3D visualization of images/objects via AR using Object Detection

A Major Project submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Technology

Submitted by

Anish Aggarwal (185039)

Gajendra Surariya (185082)

Abhav Singhal (185049)

8th sem, 4th Year, CSE

Under the guidance of

Dr. Mohit Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, HAMIRPUR

HAMIRPUR, HIMACHAL PRADESH -177005 (INDIA)

National Institute of Technology, Hamirpur

Table of Contents

CERTIFICATE

This is to certify that the Major project “3D visualization of images/objects via AR using Object Detection” report has been submitted to the Department of Computer Science and Engineering, National Institute of Technology, Hamirpur for the fulfilment of the requirement for the award of the degree of Bachelor of Technology in “Computer Science and Engineering” by following students of final year B.Tech. (Computer Science and Engineering).

Students Name (with Roll no.)

Anish Aggarwal (185039)
Gajendra Surariya (185082)
Abhav Singhal (185049)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge:

Date:

Dr. Mohit Kumar
Associate Professor

Acknowledgements

Achievements in life are always obtained through some key ingredients added into the recipe which is your work. A spoon of guidance, a touch of inspiration, a squeeze of determination, and heaps of blessings are the most crucial among them. This work would not have been possible without the constant support and motivation given to me by my teachers and Family.

We therefore take this opportunity to express our sincere gratitude to our guide and mentor, **Dr.Mohit Kumar**, Assistant Professor, Department of Computer Science and Engineering, National Institute of Technology, Hamirpur for his expert guidance, innovative suggestions, constant encouragement, and patience with which he handled me and the project. His calm minded approach to the project instilled the confidence within me and taught me to embrace the challenges that were ahead.

We are forever grateful to **Dr. T.P Sharma**, Head of the Department, Department of Computer Science and Engineering for his encouragement and considerations throughout my research work. His constant support has been invaluable to us in the pursuit of our project work.

Gajendra Surariya
Anish Aggarwal
Abhav Singhal

Abstract

The approach of the project is interactive, innovative, and quintessentially holistic. The project deals with making students visualize concepts in 3D. AR education will help students to comprehend concepts they learn thoroughly in an easy way. Using image feature projecting and calculating camera position from sequential pictures, a pose estimation approach for planar structures was suggested. Changhai et al. described a technique for reliably estimating 3D plane poses using Bayesian inference and a balanced progressive normal estimation approach. To calculate the plane pose, Donoser et al. used the characteristics of Demography is the study It There Regions to establish a positivist approach invariant frame on the closed contour. In our method, we employed viewpoint transformation to estimate a set of matching points on the gives a signal in order to estimate the object surface's basis vectors, and then used depth analysis to calculate the 3D pose by calculating the horizontal to the plane object.

Objective

It provides a common platform for students from different backgrounds to learn and explore concepts of their interest. Apart from this, it also promotes home learning which is the need of the hour in this period of turmoil. Students generally face problems in visualizing the concepts, this application develops a better understanding of the subject The objective of this project is not only to promote home learning but also quality learning since the students are using various sources for learning it is necessary for us to improve the learning methods so as to make learning more interesting, we have already seen that experiencing the things first hand provides better learning and skill development than just reading. We can take example of Labs of various subjects, the labs potential to explain a idea is greater than just reading book since the demonstration is included thus physical experience.Through our project we want to provide seamless experience to learners for visualizing the various concepts whose imagination are difficult like various 3d shapes of chemical bonds in chemistry or complex 3d figures of mathematics. We aim at attaining the visualization of complex topics through augmented reality so as to improve the quality of learning materials..

Basic Terminologies

- Unity: 3D and 2D models or games are professionally built on game engine for each platform.
- Blender: It is 3D computer graphics software used for creating visual effects and 3D printed models.
- AR Core: It is Google’s platform for building Augmented Reality experiences.
- Android Studio: Official integrated development environment for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development.

Explanation

We are using two methods in integration with AR frameworks :
1) Object Detection:- Detection of object in AR involves computer vision through camera where objects can be viewed in digital photos or videos. Take example of ball,car,shoe,vessel,pen or any object which can be detected in camera. The 2 criterias are the specifications and the location of the object. The object should be clearly visible in the image, that is , the lighting and focus should be proper and the image should not be blurred or dark. Else, it might lead to wrong detection or no detection at all. Object detection can be done by various methods like CNN(convolutional neural network) , SSD(Single-shot Multibox Detector) ,fast R-CNN and Retina-Net. These algorithms are not that efficient as compared to YOLO algorithm as it has higher accuracy and efficiency over algorithms.

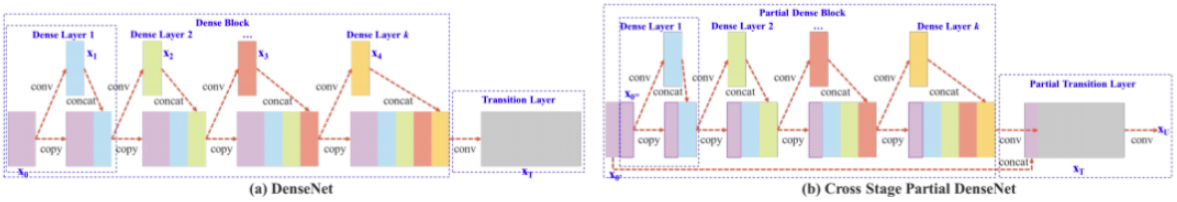
2) Pose Detection:- Pose detection has gained popularity in last few years in fields like robotics , teaching and AR. Using image feature projecting and calculating camera position from sequential pictures, a pose estimation approach for planar structures was suggested. Changhai et al. described a technique for reliably estimating 3D plane poses using Bayesian inference and a balanced progressive normal estimation approach. To calculate the plane pose, Donoser et al. used the characteristics of Demography is the study It There Regions to establish a positivist approach invariant frame on the closed contour. In our method, we employed viewpoint transformation to estimate a set of matching points on the gives a signal in order to estimate the object surface's basis vectors, and then used depth analysis to calculate the 3D pose by calculating the horizontal to the plane object.

Execution

• We will be making 3d components of circuits, and 3d shapes for math curves using Blender.
• After that we will import these models in unity and do scripting.
• Scripting will be done using C# language.
• We will write different scripts for different operations.
• Using AR we add digital content to the live camera feed making that digital content look as if it is part of the natural world

YOLO

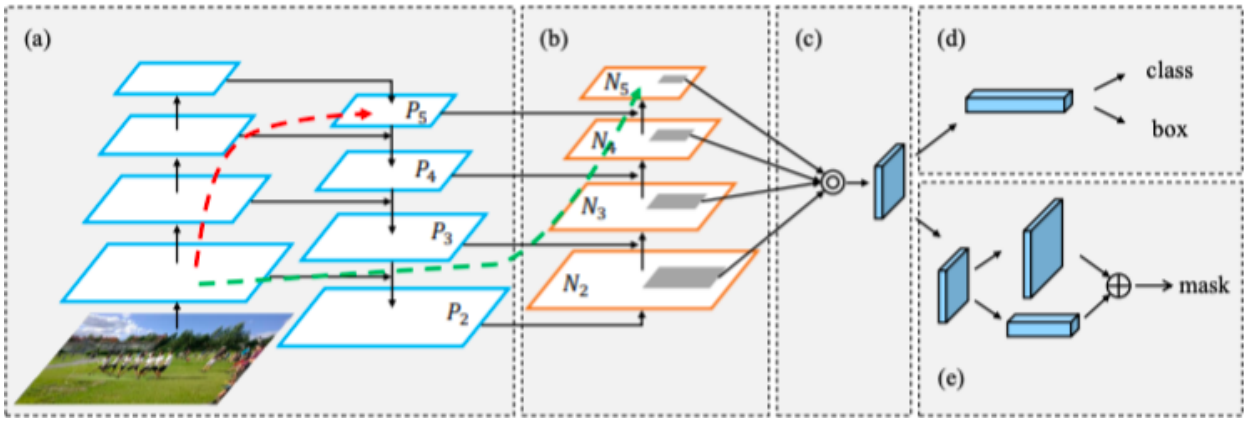
YOLO is used for object detection, it has a total of 5 versions. We used pre-trained weights trained on the COCO dataset. Object detection is done for those COCO dataset Labels. V4 was created in 2020. It is a high-speed object detector, optimized for parallel computation. It operates in a super-fast fashion, renders high accuracy, and provides remarkable detection results. It separates high context features which can be significant and cause an insignificant reduction in the operation speed of the network.



Data Augmentation is used as the backbone for increasing the variability of the input images to attain higher robustness within the object detection procedure where the images are obtained from different environments/sources,

YOLOv4 is designed on CSPDarknet53 (BackBone), Spatial pyramid pooling, and Path Aggregation Network(Neck). Here, the Head is YOLOv3. SPP is designed to get fine and coarse information by pooling different sized kernels. To increase the accuracy of the model we used mAP matrix for calculating accuracy after every epoch.

PAN takes the information from input layers and converges it into the features provided by individual backbones and feeds it to the detector.



Multiple SDKs have been provided which have commands to detect the objects from live camera feed or input images.

The detector is suitable for training on 1 GPU as classic hyperParameters are selected during the execution of genetic algorithms.

Training on MS COCO

- We're supplying MS COCO with pre-trained weights. These weights can be used as a starting point for your own neural network version. The code for training and assessment may be found in samples/coco/coco.py.
- We divided our 200K photos into three categories: 70% training, 20% validation, and 10% test. All of the photographs in this gallery are from our test set.
- Additionally, we will train it only 2 classes ,person and background, standard cross entropy loss can be used to measure the performance.

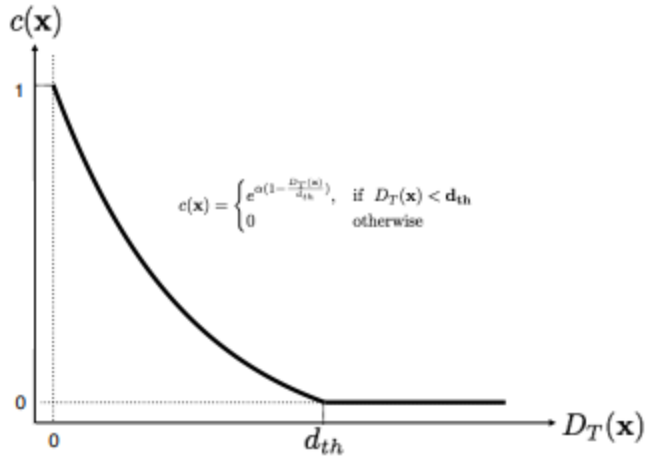
Hyper Parameters:

- Learning rate - 0.01 with a step decay scheduling strategy.
- Training Steps - 500,500
- LearnSteps - 0.1 multiplied at 400k
- Momentum - 0.9
- Weight Decay - 0.8
-

Pose Prediction of 3D Object

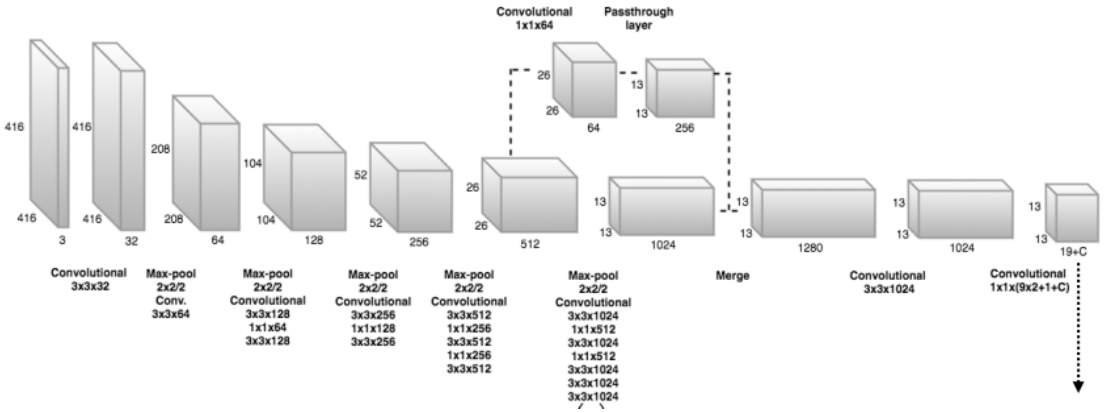
Pose detection of objects is done by 6D pose detection, cThe Pose predictions are made by using deep CNN architecture.
The architecture is fed with RGB image as input and outputs a multidimensional vector representing coordinates of bounding box projection on a 2D plane. This Technique uses CNN to process the input image and divide the image into N X N cells the. The divided cells are then tested for cCConfidence based on the expression below.

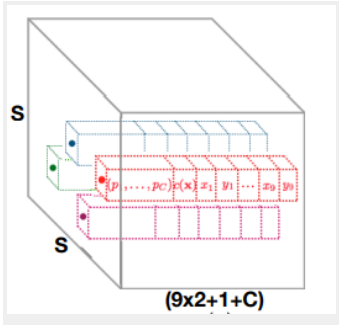
$$c(\mathbf{x}) = \begin{cases} e^{\alpha(1-\frac{D_T(\mathbf{x})}{d_{th}})}, & \text{if } D_T(\mathbf{x}) < d_{th} \\ 0 & \text{otherwise} \end{cases}$$



Along with Confidence 9 control points are also calculated for each cell, 8 of these control points are coordinates of 2D projections of predicted Object bounding box corners. The remaining one point is for the probability of class that is present in the current cell but we will not be using class probability values for that we will be using yolov4.

We need to find the coordinates of camera from the 6D orientation of Object. For this first We find the 8 coordinates of bounding box corners and centroid coordinates which represents 6D orientation of the object by passing the RGB image from the network once and the cells with low confidence are not considered for the next step as they don't contain the object. If an object is bigger than the cell then more cells will show higher confidence. Now a single pass is taken of all the cells to generalize the result over all the N X N cells. All the 8 neighbors of individual cells and the cell itself is considered for weighted average, thus recalculating the 8 control points with the help of weights. Weight used in recalculation of corner points and centroid is confidence of the cell. Through the previous step 2D projection of the centroid and bounding box around the 3D object is obtained. This information is used along with the perspective n point algorithm to estimate the position of the camera. The position of the camera is then used to render model on the selected target object.


Convolutional
1x1x64
Passthrough
layer
416
416
3
Convolutional
3x3x32
416
416
32
Max-pool
2x2/2
Conv.
3x3x64
208
208
64
Max-pool
2x2/2
Convolutional
3x3x128
1x1x64
3x3x128
104
104
128
Max-pool
2x2/2
Convolutional
3x3x256
1x1x128
3x3x256
52
52
256
Max-pool
2x2/2
Convolutional
3x3x512
1x1x256
3x3x512
1x1x256
3x3x512
26
26
512
Max-pool
2x2/2
Convolutional
3x3x1024
1x1x512
3x3x1024
1x1x512
3x3x1024
3x3x1024
3x3x1024
26
26
64
13
13
256
13
13
1024
Merge
13
13
1280
Convolutional
3x3x1024
13
13
1024
Convolutional
1x1x(9x2+1+C)
13
13
19+C


S
S
(9x2+1+C)

Evaluation

6D pose estimation is in finding the position of camera with respect to an object of interest so if you imagine you have object that you're interested in locating and you base a 3d coordinate system around . 6D estimation algorithm takes given image from camera and it finds the camera's physical position in relation to the object as well as rotation of the camera in relation to the object. There are 3 translation coordinates and those three rotation coordinates . The RGB values ranging from 0 to 255 are used as input data for our neural network used in PV new illustration. This neural network is trained to produce these two outputs semantic segmentation. The semantic labeling labels pixels according to the object that it belongs to. Instead of associating each pixel with a class , it actually associating each pixel with a set of unit vectors so these unit vectors are related to what we need to solve our problem. It gives the performance of 32 fps, which is favorable for our application to function properly. The average 3D distance and IoU score of model vertices were employed as assessment criteria.
ADD metrics are used to calculate the average of 3D distance between model vertices.These metrics calculate the average mean distance between ground truth coordinates and predicted 3D coordinates and the pose estimate is only correct when the calculated mean is greater than 10% of diameter of the object.
The ADD of the 6D pose estimation came to be 55.64.
Various IoU scores for pose prediction are as follows : AP50 of model is 96.74.

The YOLOV4 provides the performance speed of 38 fps at an AP50 of 66.

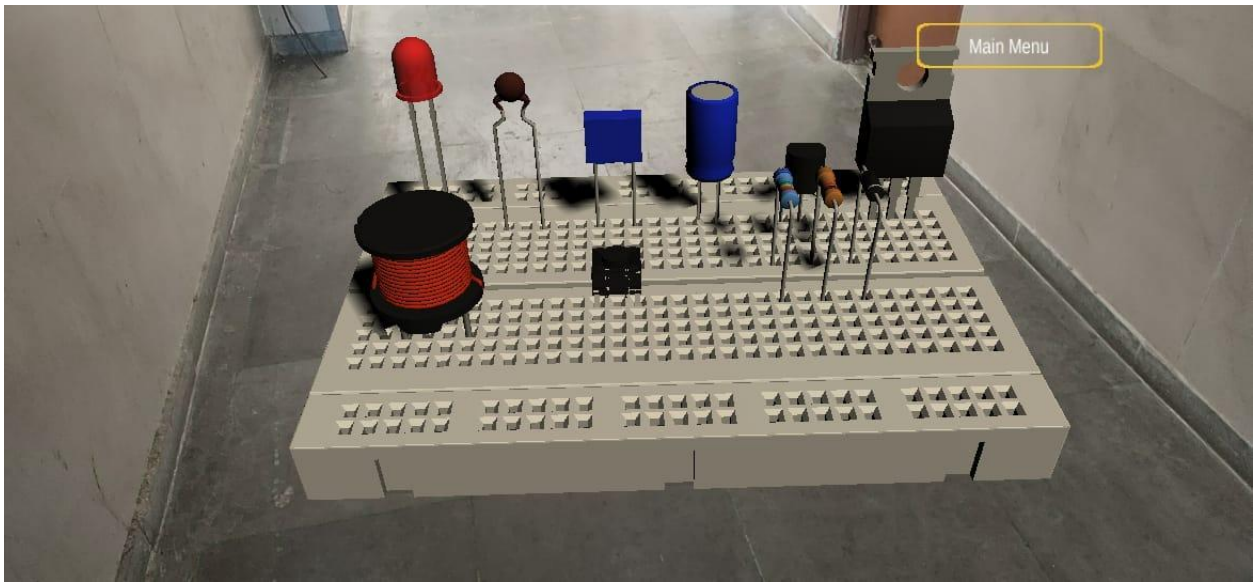
AR Android and IOS cross platform App

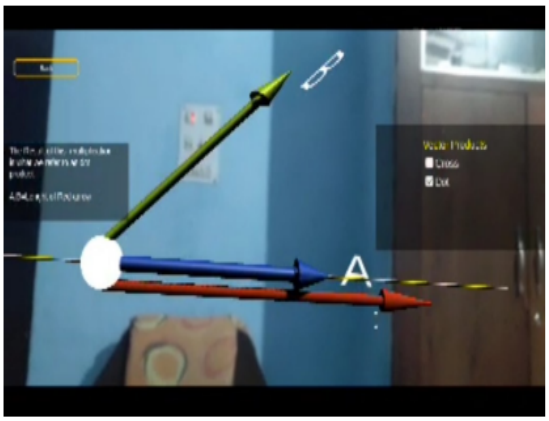
In today's world the virtual learning is increasing, with increase in virtual learning the lack of physical experience from various topics creates confusion and lack of clarity of various subjects like biology, physics etc, and various experience based knowledge like lab tools etc. To give

contrast to the Application of these methods on education and to improve the learning process we integrated the above discussed methods with augmented reality thus the user not only can watch how those concepts look and work but also interact with them. The app provides main menu through which a user can learn various information through experience.

The aim of the project is to make education easier with the help of augmented reality
Firstly we tried to understand that what are the basic requirements for this project. After researching for around 2,3 days we made a list of things that we need to learn and understand. Like 3rd modelling using blender, we have used unity 3d, learning C# (language used in unity) It was somewhat similar to the OOPs. The working procedure of the project must be like that firstly the object/Image is detected by the ML model in which input js taken from the device's camera and after that the information is passed to our unity part in which whatever the object/Image is detected by the ml model is augmented. We have then used Blender which is a software for creation of 3d models of requisite objects. We have used various versions of unity for making our project compatible with the detection and estimation models. We have used C# for writing scripts.

Below are various pictures of the application.





Code

https://drive.google.com/drive/folders/1jUraejxrLHDSYl5d9mr62PWUub3OCC9v

Applications:-

- Book Visualizer: In depth learning of some important concepts of science by visualizing the phenomenon. The image of object or concept in 3D will help students for better understanding of phenomenon.
- Circuits:3D representation of various circuits for better visualization and understanding of students to make their learning much effective and interactive.
- Mathematical Concepts: Concepts like mathematical equations and graphs, conic sections and visualisation of geometric figures like sphere, cylindrical figures. It will predominantly benefit core topics that include various graphs and 3D curves.
- It can be used in several labs to visualise the working of instruments.
- Can be used to make studies interactive.
- Instead of searching for YouTube videos each and every time students would be able to use this application to get a demonstration just by pointing the camera.

Results

The ADD of 6D pose estimation came to be 55.64.
Various IoU score for pose prediction are as follows : AP50 of model is 96.74.
The YOLOV4 provides the performance speed of 38 fps at an AP50 of 66.

Conclusion

We retrained a pre-trained mask YOLO model on the COCO dataset and evaluated it on the Pascal VOC and Cityscape datasets in this study. YOLO is used in our method for object detection and then creating a 3D bounding box in 2D image. Later-on Pose Detection is done for identifying the correct object's geometry in the 3D plane, the result is fed to AR so that the 3D figure can be plotted. Our model is capable of rendering 3D figures at a speed of 20FPS as YOLO is suitable for working at a fast speed.

Bibliography

- YOLO v1: https://arxiv.org/abs/1506.02640
- YOLO v2: https://arxiv.org/abs/1612.08242
- YOLO v3: https://arxiv.org/abs/1804.02767
- Real-Time Seamless Single Shot 6D Object Pose Prediction : https://arxiv.org/pdf/1711.08848.pdf
- CIOU - LOSS
- DropBlock: A regularization method for convolutional networks : https://arxiv.org/pdf/1810.12890.pdf
- Class Label Smoothing : https://paperswithcode.com/method/label-smoothing
- https://en.wikipedia.org/wiki/Cepstral_mean_and_variance_normalization
- Self-Adversarial Training : https://arxiv.org/pdf/2107.02434.pdf
- https://becominghuman.ai/explaining-yolov4-a-one-stage-detector-cdac0826cbd7
- CBAM: Convolutional Block Attention Module: https://arxiv.org/pdf/1807.06521.pdf
- Object Detection and Pose Estimation from RGB and Depth Data for Real-time, Adaptive Robotic Grasping