Image Detection and Recognition for Smart Glasses

in the Field of Transportation

Shengzhe Hu, Jiaman Li, Yichen Li, Dandan Zhao, Jingjing Zhu

Dec 3, 2019

Capstone in Data Analytics

Department of Applied Data Science

San Jose State University

Table of Contents

[Abstract 3](#_Toc26270532)

[Chapter 1 Introduction 4](#_Toc26270533)

[1.1 Project Background and Goals 4](#_Toc26270534)

[1.2 Analysis of Requirements 6](#_Toc26270535)

[1.3 Project Deliverables 7](#_Toc26270536)

[1.4 Technology and Solution Survey 8](#_Toc26270537)

[1.5 Literature Survey of Existing Research 14](#_Toc26270538)

[Chapter 2 Data Exploration 21](#_Toc26270539)

[2.1 Data Modeling and Plan 21](#_Toc26270540)

[2.2 Data Sources and Dataset Parameters 22](#_Toc26270541)

[2.3 Collection and Training Datasets 27](#_Toc26270542)

[2.4 Data Cleansing and Validation 29](#_Toc26270543)

[2.5 Data Transformation and Tools 30](#_Toc26270544)

[Chapter 3 Project Management 31](#_Toc26270545)

[3.1 Project Organization 31](#_Toc26270546)

[3.2 Resource Requirements 32](#_Toc26270547)

[3.3 Project Schedule 33](#_Toc26270548)

[References 34](#_Toc26270549)

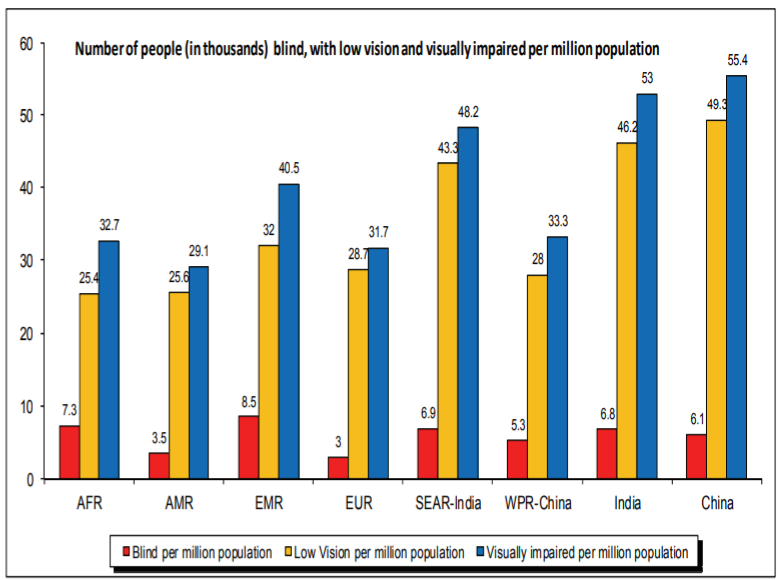
# **Abstract**

With the progression of science and technology, advanced technologies are being used in health care, which can help people with disabilities to make their lives easier. In recent years, the techniques of image identification become more mature and can be applied to our daily lives to make our lives more convenient. For people with visual disabilities, daily transportation is a tough task that restricts their travel freedom. This project aims to build analytics models to recognize five different major outdoor subjects including but not limited to utilities, vehicles, bikes, pedestrians and traffic lights to implement an image recognition system that can help with the detection of major roadblocks for people who have visual disabilities. Finally, this project plans to apply the image recognition system for smart glasses in the field of transportation to improve visual disturbances people’s safety and convenience in their daily commute for practical use.

# **Chapter 1 Introduction**

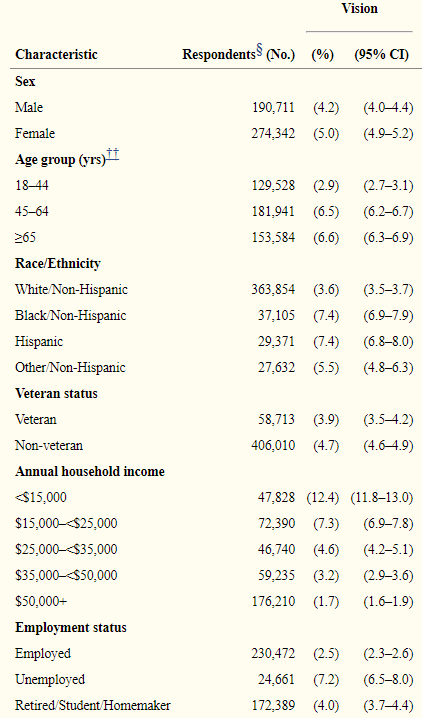
## **1.1 Project Background and Goals**

**Project background.** According to the report “Global Data on Visual Impairment 2010” provided by WHO, there are 285 million people with visually impaired and 13.7% of them are blind (WHO, 2010). Here is the distribution chart in six regions:



*Figure 1.* The number of people (in thousands) blind, with low vision and visually impaired per million population.

As the figure is shown, blindness means people unable to recognize lightness and darkness and low vision means people cannot correct their vision with glasses. In the U.S, there are 38 million people with visual impairment or blind in 2010 (Works, 2015). The report “Key Findings: Prevalence of Disability and Disability Type among Adults, United States – 2013” also mentions that among the 22% of Americans who claim that they have a disability, 4.6% of these disabled people who have a vision problem (Courtney-Long, et al., 2015). Figure 2 shows the detailed demographic information among these visually impaired people:



*Figure 2.* Demographic information among these visually impaired people.

As we can see in the previous figure, the household income of most visually impaired people is in the range between $15000 and $25000 and 7.2% of them are unemployed. Thus, there are a few alternatives provided to visually impaired people compared to costly blind or low vision treatment in the market, for example, electronic glasses.

**Project goals.** These are the reasons why we choose to work on the topic of glasses for visually impaired people. From the company side, there is a very large potential market for disabled people. According to the International Center for Corporate Accountability, the disabled people’s consumption capability is twice as much as the spending power of teens (Bates, 2017). Studies also show that disable people have customer loyalty to business which support them (Works, 2015). Most importantly, as we can see from the above table, most visually impaired glasses assist visually impaired people reading books and performing activities indoors and few of them are designed for outdoor activities. From the customers’ side, we want to improve visually impaired people's quality of life quality since many people are experiencing with a visually impaired problem. Through these projects and research, we hope we can both benefit the companies and customers.

This project aims to build image classification models used on smart glasses to help identify the most common objects on the street and enhance the company’s smart glasses development.

## **1.2 Analysis of Requirements**

This project is to build analytics models to recognize five different outdoor subjects including but not limited to utilities, all vehicles (cars), bikes (scooter, bikes, motorcycles), pedestrian and traffic lights (signs). By integrating these models, this project will create an image recognition system for smart glasses. Accuracy and F-beta scores will be used to evaluate the result of the system.

The project aims to build image recognition models that can detect objects with an accuracy of at least 80% after combining the five models into one model. Also, according to the performance of the models and the time consuming, we would choose a model that reaches our accuracy goal but no longer than 5 minutes. While considering the cost of the models, we would like to choose the simpler model which may have the same performance with the complicated ones to reduce cost. To achieve these requirements, we will need to keep cleaning and preprocessing our dataset, and at the same time tuning our model parameters to achieve the best result. The result can be tested and measured by python libraries such as Scikit-Learn to generate training and testing accuracy and score. Also, the confusion matrix, classification report, ROC curve, and AUC score will be the tools to evaluate our models.

## **1.3 Project Deliverables**

Through the analysis of the project, this study aims to build an effective image recognition system for people with visual disturbance, to improve their safety and convenience in their daily commute. Our system should be able to detect the five objects from the testing images, which are real photos taken from real scenes. The objects will be recognized by bounding boxes with labels.

We will deliver our project through the following format:

* Models - We will build models to test and validate the prediction accuracy, Google Colaboratory will be handed in
* Report - We will generate a report to summarize our research, explain our technological process and describe our output and achievements through figures, tables, and processing timeline charts.
* Prototype - We will record a short video of the street scene and break it into images to detect the five objects
* PowerPoint - We will present our results in a PowerPoint slides

## **1.4 Technology and Solution Survey**

**Introduction of object detection.** In the computer vision objects detection area, the general method is to select objects first by special methods, such as Selective Search. The process of selecting objects is called Region Proposal. Object identification is then performed for the selected items (Region Proposals) (but because the selected items may be of different sizes, the object identification may consist of only the categories or may include feature extraction plus categories). Such practice that Region Proposal needs to be found out first and then identification is usually called two-stage learning.

These two stages methods usually have a problem is that whether the selected object is too much. Assuming that one thousand items are selected in a picture, then one thousand times of classification will be needed. For example, if a single identification of an object to 0.1 seconds, then one thousand objects will need about 100 seconds. It required a powerful GPU to do parallel computing in complex systems so that can reduce the operation time. However, this type of algorithm real-time operation cannot be applied to a mobile phone. Hence, the one-stage method was developed.

One-stage Learning is that object identification can be achieved in one step, which means that a neural network can simultaneously detect object location and identify objects. Such as Single Shot Detector (SSD) released from Google in December 2015. Google's article in the first sentence and then write “We present a method for detecting objects in images using a single deep neural network.”, a depth of neural network can finish all the object detection. This method is usually very fast, but the overall identification accuracy may not be as good as that of the two stages method, but the overall identification rate is still within the acceptable range. Therefore, the method of One stage is currently developed by many people for mobile devices.

**Development of object detection technologies.** The technologies of object detection have improved a lot during these days, the following we list some technologies which include their advantages and their disadvantages.

(a) Features are computed from images with different scales independently (Featured image pyramids).

This method means image detection on different scales. If the original image size is 1000×1000, in addition to detection in the image size of 1000×1000, the image can also be reduced to detect, such as 500×500, 300×300 or 10×10 for processing. The advantage of this method is that different scales can be used for the detection of different object sizes, such as high-resolution graphs such as 500×500 for the detection of small objects, and 10×10 for the detection of large objects. If there is only one large object in the picture, it can be done well on the small scale, while if it is a small object, it needs to be resolved well on a larger scale.

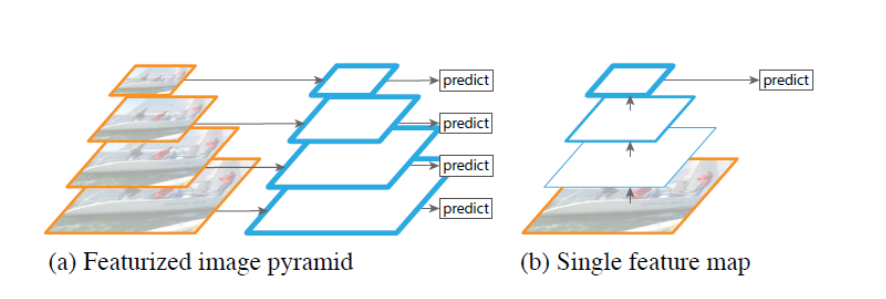
Disadvantages: a scale diagram needs to do CNN once, so it takes a lot of calculations (assuming the scale is set to 100, it needs to run CNN for 100 times), which takes a lot of time. The result is good, but the calculation takes a long time.

(b) Single scale features (Faster R-CNN and R-FCN, YOLOv1 and YOLOv2.)

This method is applied by the classical method (Faster RCNN, R-FCN, YOLOv1 and YOLOv2). When the feature extractor is matched, the feature extractor structure of CNN is generally in the following figure. The feature map is continuously shrunk and extracted by using convolution, pooling or convolution by means of stride=2. Object localization and Classification usually be done under the feature map of the last scale. Generally, the feature map under 1/32 is used for detection. For example, if the input map size is 416×416, the feature map on the last layer is 13×13 (416/32=13).

Advantages: The structure of such methods is usually simpler.

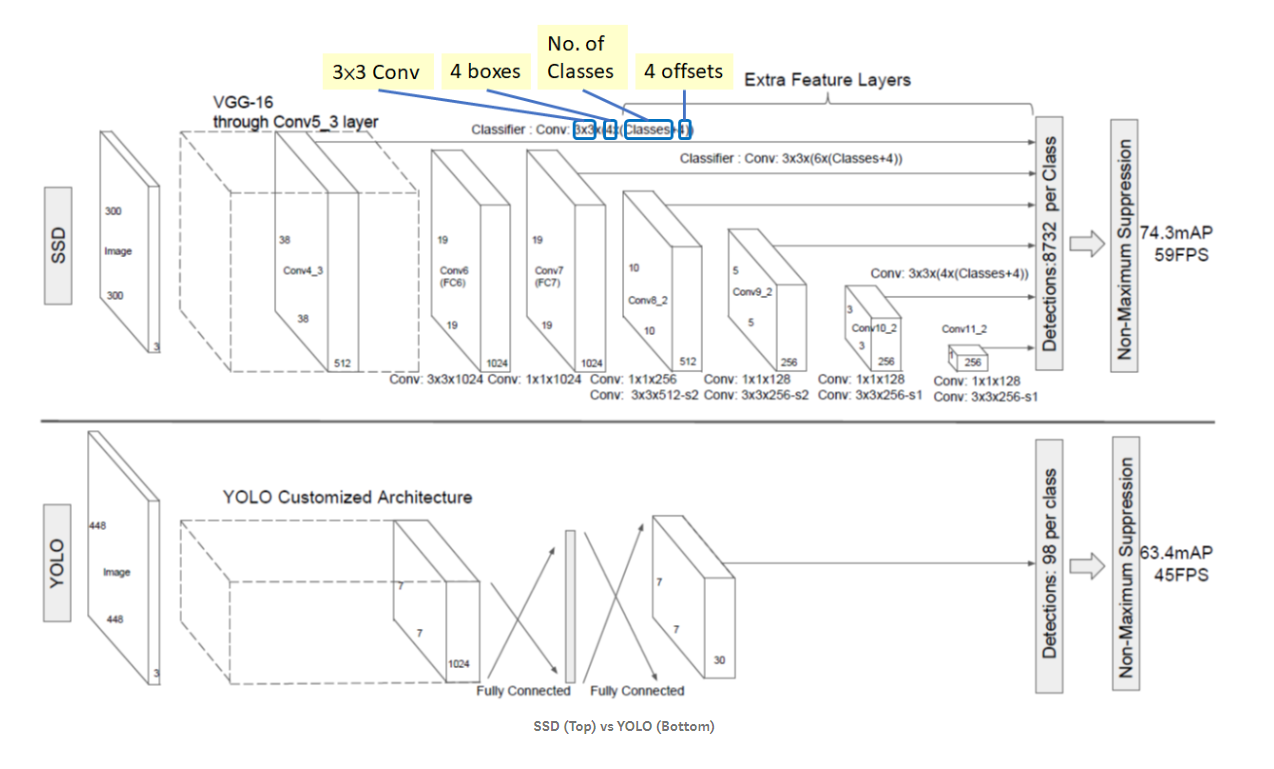
Disadvantages: Object localization and Classification are performed on a feature map of a single scale. If the scale=1/32, the detection performance of small objects is relatively poor. It is possible that features of small objects disappear when down-scale occurs.



*Figure 3.* Comparison of different scales features with a single feature.

(c) Multi-scale features (Pyramidal Feature network, SSD).

A multi-scale feature map can solve such problems as (b). Object localization and Classification can be performed under feature maps of different scales. Single Shot MultiBox Detector (SSD) is the most classic method of this Classification.



*Figure 4.* The architecture of SSD vs YOLO

(d) Feature fusion method (Feature Pyramid network, YOLOv3).

This method integrates (b) the advantages of feature map fusion and (c) the advantages of multi-scale detection. In the small-scale feature map (after a lot of layers of the convolution results) image semantic is enough but space explanation ability is poor. At this time using unsampled (methods: deconv or transpose conv., nearest neighbor, bilinear, etc.) will zoom in the feature map. Then detections can be done by combining fusion with feature maps of different scales.

Therefore, this method not only retains the advantages of feature map fusion but also integrates the advantages of multi-scale detection.

Advantages: The semantic features of different scales can be extracted enough, and the detection effect is good.

Disadvantages: Seemingly perfect solution, but the model is much more computationally intensive, with poor execution time and memory utilization.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 1  *Comparison of Multiple Algorithms* | | | | | |
|  |  | Features are computed from images with different scales independently | Single scale features | Multi-scale features | Feature fusion method |
| Purpose | Detection | Yes | Yes | Yes | Yes |
| Classification | Yes | Yes | Yes | Yes |
| Scales | Various  scales | Yes |  | Yes | Yes |
| Single  scale |  | Yes |  |  |
| Computation intensive |  | Yes |  |  | Yes |

**Industry solutions.** Table 2 shows the current industry solutions regarding smart glasses for blind people. The main functions, features, and prices are listed below (Malik, 2019):

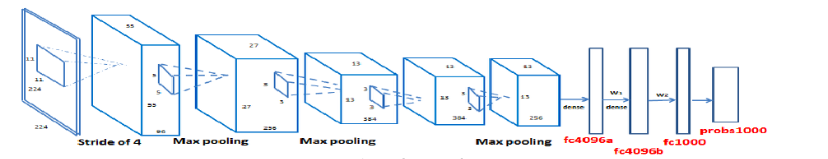
|  |  |  |  |
| --- | --- | --- | --- |
| Table 2  *The Top 5 Electronic Glasses for the Blind and Visually Impaired People* | | | |
| Name | Description | Feature | Price |
| IrisVision | Design for reading books and watching TV | - 14 X magnification | $2950 |
| Acesight | Design for reading | - Float reading mode  - HD display  - 15 X magnification  - Customize color and contrast  - 45-degrees field of view | $4995 |
| NuEyes Pro | Design for reading | - Design for reading  - Can be controlled through voice commands or wireless hand controller  - 12 X magnified images | $5995 |
| MyEye2 | Design for reading, writing, recognizing faces | - Attached camera  - Wireless operations  - Color and page detection  - Face, barcode, and product identification | $3500 |
| eSight | Design for reading and recognizing faces | - Forward-facing camera  - Head-mounted goggles designed | $5950 |

**Future work**. Improvement of the structure of object detection should be limited in the future. The research direction may tend to perform the application more accurate and faster. Future trend: (1) Change the feature extractor. (2) As mentioned in (d) Feature Fusion, Change the structure of detection to make the performance good but reduce the overall amount of calculation and parameters.

## **1.5 Literature Survey of Existing Research**

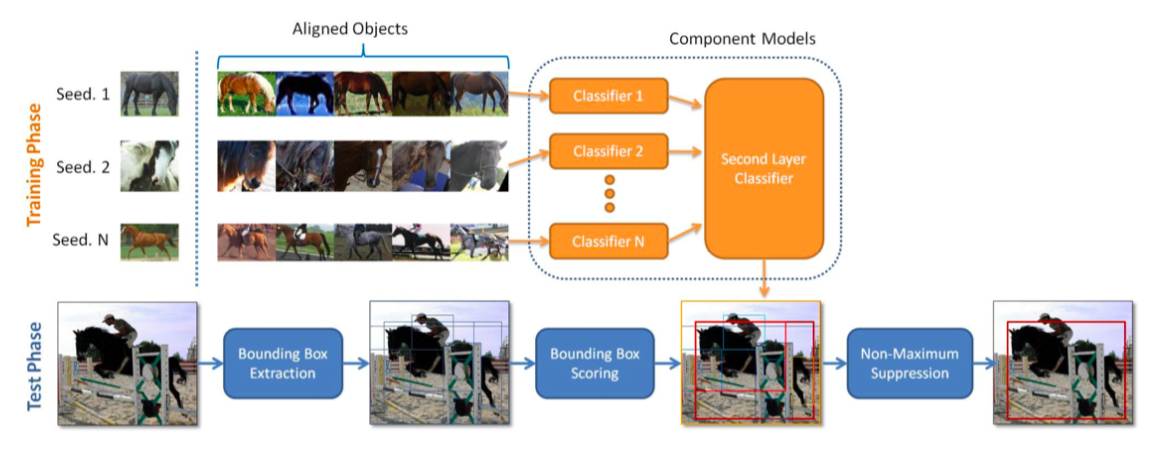
There were various kinds of approaches being used to perform objection detection on images, and different machine learning or deep learning models were being used depends on the problem context. Based on the problem that every research paper is trying to solve, researchers selected the model or multiple that can meet the requirement, and the performance of different models are being compared.

**Convolutional neural network in AlexNet architecture.** In a research paper about image classification using deep learning (Manoj, Neelima, Harshali & Rao, 2018), the convolutional neural network in AlexNet architecture was being used for image classification to show the effectiveness of using deep learning. The author selected four images from the ImageNet database to experiment. The four images were cropped into separate portions to do the validation, and all the classification results were successful. This paper shows the effectiveness of using CNN in AlexNet architecture in the image classification area.



*Figure 5.* The model design for Convolutional Neural Network in AlexNet architecture.

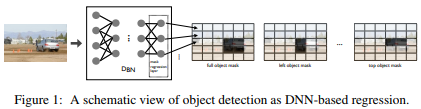
**Multi-component Models for Object Detection.** Gu, ArbelAez, Lin, Yu, Malik (2012) in their research paper performed object detection by firstly clustering components that are tightly similar due to intra-class variations. This could include subcategories, different viewpoints or poses. After the first layer of N classifiers, the second layer will be another classifier that aggregates the components on the category level. On the test level, they generated a small number of bounding boxes on each image and each box is scored by their two-layer model. They used binary Intersectional Kernel SVM for each category to do the PASCAL detection challenge and they got a competitive performance.



*Figure 6.*Structure of the two-layer classification training model and testing strategy.

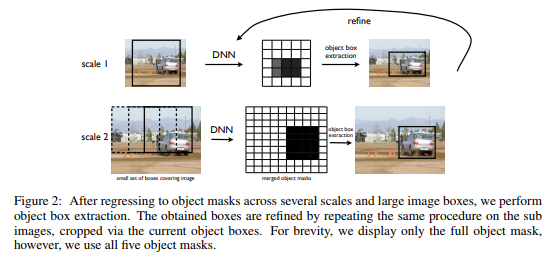
**Contextual models for object detection using boosted random fields.** In the research paper “Contextual models for object detection using boosted random fields”, Torralba, Murphy & Freeman (2005) try to create a vision system to distinguish small objects in images. They model the relation between object classes by using Boosted Random Fields, which is based on boosting and conditional random field (CRF). All pixels in the image are given a label and 100 images are selected from the office and street scene category to be training data. During data training, only local potentials are updated and increased local evidence in the first 5 rounds. Compatibility functions were updated only after the 5th iteration. To have more complicated local potentials, only local potential and compatibility functions associated with a single object class are updated. The algorithm first learns to identify easy and big objects because this is the quickest way to eradicate the error of all classes. The easy-to-detect objects can then pass information to the harder ones. The BRF algorithm is more proficient for training and inference. Moreover, it can detect information and reject certain objects quickly (Torralba, Murphy, & Freeman, 2005).

**Deep neural networks for object detection.** Szegedu, Toshev, & Erhan (2013) in their research paper utilized deep neural networks (DNNs) to distinguish numerous objects of different sizes, within the same images. The authors choose the DNN model because it has the capability to handle complex models. It is also able to automatically learn dominant object representations and deal with ImageNet classification tasks across various classes. They changed the final layer with a regression layer to record strong geometric information. They formulate a DNN-based regression to predict the bounding boxes of multiple objects in one image and outputs a binary mask of the bounding box, shown in Figure 7.



*Figure 7.* An object detection as DNN-based regression.

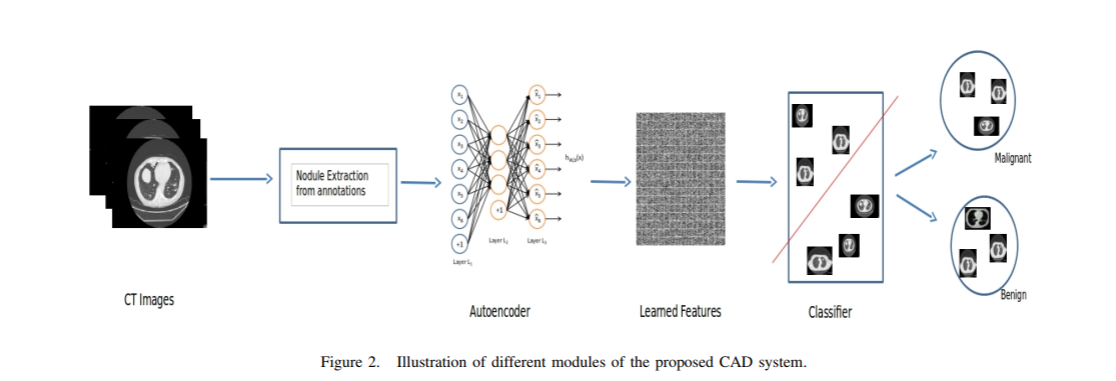
DNN mask generation in a multi-scale fashion was applied on both the full image and a few large image crops to enhance localization accuracy, as shown in Figure 8. As a result, they can predict a low-resolution mask, but it is limited by an output layer (Szegedu, Toshev, & Erhan, 2013).



*Figure 8.* The process of object box extraction.

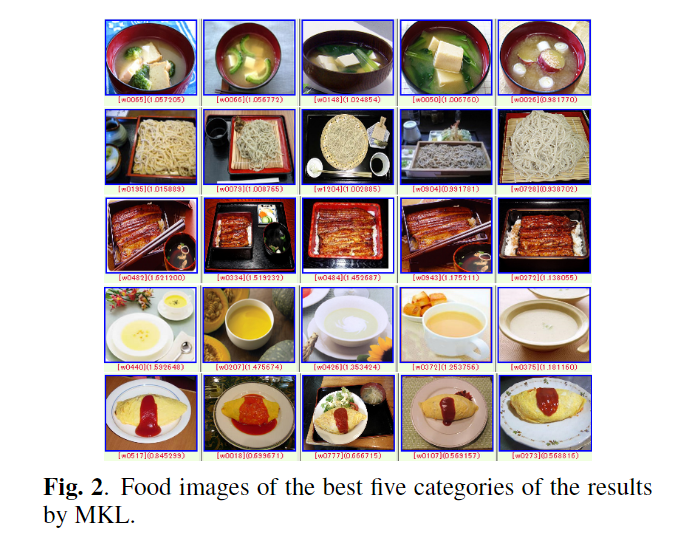
**Image classification: sparse coding and deep learning model.** Druzhkov and Kustikova (2016) mentioned two methods in image classification: sparse coding and deep learning model. Before deep learning’s development, BagofWord is the most popular and subsequent method to solve classification problems. One of the most advanced algorithms is sparse coding. However, sparse coding cannot build features’ hierarchies. For a deep learning model, Druzhkov and Kustikova (2016) introduced two unsupervised learning models: Autoencoder (AE) and Restricted Boltzmann Machine (RBM). Autoencoder can perform coding when the information is lost. Therefore, the following decoding result is close to the original data. Besides, Autoencoder sometimes could filter some unimportant details to build a visional object model and produce the hierarchy of features. On the other side, RBMs are stochastic neural models, which have two layers corresponding to the visible and hidden probabilistic system. RMMs can be used for convolutional neural networks pre-training (Druzhkov & Kustikova, 2016).

**Image classification using a computer-aided diagnosis.** Another paper Lung Nodule Classiﬁcation Using Deep Features in CT Images (2015), developed a Computer-aided diagnosis (CAD) system to help the radiologist to make a decision more precisely, the system uses the deep features extracted from an autoencoder to make a classification, which recognize lung nodules as malignant or benign. During the process of building the system, they labeled the nodes from CT images and fed the nodule area to the autoencoder and extract the features of the nodule, which as the training data, the classifier is the binary tree classifier, so the result either positive or negative. This paper explains the data extraction and deep feature extraction for image recognition.



*Figure 9.* Different modules in the proposed CAD system.

**Food image recognition system using a multiple kernel learning method.** Being the first report of implementing a food image recognition system, Joutou and Yanai (2009) adopted a multiple kernel learning method to achieve recognition with many kinds of food categories. They selected 50 different food categories to train the SVM model and achieved the 61.34% recognition rate. Finally, they implemented a prototype system for users to upload food pictures taken by their cellular phone and reached the 37.35% recognition rate. This paper provides a good example of food image recognition for practical application in the future.

****

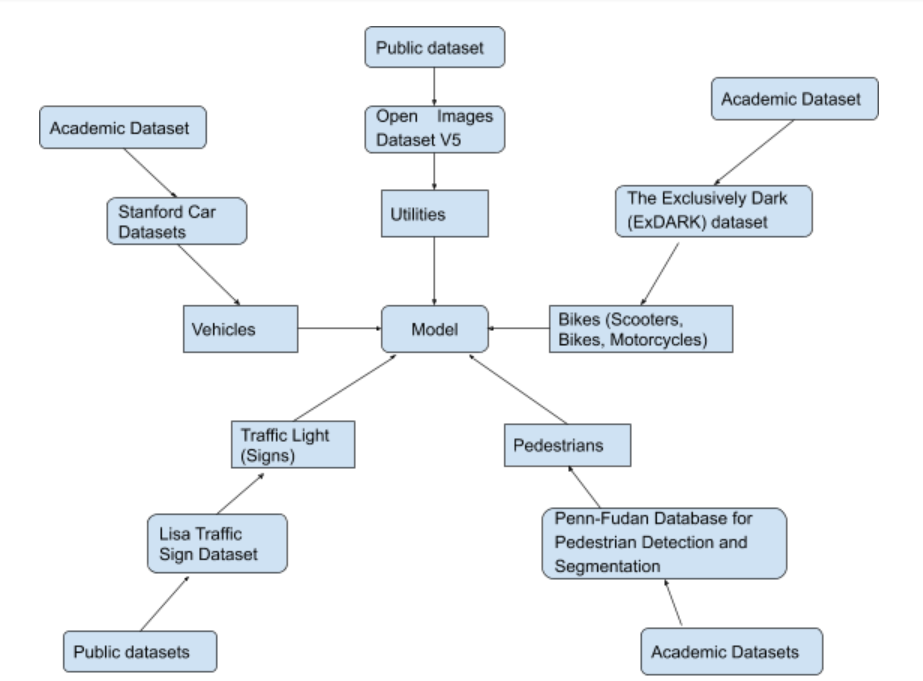
*Figure 10*. The best five results by MKL in food image recognition.

**Summary of literature survey.** Table 3 below is a summary of important information of all the literature reviews:

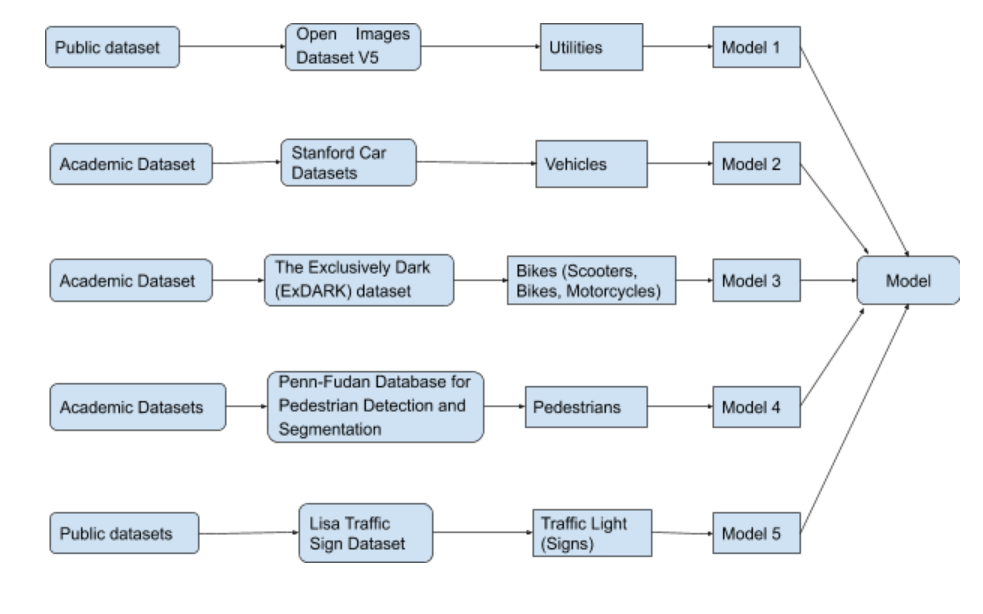
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3  *Summary of Literature Review* | | | | | |
| Research Paper | Research Area | Model | Result | Model Advantage | Model Disadvantage |
| Image classification using deep Learning | Image classification | AlexNet | The classification is successful on cropped images | Powerful on extracting image features and patterns |  |
| Multi-component models for object detection | Object detection | Two-layer model | Competitive on PASCAL detection challenge | More accurate and finer-grained | Limitation on number of components |
| Research Paper | Research Area | Model | Result | Model Advantage | Model Disadvantage |
| Contextual models for object detection using boosted random fields | Object detection | Boosted random fields | Small items are detected from images | Easy for training and inference |  |
| Deep neural networks for object detection | Object detection | Deep neural networks (DNN) | Achieve state-of-art bounding box localization | Produce high-resolution object detections at a low cost, easy application |  |
| A Survey of deep learning methods and software tools for image classification and object detection | Image classification | Sparse coding, autoencoder (AE), restricted Boltzmann machine (RBM) | Compared deep learning models in image classification | Autoencoder model can perform coding when information is missing | Sparse coding cannot build features’ hierarchies |
| Lung nodule classiﬁcation using deep features in CT Images | Image classification | Binary tree classifier | Outperformed the state-of-the-art method on overall accuracy metric | Use a binary decision tree as a classified as it can handle missing information in the input as well |  |
| A food image recognition system with multiple kernel learning | Image recognition system | A multiple kernel learning method | Achieved the 61.34% recognition rate | A good example of food image recognition for practical application |  |

# **Chapter 2 Data Exploration**

## **2.1 Data Modeling and Plan**

We will use two methods to build the model: one method is to select one common model and train the five targeted objects differently; Another method is to train 5 models using different five targeted objects and then integrate them into one model to be used in the testing set. We will compare the results regarding the different methods and models we selected. Figure 11 and Figure 12 shows the process of data modeling plan in these two methods.

*Figure 11.* Data modeling for the first method.

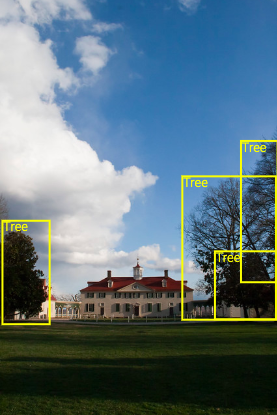
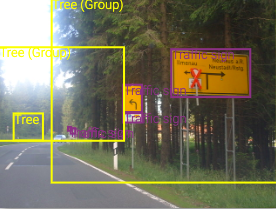
* Figure 12.* Data Modeling for the second method.

## **2.2 Data Sources and Dataset Parameters**

There will be five data sources in our project: utilities, all vehicles (cars), bikes (scooter, bikes, motorcycles), pedestrians, traffic lights (signs). Each object will include 1000 images for the training set and 200 images for the testing set. In total, there are 6000 images (5000 for the training set, and 1000 for the testing set). The parameters and the corresponding descriptions for all the objects are shown in the following table:

|  |  |
| --- | --- |
| Table 4  *Data Parameters Description* | |
| Parameters | Description |
| Image Size and Resolution | Vary from 640 \* 480 to 1280 \* 960 pixels |
| Color Scale | The image is grayscale or color scale (R, G, B). |
| Format | The image is in the format of .jpg or .png. |
| Label | The image is with the label originally or need to label manually. |
| Bounding Box | The objects in the image have the bounding box or not. |
| Annotation File | The images attach the annotation file or not, which describes the image id, file name, object coordinates, etc. |
| Background | The images are in different backgrounds or single background, like day and night. |
| Source | The images come from video clips, the photograph, or the image. |

**Utilities.** The utility data will include trees, palm trees, Christmas trees, and street lights. The source of the dataset will be Open Images Dataset V5, which includes about 9M images in total with image-level labels and bounding boxes provided. The dataset includes training, testing and validation parts which are easier to use. This data source has 16M bounding boxes included in 1.9M images data with 600 classes, so approximately 3667 images for each class. For each labeled image, there are several labeled bounding boxes. For the dataset with bounding boxes, the columns include label names and coordinators (XMin, YMin, XMax, YMax) for the bounding boxes.

*Figure 13.* Example of the target object “utilities”. 

**Vehicles.** The data type of vehicle datasets is a 3D object with labels. We can use them to detect multi-view object class and understand the scene. Computers can use fine-grained recognition to distinguish appearance differences (Li, 2018). All of the images are pre-processed. The founders use bounding boxes to crop the images to keep the interesting object in the center of the image (Krause, Stark, Deng, & Li, 2013). Classes include:

* Make
* Model
* Year.



*Figure 14.* Examples of the target object “vehicles”.

**Bikes (scooter, bikes, motorcycles).** Like vehicles, detection of bikes is also important in our recognition system. Because bikes have a smaller size than regular vehicles and sometimes, they are put in everywhere that they are more likely to be considered as obstacles for visual disturbances’ people. The datasets are mainly from these two sources:

* The Exclusively Dark (ExDARK) dataset: It is a collection of 7,363 low-light images from very low-light environments to twilight (i.e 10 different conditions) with 12 object classes (contains bikes) annotated on both image class level and local object bounding boxes
* Kaggle image classification competition: It contains cars, bikes, and motorbikes with 200 of each. Images are just taken from our daily life in all kinds of scenarios.



*Figure 15.* Examples of the target object “bikes”.

**Pedestrians.** Pedestrian Detection is a popular and widely applied algorithm in the field of image classification. The pedestrian dataset contains images from several different sources:

* Penn-Fudan Database for Pedestrian Detection and Segmentation: In this dataset, the image is selected from the street scene and campus street, and each image has at least one pedestrian.
* Caltech Pedestrian Detection Benchmark: This dataset is acquired by a vehicle camera and is divided into two data sets: detection and classification. It is shot by a car camera for about 10 hours. The resolution of the video is 640× 480, 30 frames/second.
* INRIA Person Dataset: This database is the most widely used static pedestrian detection database, providing original images and corresponding annotation files.
* Google Images are also a good data source to extract pedestrians’ images.



*Figure 16.* Examples of the target object “pedestrian”.

**Traffic lights (signs).** Traffic lights and traffic signs are an integral part for the blind person when they walk on the street, we filter out various of traffic signs, which are unnecessary for the blind person, since most of the traffic signs for drivers, thus only leave the “stop” sign and the traffic lights, more useful when they pass across the intersection. For traffic light datasets, we collected the images from day and night background, since traffic lights show different colors in the days and the nights, and these images are all attached the annotation files contains the information of the bounding box for each light position in the image as well as the stop sign images.

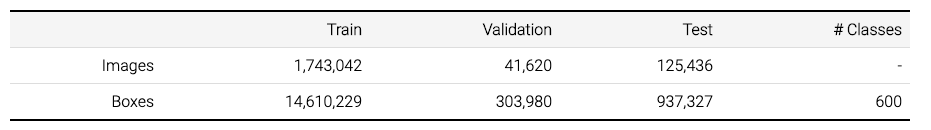


*Figure 17.* Examples of the target object “traffic lights”.

## **2.3 Collection and Training Datasets**

1200 images will be collected by data mining pipeline for each object and split to 1000 training data and 200 testing data. We will look for some CNN image packages and modify the Facebook detectron model to train our dataset.

**Utilities.** The utility dataset will include images, image labels, bounding boxes, and bounding boxes labels. After downloading the image level dataset and bounding box dataset, we will need to join the two tables to get both image labels and bounding box labels. Since the dataset includes 1.9M images and 16M bounding boxes with 600 classes. We will need to filter out for the classes we want, approximately 3667 images for each class. These data are split into training, testing, and validation.



*Figure 18.* The structure of training, validation and testing dataset of images and boxes for utility dataset.

**Vehicles.** The vehicle dataset is collected from Stanford Car Database (Krause, Stark, Deng, & Li, 2013). The car dataset contains 16,185 images in 196 classes of cars. These are split to 50-50 for training and testing. We will only use the training images as our training data.

**Bikes (scooter, bikes, motorcycles).**

* The Exclusively Dark (ExDARK) dataset: It is a collection of 7,363 low-light images from very low-light environments to twilight (i.e. 10 different conditions) with 12 object classes (contains bikes) annotated on both image class level and local object bounding boxes
* Kaggle image classification competition: It contains cars, bikes, and motorbikes with 200 of each. Images are just taken from our daily life in all kinds of scenarios.

**Pedestrians.** Since pedestrian detection is one part of our model design, we need to select 1200 images which are better and clear to train the model. From the pedestrian data source, we choose, we will use 1000 images to be the training set, and 200 images to be the testing set.

* Penn-Fudan Database for Pedestrian Detection and Segmentation: In this dataset, the image is selected from street scene and campus street, and each image has at least one person, containing 345 pedestrians with a total of 170 images.
* Caltech Pedestrian Detection Benchmark: This dataset is acquired by a vehicle camera and is divided into two data sets: detection and classification. In total, there are 15560 images which contain at least one person, and 6744 images meaning negative samples from the training set. For the testing set, it consists of about 27 minutes of video, including 21790 images, and 56492 pedestrians are labeled.
* INRIA Person Dataset: In total, there 711 images and 3542 pedestrians who are labeled, containing original images and corresponding annotation files.

**Traffic lights (signs).** For the traffic light and the stop sign dataset, they were collected by downloading from the internet and by using video to jpg converter to get images from the video. As the training data, there are 780 stop signs and 2161 traffic light images, 200 stop signs and 700 traffic light images as the validation data, and 150 stop signs and 500 traffic light images as the test data.

## **2.4 Data Cleansing and Validation**

The purpose of data cleaning is to improve the quality of the dataset by removing duplicates and choosing the images to be easier for training. After data cleaning, the dataset should be clearer and more reliable for analytical use.

|  |  |  |
| --- | --- | --- |
| Table 5  *Data Cleansing and Used Tools* | | |
| Images | Data Cleansing | Tools |
| Utilities | - We will collect our raw image data and remove duplicated images.  - Then, we will scan the images one by one to find whether there are some invalid pictures like the picture background is very blurry or the target object is not very clear for us to train the model.  - Finally, we will label the images which miss the correct labels and choose 5,000 training images, with each object containing 1,000 images. | - Labelme (an image annotation tool, implemented by Python and PyQt5).  - d2l library to draw the bounding box.  - Numpy library to randomly print and validate images. |
| Vehicles |
| Bikes |
| Pedestrians |
| Traffic Lights |

For the validation process, we will choose 8 sample images for each object randomly, and manually validate the images with their corresponding labels to verify whether each image matches their labels and annotations and to check whether each class has the same number of images.

## **2.5 Data Transformation and Tools**

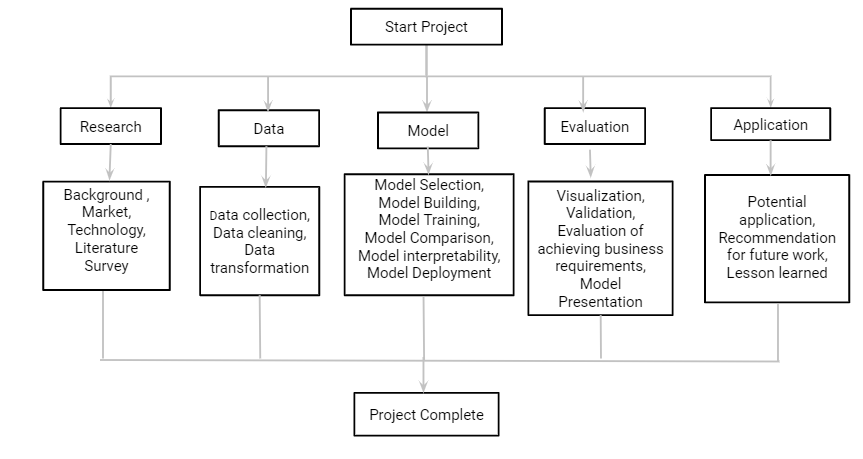
Since our data are all images with the corresponding labels, containing the fives objects of utilities, vehicles, bikes, pedestrians, and traffic lights, we will transform the data in a universal and general approach according to the image transformation principles:

* We will need to transform the images to be suitable for model training, as the images should have the same dimensions. Also, the color scale should be identical as well to make the images standardized. To do so, we will build data pipelines to load and transform the original images into an acceptable format for our models. We will use tools like Numpy, Scikit-Learn, PIL, Torchvision and other packages in Python Libraries in our project.
* There are numerous preprocessing techniques to transform our images to be easier to be recognized by the machine learning model: One method is to remove the background color of the images to reduce noise; Another method is to only highlight the target object to make the images more pointed. Python Libraries including Pytorch, Numpy, and other packages will be used.
* In order to let the images be trainable by machine learning models, we will need to transform the image dataset into an array form of numbers that tells the color information. OpenCV Python interface will be used to convert RGB images to Numpy arrays.
* For training and data validation, we will randomly split the remaining non-test data into a training set and a validation set, containing 85% and 15% of the images, respectively. The tools we will use containing Python Libraries, Ubuntu, Spark, and other techniques.

# **Chapter 3 Project Management**

## **3.1 Project Organization**

We use a breakdown workflow to show our project structure clearly.



*Figure 19.* Project organization workflow.

## **3.2 Resource Requirements**

**Data Requirement.** We will use image datasets that all publish publicly which can be easily used and accessed.

* Open Images Dataset V5: A public dataset for large-scale multi-label and multi-class image classification.
* Stanford Car Datasets: Stanford University AI Lab.
* The Exclusively Dark (ExDARK) dataset: University of Malaya.
* Penn-Fudan Database for Pedestrian Detection and Segmentation: University of Pennsylvania and Fudan University.
* Caltech Pedestrian Detection Benchmark: California Institute of Technology (CIT).
* INRIA Person Dataset: Used in the ICCV 2009 paper on image auto-annotation.
* Lisa Traffic Sign Dataset: Laboratory for Intelligent & Saft Automobiles.

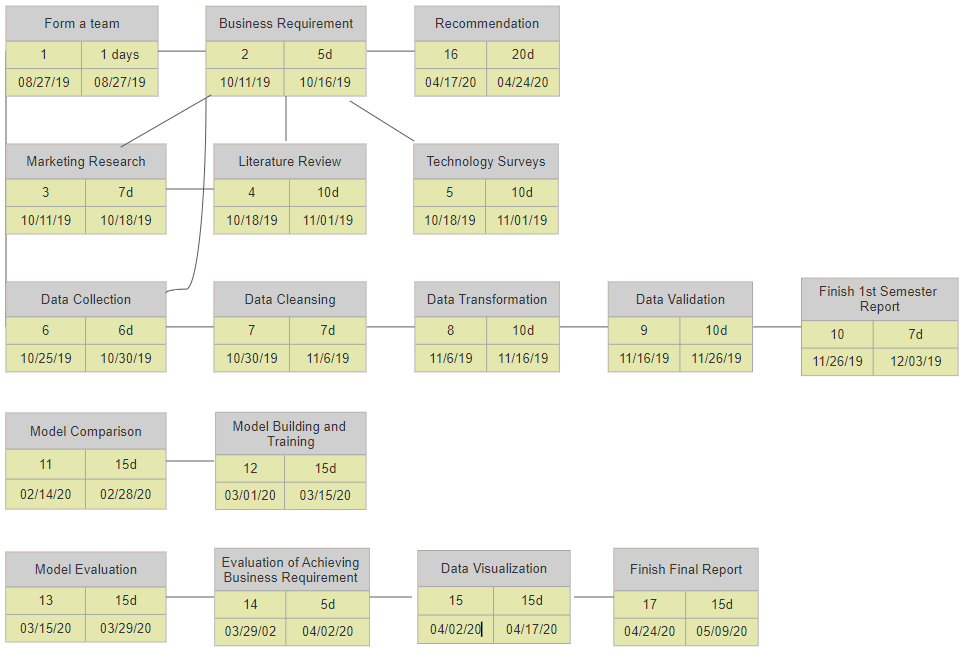
**Tools Used.** Python will be our major programming language, and Google Colab will be used to share our code and keep the result. Due to the libraries and the packages, it depends on the models we chose, we will leverage the hardware, tools, software, and packages in general, which are shown in the following tables:

|  |  |  |
| --- | --- | --- |
| Table 6  *Hardware Requirements* | | |
| Hard Drive | CPU | Memory |
| 1T | Intel i5/i7 | Colab: 24GB GPU RAM |

|  |  |
| --- | --- |
| Table 7  *Software Requirements* | |
| Tools, platform, and software | Packages and libraries |
| - Google Colaboratory  - Python (language)  - Google Drive (document sharing)  - Labelme (label the images) | Numpy, Pandas, Keras, Scikit-Learn, Matplotlib, TensorFlow, Seaborn, and other necessary packages. |

## **3.3 Project Schedule**

We use a PERT chart to organize our project’s tasks and timelines.

****

*Figure 20.* Project detailed schedule.

# **References**

Bates, D. (2017, February 22). 10 Ways to Engage the Disability Market and Keep a Competitive Edge. Retrieved from <https://austinstartups.com/10-ways-to-engage-the-disability-market-and-keep-a-competitive-edge-62a725634ecf>

Braille Works. (2015, March 5). Accessible Marketing – Are You Missing 38 Million Americans? Retrieved from <https://brailleworks.com/accessible-marketing-are-you-missing-38-million-americans/>

Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1209-1218).

Caltech Pedestrian Detection Benchmark. (n.d.). Retrieved from <http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/>

Courtney-Long, E. A., Carroll, D. D., Zhang, Q. C., Stevens, A. C., Griffin-Blake, S., Armour, B. S., & Campbell, V. A. (2015). Prevalence of disability and disability type among adults—United States, 2013. *MMWR. Morbidity and mortality weekly report*, *64*(29), 777.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743-761.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2009). Pedestrian detection: A benchmark.

Druzhkov, P. N., & Kustikova, V. D. (2016). A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, *26*(1), 9-15.

Gandhi, R. (2018, July 9). R-CNN, Fast R-CNN, Faster R-CNN, YOLO - Object Detection Algorithms. Retrieved from <https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e>

Global data on visual impairment. (2017, December 8). Retrieved from <https://www.who.int/blindness/publications/globaldata/en/>

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Gu, C., Arbeláez, P., Lin, Y., Yu, K., & Malik, J. (2012, October). Multi-component models for object detection. In *European Conference on Computer Vision* (pp. 445-458). Springer, Berlin, Heidelberg.

Hussin, R., Juhari, M. R., Kang, N. W., Ismail, R. C., & Kamarudin, A. (2012). Digital image processing techniques for object detection from complex background image. *Procedia Engineering*, *41*, 340-344.

ImageNet. (n.d.). Retrieved from <http://image-net.org/index>

INRIA Person Dataset. (n.d.). Retrieved from <http://pascal.inrialpes.fr/data/human/>

Joutou, T., & Yanai, K. (2009, November). A food image recognition system with multiple kernel learning. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (pp. 285-288). IEEE.

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 554-561).

Kumar, D., Wong, A., & Clausi, D. A. (2015, June). Lung nodule classification using deep features in CT images. In *2015 12th Conference on Computer and Robot Vision* (pp. 133-138). IEEE.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

LISA traffic lights and traffic signs dataset. (n.d.). Retrieved from <http://cvrr.ucsd.edu/LISA/datasets.html>

Malik, U. (2019, October 21). 5 Electronic Glasses for the Blind and Visually Impaired. Retrieved from <https://irisvision.com/electronic-glasses-for-the-blind-and-visually-impaired/>

Manoj, M., Neelima, M., Harshali, M., & Rao, M. V. G. (2018). Image classification using Deep learning. *vol*, *7*, 614-617.

Open Images Dataset V5. (n.d.). Retrived from <https://storage.googleapis.com/openimages/web/factsfigures.html>

Penn-Fudan Database for Pedestrian Detection and Segmentation. (n.d.). Retrieved from <https://www.cis.upenn.edu/~jshi/ped_html/>

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Reid, O. (2018, May 12). Facebook Detectron. Retrieved from <https://dlfor.blogspot.com/2018/05/facebook-detectron.html>

Sachan, A. (2017). Zero to Hero: Guide to Object Detection using Deep Learning: Faster R-CNN, YOLO, SSD.

Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. In *Advances in neural information processing systems* (pp. 2553-2561). Yolo: Real-time object detection. (n.d.). Retrieved from <https://pjreddie.com./>

Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). Contextual models for object detection using boosted random fields. In *Advances in neural information processing systems* (pp. 1401-1408).

Xu, M., Cui, L., Lv, P., Jiang, X., Niu, J., Zhou, B., & Wang, M. (2018). Mdssd: Multi-scale deconvolutional single shot detector for small objects. *arXiv preprint arXiv:1805.07009*.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*.