (i.e. estimating monocular depth) a serious problem. The deep learning-assisted monocular depth estimation has been extensively studied. In terms of precision, it has already performed very promisingly. In the meantime, the dense depth maps from individual images are estimated by deep neural network in a format that various forms of network frames, loss functions and strategies are then proposed to increase the accuracy of the depth estimation. Therefore, during this review, the researchers examined these deep learning-assisted monocular depth estimation methods.

According to Fernandes et al. [2] The research paper provides an accurate process for computing the scale of the boxes directly from perspective projection images acquired by conventional cameras. The approach is dependent on projective geometry and calculates the dimensions of the box by using the data. The facts, figures and data is extracted from the silhouette of the box with the prediction of two parallel laser beams on one of the faces of the box. To identify the silhouette of the object, a statistical model is developed for homogeneous background colour removal that works with a moving camera. A voting scheme for the Hough transform is incorporated in the archetype that ascertains the collinear pixels' groups. The efficiency of the proposed approach is reflected when the dimensions of real boxes are calculated using a scanner prototype that implements the algorithms and the approaches described within the paper.

According to Wang et al.[3] In-field mango fruit sizing is beneficial for estimating the fruit maturation and size distribution, as it suggests harvesting, produce resourcing (e.g., tray insert sizes), and marketing. In-field, machine vision imaging is used for fruit count, and now it is being used for the analysis of fruit sizes from images. The low-cost examples of 3 technologies for estimating the distance of a camera to fruit is assessed. The RGB-D camera is used due to cost and performance sanctioned, but it operated poorly under direct sunlight. For detection of fruits, a cascade detection with the histogram of oriented gradients (HOG) feature is used, then Otsu's method, followed by colour thresholding application to get rid of the background objects. Finally, fruit lineal dimensions are premeditated using the RGB-D depth statistics, fruit image extent and the thin lens formulation. The authors believe that this work signifies the leading practical execution of machine vision fruit sizing infield, with realism evaluated in terms of capital and ease of operation.

According to Standley et al.[4] A successful robotic influence of real-world entities necessitates thoughtful adaptability of the physical properties of these objects. The authors have proposed a model which estimates, pass, from the image of an object. They have compared a variety of baseline replicas for an image-to-mass problem then they were trained on this dataset. The authors have also characterized the performance of a human on this problem. Finally, a model is presented which interprets the 3D shape of the object.

According to Zhu et al.[6] In this work, the advantages of two common computer vision tasks: semantic segmentation and self-supervised depth estimation from images is studied. The authors proposed a technique to measure the regularity of border explicitly between depth and segmentation and diminish it greedily by iteratively administering the network towards a sectional optimal solution. Through expansive examinations, this suggested approach advanced to the state of the art on unsupervised monocular depth estimation on the KITTI dataset.

According to Liu et al.[7] In this paper, the matter of depth approximation from solitary monocular images is addressed. Estimating depth in monocular images is challenging as compared to estimating depth using numerous images. Earlier work focuses on making the most use of additional sources of data. The authors have put forward a deep architecture learning strategy that acquires the pairwise and unary potentials of continuous Conditional Random Field in a amalgamated deep CNN framework. Then further the proposal of an equally effective model based on fully convolutional networks and a novel super pixel pooling method is made. This method is used for estimating basic scenes with no geometric priors or any extra information injected.

According to Ma et al.[9] The authors have considered the prediction of dense depth from a sparse set of dimensions of depth and an RGB image. Since the estimation of depth from monocular visual descriptions is intrinsically ambiguous and unreliable, to achieve a better level of robustness and accuracy, additional sparse depth samples were introduced, which are obtained by visual Simultaneous Localization and Mapping (SLAM) algorithms. The experiments conducted by the team highlights that in comparison to the usage of RGB images, the addition of 100 spatially random depth samples reduces the prediction of root-mean-square error (RMSE) by 50% on the NYU-Depth-v2 indoor dataset. It boosts the section of reliable prediction from 59% to 92% on the KITTI dataset.

According to Li et al. [13], Self-supervised depth estimation has shown significant possibilities in hypothesizing 3D structures using entirely un-annotated images. However, the performance significantly decreases when the trained on images with varying brightness /moving objects. In this paper, the academicians have accredited this matter by intensifying the strength of the self-supervised criterion using a set of image-based/geometry-based constraints. Firstly, a gradient-based robust photometric loss framework is proposed. Secondly, the irregular regions are cleared by leveraging the inter-loss consistency and the loss gradient consistency. Thirdly the motion estimation was repressed to generate across-frame consistent motions via proposing a triplet-based cycle consistency constraint. Expansive examinations directed on KITTI, Cityscape and Make3D datasets indicated the supremacy of this method.

According to Eigen et al.[15], Calculating depth is an indispensable constituent in realizing the 3D geometry depicted in a certain scene. Furthermore, the chore is characteristically abstruse, with a great source of indecision from the inclusive scale. In this paper, the journalists have conferred a novel

technique that discusses this assignment by retaining two deep network stacks: making an abrasive universal estimation established on the grounds of entire image and additional that enhances this prediction locally.

According to Ponce et al.[16], Fruit grading is a necessary post-harvest chore in the olive industry. The combination is based on size, assists and mass in the processing of high-quality table olives. The study presents a technique fixated on assisting olive grading by computer vision techniques and feature modelling. The sum total of 3600 olive fruits from nine variations was captured, stochastically distributing the individuals on the scene, using an improvised imaging chamber. Then, an image analysis algorithm was invented to divide olives and extract descriptive characteristics to evaluate their minor and major axes; and their mass. Determining the accurate performance for the individual division of the olive fruits, the algorithm was proven through 117 captures containing 11606 fruits, producing only six fruit-segmentation mistakes.

According to Ege et al.[17], Analysis of the estimation of images of food for precise food calorie estimation comprises three prevailing techniques with two new methods: (1) CalorieCam — It evaluates real food size on the foundation of a reference object, (2) Region Segmentation — It employs food calorie estimation, (3) AR DeepCalorieCamV2 — It is about visual-inertial odometry built-in the iOS ARKit library, (4) DepthCalorieCam — It is based on camera (stereo) such as iPhone X/XS, and (5) RiceCalorieCam — It utilizes rice grains as reference articles. The last two approaches attained 10% or less estimation error, which is good for estimating food calorie.

3.2 METHODOLOGY

The proposed methodology adopted is described in Figure 3.1:

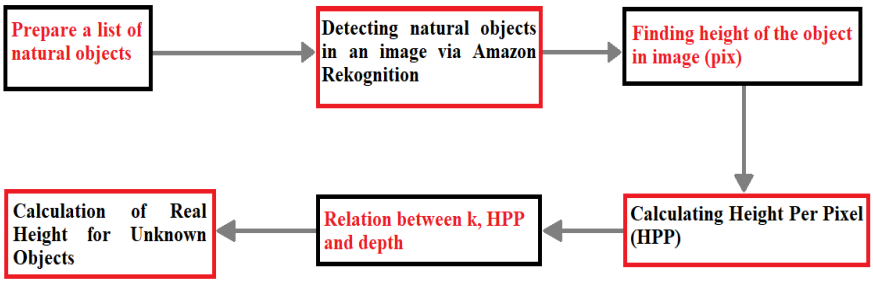

Figure 3.3: Proposed Methodology

3.2.1 Dataset

The dataset has been attained by Karlsruhe Institute of Technology. They have used autonomous driving platform known as Annieway for the development of unique and inspiring real-world computer vision benchmarks. Annieway is equipped with two high-resolution colour and grayscale video cameras on its hood. The accurate and precise ground truth is supplied by a Velodyne laser

scanner and a GPS localization system. All the datasets are captured by driving around the mid-size city of Karlsruhe, Germany.

Figure 3.4: Autonomous Driving Platform for KITTI

The info set used for this work is Data Scene Flow 2015 / Stereo 2015 which has been downloaded from KITTI Dataset and consists of 200 training scenes and 200 test scenes (4 colour images per scene, saved in loss less PNG format). The estimation server calculates the proportion of bad pixels intermediated over all ground truth pixels of all 200 test images. In the dataset, the images contain up-to 15 cars and 30 pedestrians.

Sample imageries from the dataset are as follows:

Figure 3.5(i): Sample Image 1

Figure 3.5(ii): Sample Image 2

Figure 3.5(iii): Sample Image 3

3.2.2 Algorithm

1. Creation of table of natural objects in the image.
2. Detecting natural objects in an image.
3. Once the objects are detected the table is prepared.
4. The subsequent stage is to calculate HPP which is height per pixel. Its unit is cm/pix. This can be calculated by applying $= \frac{\boldsymbol{Real\ Height}}{\boldsymbol{No.\ of\ Pixels}}$.
5. *k* is calculated by applying the formula $\boldsymbol{k} = \frac{\boldsymbol{HPP}}{\boldsymbol{depth}}$.
6. Then $\boldsymbol{Real\ Height} = \boldsymbol{k} * \boldsymbol{Depth} * \boldsymbol{Pixel\ Height}$ is calculated which is Predicted Height.

3.2.3 Steps

1. **Preparing the list of natural objects:** The images were perused and the natural objects were noted. After that, the standard height of these objects was appropriated from reliable web sources like websites of construction companies, Wikipedia, car manufacturing websites [20][23][24][25]. For trees the average height was taken from research paper [21]. Then the list was processed along with the mean height of the objects.

TABLE 3.2 List of Natural Objects

Sr No.	Natural Objects	Mean Height (cm)
01	Train	402
02	Car	170
03	Person	172
04	Traffic Signal	210
05	Truck	280
06	Wheel	83

The mean height is taken in cm.

2. **Detecting natural objects in an image:** Amazon Rekognition tool is brought in usage to detect objects in the KITTI images. It is provided by AWS.

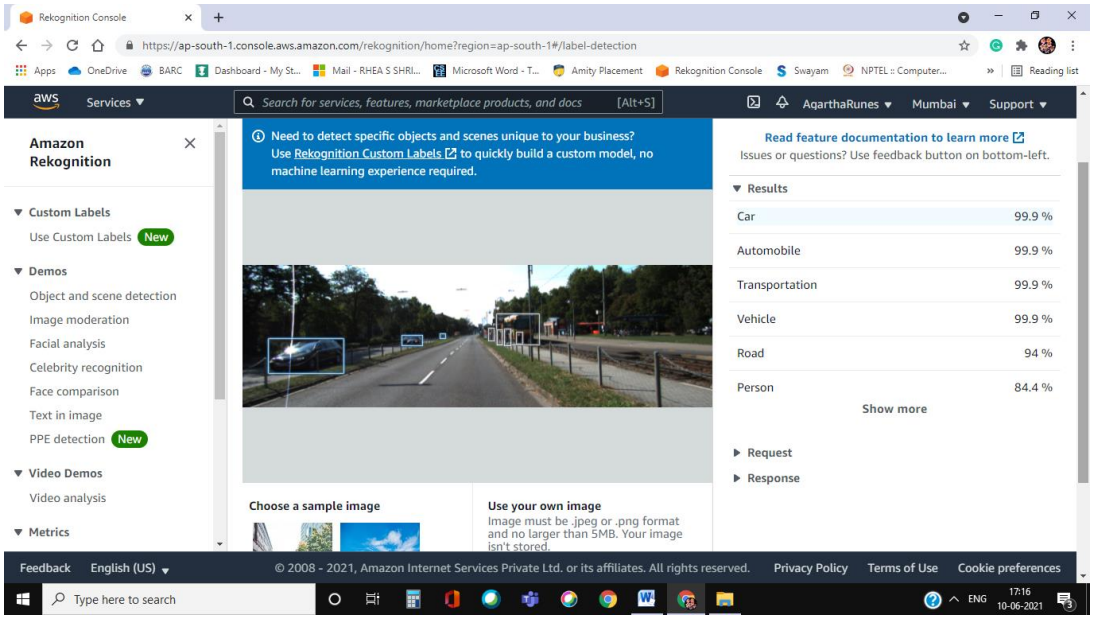

Figure 3.6: Object detection working on Image from dataset

From figure 3.6, it is evident that it is detecting articles predominant in KITTI dataset image. The bounding box appears on some items for determining the image scale.

3. **Finding height of the object in image (in pix):** This step is carried out by the help of bounding box. When AWS recognition is used, the responses includes numerical data pertaining to width, height, left and top[22, 32].

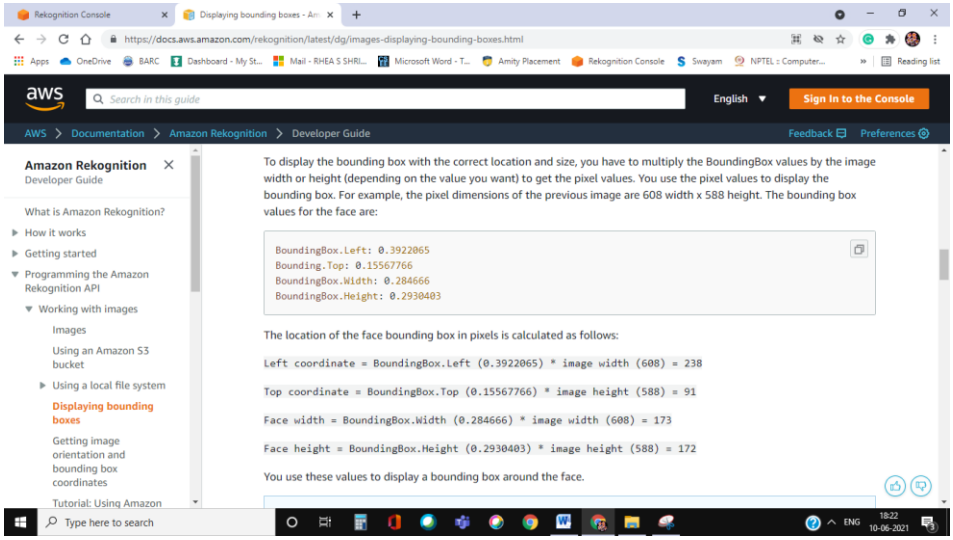

Figure 3.7: Snapshot of the AWS Developers Guide explaining Bounding Box calculations [22]

It is to procure the needed information. The bounding box returns the quantity of image elements to express an object. Only height will be considered for further calculations as it indicates the ratio of overall image height which is the requirement of this stage.


```
{
   "Labels": [
       {
           "Name": "Car",
           "Confidence": 99.9004135131836,
           "Instances": [
               {
                   "BoundingBox": {
                       "Width":
0.1577451378107071,
                       "Height":
0.2390778809785843,
                       "Left":
0.05361174792051315,
                       "Top":
0.5120320916175842
                   },
                   "Confidence":
99.9004135131836
               },
```


Figure 3.8: The response after Image (in Figure 3.5 (i)) was uploaded in Amazon Rekognition Object Detection

4. The real height of the object will remain same irrespective of its position in the iilustration which is understandable. The formula which is applicable in this step is as follows:

- $\boldsymbol{HPP = \frac{Real\ Height}{No.\ of\ Pixels}}$ **; - Equation 1,** Real Height is taken from the table created in Step 1.
- $\boldsymbol{HPP \propto \frac{1}{No.\ of\ Pixels}}$; This indicates that if HPP (height per pixel) is increased then no. of pixels will decrease and vice versa.

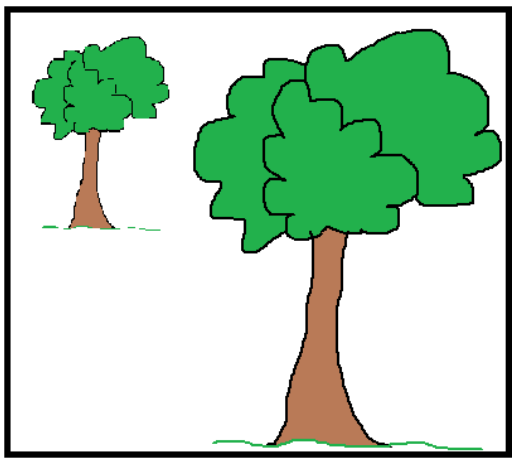

Figure 3.9: Understanding the real height of the object

Irrespective of the location of the article in the photograph, its real height or natural height remains same. The only modification is the pixel representation of the interested object. In the above figure

3.7, it's unmistakably noticeable that the closer the object is to the camera more number of pixels will be used to represent it, hence the Height Per Pixel will decrease and vice versa.

5. **Relation between k, HPP and depth:** Through step 4 and from Equation 1, we can say that,

- $\boldsymbol{HPP \propto depth}$ **- Equation 2**
- $\boldsymbol{HPP = k * depth}$; *k* is a constant which is to be calculated.
- $\boldsymbol{k = \frac{HPP}{depth}}$ **- Equation 3**

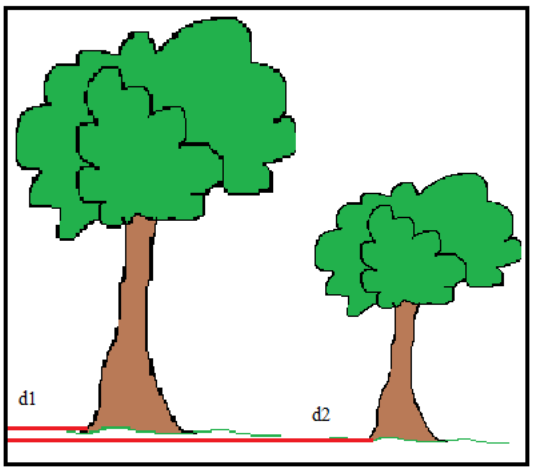

Figure 3.10: Pictorial representation of how depth is related to no. of pixels to represent an object in the image

According to Equation 2, Height per pixel will increase with increase in depth and will decrease with decrement in depth. This is done to evaluate *k* which is an important component of this whole research [37]. *k* remains constant throughout the image and is different for different images. Also, through availability of *k*, the calculation of the scale of unknown substances in the image has become easy.

6. **Calculation of Real Height for Unknown Objects:** The last step uses all the acquired details from above steps and substitutes them in Equation 4 to get the results.

$\boldsymbol{Real\ Height = k * Depth * Pixel\ Height}$ **- Equation 4**

In this equation, *k* attained for the entire image will be substituted along with the depth of the entity (step 5) and pixel height (step 3). This provides us with the desired outcome.

3.3 IMPLEMENTATION

Step 1:

First step to implement the methodology (ref. Figure 3.3) is to create a table (ref. Table 3.2) for natural objects with their mean height. This is thoroughly explored and the dataset being used is

from Karlsruhe, Germany. Therefore, the estimations and calculations have been done keeping this in the forefront of the research.

Step 2:

After the list was prepared, the next task was to identify the articles in the images. This was done with the assistance of Amazon Rekognition.

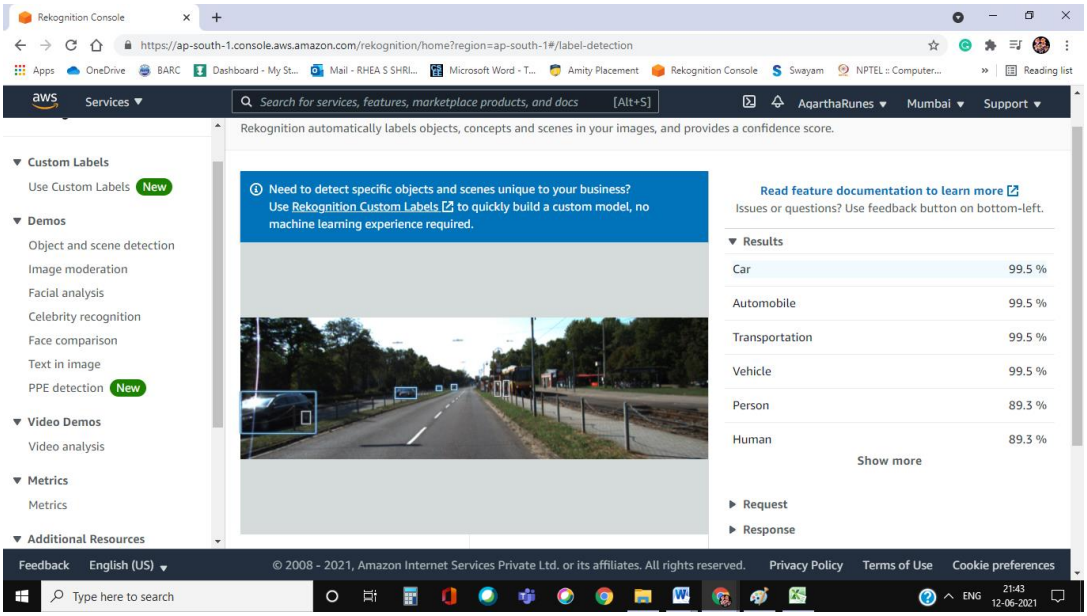

Figure 3.11: Amazon Rekognition recognising the objects in the KITTI dataset image

Step 3:

After the objects are detected the table is prepared using Microsoft Excel as follows:

ImageID	Natural Objects	Height Ratio	Dimension Of Image	Height Ratio x 375 (pix) pix value
000000_10	Car 1	0.1181	1242x375	44.2875
000000_10	Car 2	0.1261	1242x375	47.2875
000000_10	Car 3	0.0857	1242x375	32.1375
000000_11	Car 1	0.1416	1242x375	53.1
000000_11	Car 2	0.0937	1242x375	35.1375
000000_11	Car 3	0.1164	1242x375	43.65
000000_11	Car 4	0.1883	1242x375	70.6125
000000_11	Car 5	0.0586	1242x375	21.975
000003_10	Car 1	0.239	1242x375	89.625
000003_10	Car 2	0.0656	1242x375	24.6
000003_10	Car 3	0.0281	1242x375	10.5375
000003_10	Person 1	0.1136	1242x375	42.6
000003_10	Person 2	0.0994	1242x375	37.275
000003_10	Person 3	0.0851	1242x375	31.9125
000003_10	Person 4	0.077	1242x375	28.875
000003_10	Train	0.214	1242x375	80.25
000003_11	Car 1	0.2851	1242x375	106.9125
000003_11	Car 2	0.0742	1242x375	27.825
000003_11	Car 3	0.0273	1242x375	10.2375
000003_11	Car 4	0.0355	1242x375	13.3125
000003_11	Person 1	0.1025	1242x375	38.4375
000003_11	Person 2	0.1027	1242x375	38.5125
000003_11	Wheel	0.0827	1242x375	31.0125

Figure 3.12: Table depicting the values garnered from Amazon Rekognition Software

The column "*Natural Objects*" and "*Height Ratio*" has been generated by the results from Amazon's software. The "*Dimension Of Image*" is already known to us. The column with heading "*Height Ratio x 375 (pix) pix value*" is indicative of step 3. Here, the values are in pixel units: pix and define the extent of pixels that define an object restrained by bounding box.

Step 4:

The subsequent stage is to calculate HPP which is height per pixel. Its unit is cm/pix. This can be calculated by applying Equation 1. Once it's done the resultant column is achieved as follows:

HPP (cm/pix)
3.8385
3.595
5.2897
3.201
4.8381
3.8946
2.4075
7.736
1.8967
6.9105
16.1328
4.0375
4.6143
5.3897
5.9567
5.0093
1.59
6.1096
16.6056
12.7699
4.4747
4.466
2.6763

Figure 3.13: HPP column

Step 5:

Upon estimating HPP, we calculate *k*. This is done with the help of Equation 3. The depth is already given and the required details are substituted in the formulae. This effect in generation of column *k* which is showed in figure 3.14 below:

Depth (cm)	k=HPP/Depth	k_Mean
150	0.022666667	
300	0.023333333	
1500	0.024	
750	0.022933333	0.026966
80	0.0215	
50	0.022933333	
38	0.022631579	
40	0.052894737	
39	0.063173541	
375	0.004533333	
975	0.005128205	
1025	0.004738676	0.005826
375	0.008493827	
385	0.013139801	
160	0.014409722	

Figure 3.14: Depth and constant *k*

Step 6:

Then Equation 4 is used and the column "Predicted Real Height" is generated. These values are then paralleled with the natural heights of real world objects which provide insight as to how much the predicted height is varying from real height.

Pred H=k*depth*pixh	Actual Height (cm)
170	170
175	175
180	180
172	172
172	172
172	172
172	172
402	402
170	170
170	170
170	170
170	170
172	172
172	172
83	83

Figure 3.15: Natural Height vs. Predicted Real Height

3.4 RESULTS AND CONCLUSION

The attached screenshot below describes the requirements for calculation.

Image	Object	Actual Height (cm)	k=HPP/Depth	k_Mean	Depth (cm)	Pixel Height	Pred H=k*depth*pixh
000003_10	Car 1	170	0.022666667		150	50	170
000003_10	Car 2	175	0.023333333		300	25	175
000003_10	Car 3	180	0.024		1500	5	180
000003_10	Person 1	172	0.022933333	0.026966	750	10	172
000003_10	Person 2	172	0.0215		80	100	172
000003_10	Person 3	172	0.022933333		50	150	172
000003_10	Person 4	172	0.022631579		38	200	172
000003_10	Train	402	0.052894737		40	190	402
000003_11	Car 1	170	0.063173541		39	69	170
000003_11	Car 2	170	0.004533333		375	100	170
000003_11	Car 3	170	0.005128205		975	34	170
000003_11	Car 4	170	0.004738676	0.005826	1025	35	170
000003_11	Person 1	172	0.008493827		375	54	172
000003_11	Person 2	172	0.013139801		385	34	172
000003_11	Wheel	83	0.014409722		160	36	83

Figure 3.16: Necessary columns for Calculation

According to preceding steps *k* was calculated and its mean was taken. Now, for every image there is a constant *k.* This *k* when substituted in the formulae gives the desired outcome which is the Predicted Size. All articles in the photograph are at a certain depth from the camera's focal point so there's a requirement of taking mean depth which is illustrated in the figure below. The mean depth will aid in estimating the height of the unknown objects in the image.

Image	Mean Depth
000003_10	363.5
000003_11	476.28571

Figure 3.17: Mean Distance of Images

Lastly, Mean Squared Error of the images is calculated. The formula for MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y')^2$$

Where, MSE is Mean Squared Error, n = number of data points, y = Observed values and y′ = Predicted values. This results in the values as mentioned in Figure 3.18:

Image	K	Mean Depth	Mean Squared Error
000003_10	0.026966	363.5	0
000003_11	0.005826	476.2857143	0

Figure 3.18: Mean Squared Error of the Images

The data so far calculated helps in estimating the size of the unknown objects in the image. Further, calculation of MSE is carried out and it is depicted below. The predicted height is subtracted from the actual height.

Image	Object	Actual Height (cm)	Pred H=k*depth*pixh	Error = Actual - Predicted	Sq(Error)
000003_10	Car 1	170	170	0	0
000003_10	Car 2	175	175	0	0
000003_10	Car 3	180	180	0	0
000003_10	Person 1	172	172	0	0
000003_10	Person 2	172	172	0	0
000003_10	Person 3	172	172	0	0
000003_10	Person 4	172	172	0	0
000003_10	Train	402	402	0	0
000003_11	Car 1	170	170	0	0
000003_11	Car 2	170	170	0	0
000003_11	Car 3	170	170	0	0
000003_11	Car 4	170	170	0	0
000003_11	Person 1	172	172	0	0
000003_11	Person 2	172	172	0	0
000003_11	Wheel	83	83	0	0
				Mean Square Error	0

Figure 3.19: Calculation of MSE for all objects in images

It is understandable that MSE governs the correctness and accuracy of the model, system. It is a loss of calculates the amount of error. It is said that MSE and accuracy are inversely related to one another as MSE's ideal value 0 corresponds to utmost accuracy. According to Wikipedia [27], "The MSE is a measure of the quality of an estimator. As it is derived from the square of Euclidean distance is

constantly a positive value with the error decreasing as the miscalculation approaches zero." That can be seen from the image. Since MSE is 0, the algorithm works perfectly well with the dataset. Therefore, it can be said that the methodology and the algorithm are working with full efficiency and desired outcome is achieved. This study has been very beneficial as the visual descriptions without scale are now interpreted with ease which is helpful in understanding them, whatever may be the purpose. This is going to aid future researchers and their analysis in comprehending the true essence of varied photographs for experimentation. It is further going to open new horizons into undiscovered or less explored terrain of research.

3.5 FUTURE SCOPE

Size estimation of material world objects through images is a progressive research community and within a span of few years it is anticipated to experience extensive scrutiny. It is indicative of a breakthrough in a multitudinous of fields ranging from computer vision, image processing, artificial intelligence, military science, medical research, criminology, food technology, space research, oceanography, volcanology and many more. It has introduced a way for novel and generative ideas which results in progressive research. The scheme of ideas pertaining to future research entails carrying out the research on larger dataset and on a variety of images unlike the executed one. Considering the effort in this field has just begun there exists a colossal scope and potential in it. It will be truly expansive and skyrocketing years from now.

KEYWORDS

Image Processing
Machine Learning
Size Estimation
Size Determination
KITTI
Stereo Images
Natural Objects

References

1. Zhao, C, Sun, Q, Zhang, C, Tang, Y & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences 63*. 1612–1627. Retrieved from https://link.springer.com/article/10.1007/s11431-020-1582-8
2. Fernandes, L. A. F., Oliveira, M. M., Da Silva, R & Crespo, G. J. (2006). A fast and accurate approach for computing the dimensions of boxes from single perspective images. *Journal of the Brazilian Computer Society, 12*. 19-30. Retrieved from https://link.springer.com/article/10.1007/BF03192392

3. Wang, Z, Walsh, K. B. & Verma, B. (2017). On-Tree Mango Fruit Size Estimation Using RGB-D Images. *Sensors in Agriculture and Forestry*, *17*(12). 2738. doi:10.3390/s17122738
4. Standley, T, Sener, O, Chen, D & Savarese, S. (2017). image2mass: Estimating the Mass of an Object from its Image. *Proceedings of the 1st Annual Conference on Robot Learning, PMLR 78*, 324-333. Retrieved from http://proceedings.mlr.press/v78/standley17a.html
5. Ingvander, S, Brown, I. A., Jansson, P, Holmlund, P, Johansson, C & Rosqvist, G. (2018). Particle Size Sampling and Object-Oriented Image Analysis for Field Investigations of Snow Particle Size, Shape, and Distribution. *The Arctic, Antarctic, and Alpine Research, An Interdisciplinary Journal*, *45*(3), 330-341. Retrieved from https://doi.org/10.1657/1938-4246-45.3.330
6. Zhu, S, Brazil, G & Liu, X. (2020). The Edge of Depth: Explicit Constraints between Segmentation and Depth, presented at Seattle, WA, USA, 2020. IEEE.
7. Liu, F, Shen, C, Lin, G & Reid, I. (2016). Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *38*(10), 2024-2039. doi**:** 10.1109/TPAMI.2015.2505283
8. He, L, Wang, G & Hu, Z. (2018). Learning Depth From Single Images With Deep Neural Network Embedding Focal Length. *IEEE Transactions on Image Processing*, *27*(9), 99. doi: 10.1109/TIP.2018.2832296
9. Ma, F & Karaman, S. (2018). Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image, presented at IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 2018. IEEE.
10. Mahjourian, R, Wicke, M & Angelova, A. (2018). Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Retrieved from https://arxiv.org/abs/1802.05522
11. Wang, C, Buenaposada, J. M., Zhu, R & Lucey, S. (2018). Learning Depth from Monocular Videos using Direct Methods, presented at 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018. IEEE.
12. Cao, Y, Wu, Z & Shen, C. (2018). Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 11, 3174-3182. doi: 10.1109/TCSVT.2017.2740321
13. Li, R, He, X, Zhu, Y, Li, X, Sun, J & Zhan, Y. (2020). Enhancing Self-supervised Monocular Depth Estimation via Incorporating Robust Constraints. *Poster Session F2: Media*

Interpretation & Mobile Multimedia, 3108-2117. Retrieved from https://doi.org/10.1145/3394171.3413706

14. Godard, C, Aodha, O.M., Firman, M & Brostow, G. (2019). Digging Into Self-Supervised Monocular Depth Estimation, presented at IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019. IEEE.
15. Eigen, D, Puhrsch, C & Fergus, R. (2014). Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *Computer Vision and Pattern Recognition*, Vol. 2, 2366-2374. Retrieved from https://dl.acm.org/doi/10.5555/2969033.2969091
16. Ponce, J, Aquino, A & Millan, B. (2019). Automatic Counting and Individual Size and Mass Estimation of Olive-Fruits through Computer Vision Techniques. *IEEE Access*, Vol 7, 59451-59465.
    doi: 10.1109/ACCESS.2019.2915169
17. Ege, T, Ando, Y, Tanno, R, Shimoda W & Keiji Yanai. (2019). Image Based Estimation of Real Food Size for Accurate Food Calorie Estimation, presented at IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019. IEEE.
18. Dubey, A. (2015, Oct 14). The challenges and opportunities for ADAS stereo vision applications, part I. *EDN*. https://www.edn.com/the-challenges-and-opportunities-for-adas-stereo-vision-applications-part-i/
19. *APA citation guidelines* (n.d.). KITTI. http://www.cvlibs.net/datasets/KITTI/
20. Laura. (2020, Oct 30). Understanding Car Size and Dimensions. *Nationwide Vehicles Contract.* https://www.nationwidevehiclecontracts.co.uk/guides/ask-nvc/understanding-car-size-and-dimensions
21. I. Balenović, A. Seletković, R. Pernar, A. Jazbec, "Estimation of the mean tree height of forest stands by photogrammetric measurement using digital aerial images of high spatial resolution" in Annals of Forest Research, 2015.
22. *APA citation guidelines*. (2021). AWS. https://docs.aws.amazon.com/Rekognition/latest/dg/images-displaying-bounding-boxes.html
23. Adams, S. (2021, May 24). What is the Average Car Length, The In-depth Guide. *Curate View*. https://curateview.com/average-car-length/
24. *APA citation guidelines*. (2013, April 25). The Economic Times. https://economictimes.indiatimes.com/infrastructure/chennai-bangalore-ac-double-decker-express-train-starts-operations/height-width-of-train/slideshow/19720500.cms
25. Dahl, T. (2017, Feb 9). How to Read a Tire Size. *Popular Mechanics*. https://www.popularmechanics.com/cars/a25156/how-to-read-a-tire-size/
26. Lane. (2021).Retrieved June 12, 2021, from https://en.wikipedia.org/wiki/Lane

27. Mean Squared Error. (2021). Retrieved June 14, 2021. from https://en.wikipedia.org/wiki/Mean_squared_error
28. De Vries, G, Verbeek, P.W. (2002). Scale-adaptive landmark detection, classification and size estimation in 3D object-background images, presented at Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Barcelona, Spain, 2000. IEEE.
29. Naritomi, S & Yanai, K. (2020). CalorieCaptorGlass: Food Calorie Estimation based on Actual Size using HoloLens and Deep Learning, presented at 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 2020. IEEE.
30. Qiao, Y, Lew, B. V., Lelieveldt, B. P. F. & Staring M. (2016). Fast Automatic Step Size Estimation for Gradient Descent Optimization of Image Registration. *IEEE Transactions on Medical Imaging*, *35*(2), 391-403.
doi: 10.1109/TMI.2015.2476354
31. Lo, F. P. W., Sun, Y, Qiu, J & Lo, B. (2020). Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1926-1939.
doi:10.1109/JBHI.2020.2987943
32. Xiong, Z, Yuan, Y & Wang, Q. (2019). RGB-D Scene Recognition via Spatial-Related Multi-Modal Feature Learning. *IEEE Access 7*, 106739-106747.
doi: 10.1109/ACCESS.2019.2932080
33. Sabahat, N, Malik A. A. & Azam, F. (2017). A Size Estimation Model for Board-Based Desktop Games. *IEEE Access 5*, 4980-4990.
doi: 10.1109/ACCESS.2017.2678459
34. Shui, P. L. & Zhang, K. (2019). Ship Radial Size Estimation in High-Resolution Maritime Surveillance Radars via Sparse Recovery Using Linear Programming. *IEEE Access 5*, 70673-70688.
doi:10.1109/ACCESS.2019.2919242
35. Raju, V. B. & Sazonov, E. (2020). A Systematic Review of Sensor-Based Methodologies for Food Portion Size Estimation. *IEEE Sensors Journal 21*(11), 12882-12899.
doi:10.1109/JSEN.2020.3041023
36. Daud, M. & Malik, A. A. (2021). Improving the Accuracy of Early Software Size Estimation Using Analysis-to-Design Adjustments Factors (ADAFs). *IEEE Access 9*, 81986-81999.
doi:10.1109/ACCESS.2021.3085752
37. Feng, C, Zhang, C, Chen, Z, Li, M, Chen, H & Fan, B. (2020). LW-Net: A Lightweight Network for Monocular Depth Estimation. *IEEE Access 8*, 196287-196298.
doi:10.1109/ACCESS.2020.3034751