Object Detection & Image Mosaicing for Video Content Summarization

Abhanshu Gupta
Department of Computer Sciences
University of Wisconsin-Madison
abhanshu@cs.wisc.edu

Varun Sah
Department of Computer Sciences
University of Wisconsin-Madison
varun@cs.wisc.edu

April 3, 2018

1 Introduction and Motivation

With the advent of smart phones with high-quality cameras, video and image content is being generated at an unprecedented rate, rendering the task of manual content identification and classification virtually impossible. Content identification and classification tasks impact human lives in many ways. Content categorization enables delivery of personalized content based on user preferences. At the same time, it is instrumental in maximizing revenue for several industries. For instance, the advertising and marketing industry is a particularly visual world, with millions of images and videos displayed everyday within websites, television programs, and movies, in order to expose consumers to the latest trends and products.

The sheer volume of media existing today motivates the need for automated software systems that leverage state-of-the-art artificial intelligence (AI) and computer vision (CV) techniques for accomplishing the content cataloging task efficiently. AI and CV based software allow for visual product discovery and categorization, thereby reducing the reliance on manually entered, subjective, noisy product meta-data. Their ability to group similar products based on their visual affinity makes the process of categorization objective, noise-free and exponentially faster as compared to methods that require human intervention. However, like many other computer vision problems, a single approach that is universally considered the obvious or "best" method to address the problem of content categorization efficiently and effectively is lacking.

Another problem closely associated with video and image categorization is that of content summarization. There is a growing need for automatically generating aesthetically pleasing visualizations that are representative of categorized content and provide the viewer with a preview or overview of the actual content in the media. This project is aimed at efficient and effective summarization of video content by means of real time object detection and image mosaicing.

2 Problem Description

This project is aimed at enabling automatic content summarization of videos by creating a representative mosaic of the video's object of focus. We intend to do this in a three-step process involving real time object detection, salient object identification and image mosaicing. Each of these steps is described in detail in the forthcoming section.

3 Methodology

Our proposed content summarization pipeline comprises of three modules namely Object detection, Salient Object Identification and Image Mosaicing.

3.1 Object Detection

Object detection is the problem of finding and classifying a variable number of objects on an image. Object detection has proven to be a hard problem as compared to classification, since its output is variable in dimensions due to the inherent differences in object size and number across images. Object detection in videos is an even more challenging task than object detection in images primarily due to the motion and blur effects that result in detection failures on certain frames.

A traditional method for object detection is using Histogram of Oriented Gradients(HOG) features and Support Vector Machine (SVM) for classification. It requires a multi-scale sliding window making it much slower. In recent years deep learning has been a real game changer in computer vision. Deep learning models have virtually replaced classical techniques for the tasks of image classification and object detection and are currently an active area of research in computer vision. Many deep learning models are already in place that are state of the art in object detection. These models are fast and provide high accuracy and detection efficiency. In this phase, we will modify and build on state-of-the-art frameworks like R-CNN. R-FCN, YOLO and SSD for application-specific object detection.

3.2 Salient Object Identification

After obtaining the output of the object detection module, we will extract information about the content and salient objects of the video based on a mixture of heuristics and learned decisions. In this phase, we intend to determine the content of the video by analyzing identified object categories, their frequency of occurrence and the mutual relationship between objects detected in the video. The output of this module will be a list of tags or categories and salient objects along with their frames of occurrence present in the video.

3.3 Image Mosaicing

Mosaicing is an old art technique where pictures or designs are formed by inlaying small bits of colored stone, glass, or tile. These small bits are visible close up, but the boundaries of the bits will blend and a single recognizable image will show at a distance. In the modern digital world, this ancient art form has been transformed and combined with new technologies. Instead of using pure-colored blocks, entire images can be used as tiles to make an overall pictures.

After obtaining the output of the Object Detection and Salient Object Identification modules, we will choose a representative image as our target image. Using other occurrences of the same object (as well as other significant objects), we will reconstruct the target image in the form of a mosaic. This aesthetic visualization would be our final output that succinctly provides a visual preview of the video content.

4 Results

The object detection and salient object identification modules were tested on 4 commercials advertising a car (Mercedes Benz), a laptop (Microsoft Surface), a phone (Windows), and a tennis racket (Wilson). Each advertisement was obtained from the respective manufacturer's youtube channel.

4.1 Object Detection

After exploring several techniques for reliable object detection,we chose to leverage a yolov2 [RF16] model. The model was pre-trained for object recognition of 80 classes present in Common Objects in Context (COCO) [LMB+14] dataset. Architecturally, the trained network consists of 23 convolution layers, 2 route layers and 1 recognition layer. In particular, we chose the darkflow python implementation using the Tensorflow framework for object detection in this project. For each of the commercials under consideration, we down-sampled the number of frames per second (fps) by a factor of 3, reducing 24 fps to 8 fps. While the network itself has the potential of supporting processing upto 40 fps, in the absence of access to machines with GPUs, we ran the model on a 4-core CPU machine with a processing speed of about 1 fps.

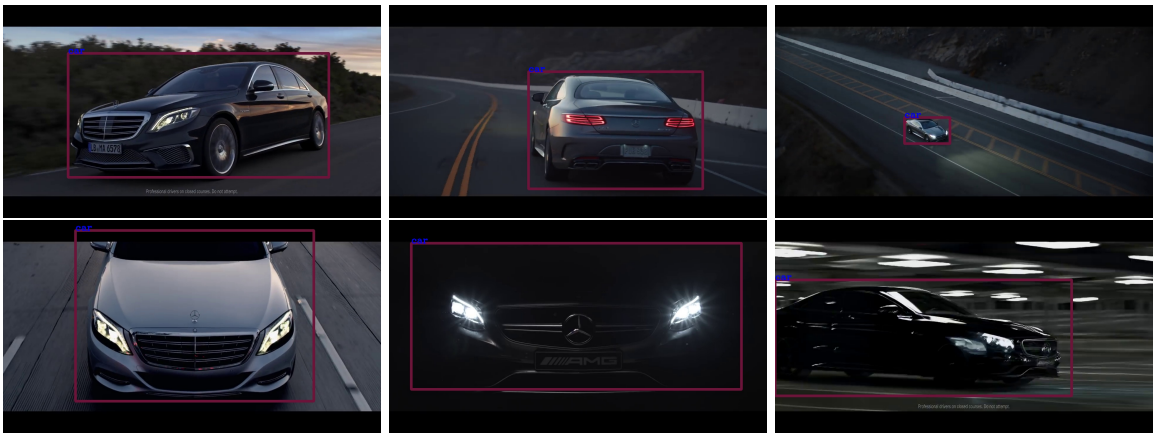

Figure 1: Results on the Mercedes-Benz car commercial: Detection of cars when viewed partially or completely, from different angles & distances, under varying lighting & blur conditions

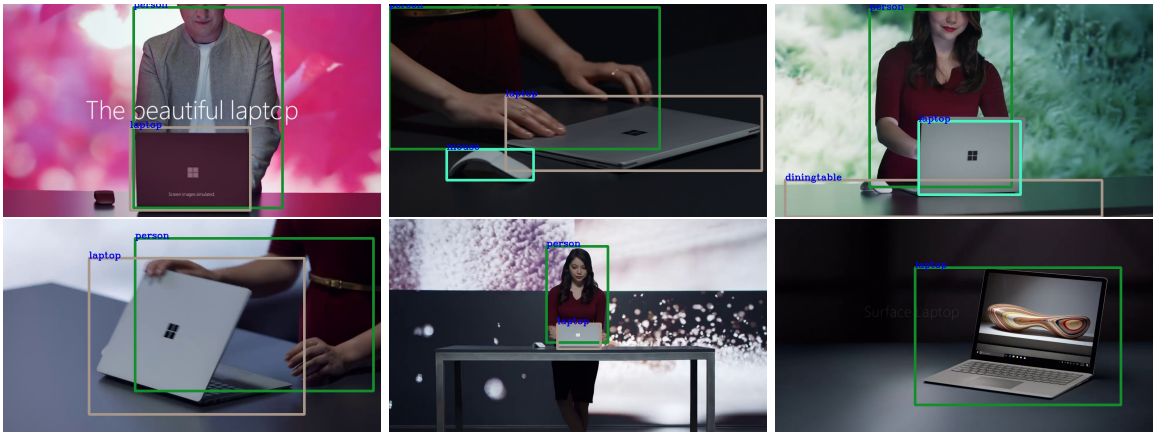

Figure 2: Results on the Microsoft Surface laptop commercial: Detection of laptops when viewed partially or completely, from different angles & distances, under varying lighting & blur conditions

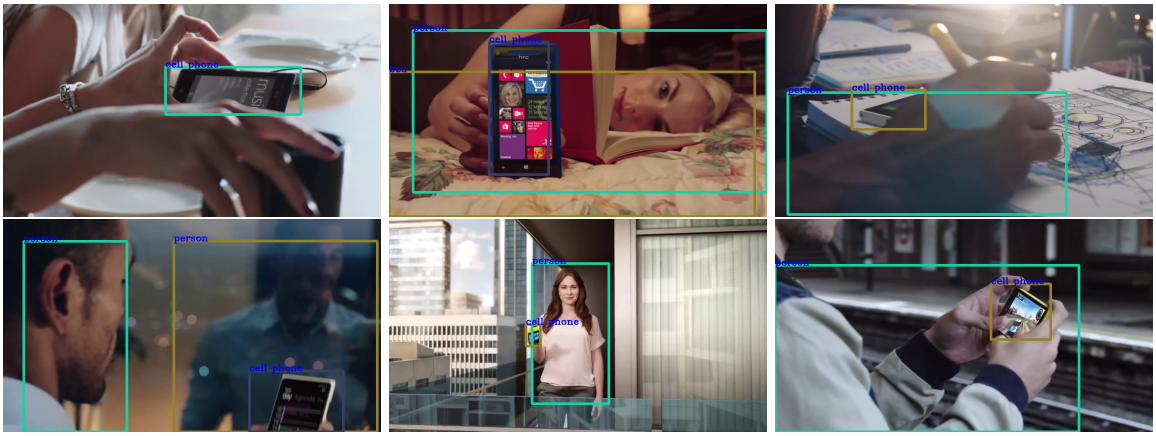

Figure 3: Results on the Microsoft Windows phone commercial: Detection of cell phones when viewed partially or completely, from different angles & distances, under varying lighting & blur conditions

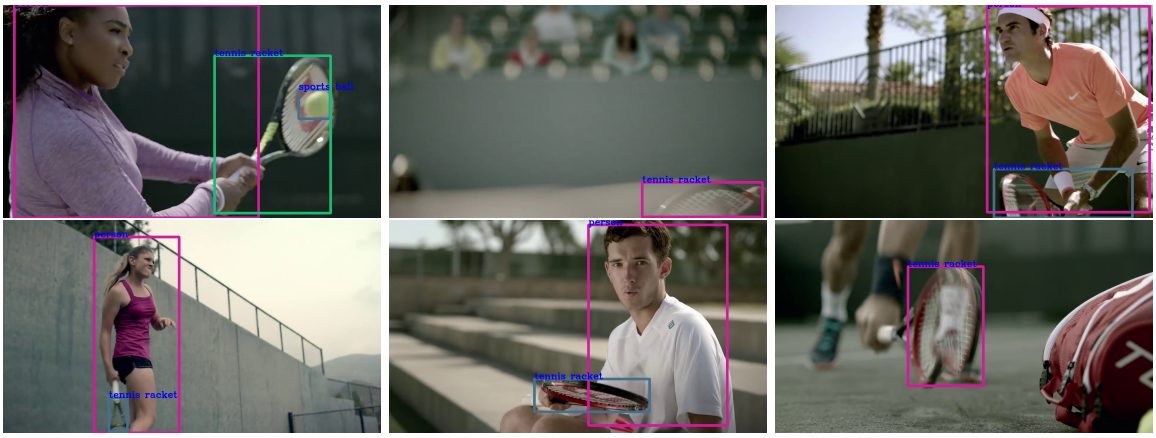

Figure 4: Results on the Wilson tennis racket commercial: Detection of tennis rackets when viewed partially or completely, from different angles & distances, under varying lighting & blur conditions

Figures 1, 2, 3, and 4 illustrate the performance of the object detection module on 4 commercials advertising a car (Mercedes Benz), a laptop (Microsoft Surface), a phone (Windows), and a tennis racket (Wilson) respectively. It is worth noting that several of these objects are correctly detected even in unfavorable visual conditions such as poor lighting, motion-blur, and significant obstruction by other objects. While the accuracy is remarkable, the speed of classification leaves much to be desired due to the limited computational resources available to us.

4.2 Salient Object Identification

For identifying the object of focus in commercial advertisements, we started by filtering objects identified with a confidence of less than 50%. Additionally, we filtered out object classes based on their frequency of occurrence in the video. Objects detected in only a handful of frames were successfully filtered out due to our thresholding strategy.

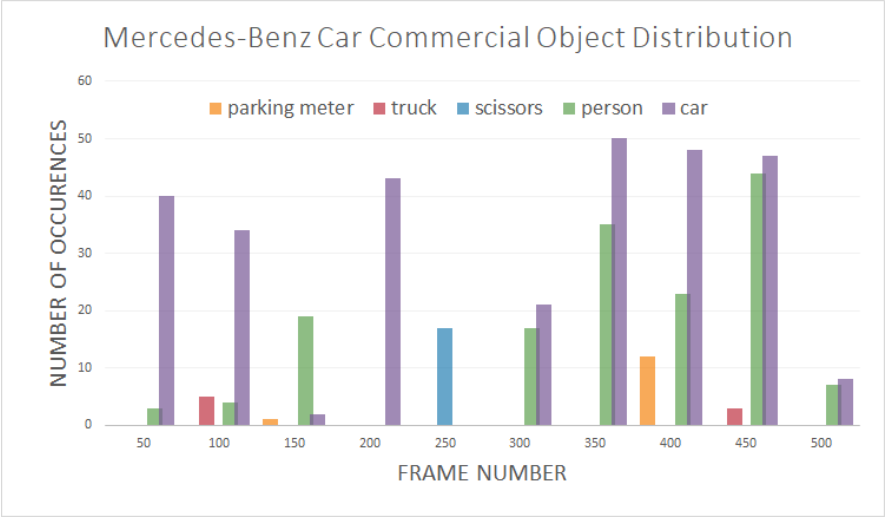

Figure 5: Results on the Mercedes-Benz car commercial: Binned frame-wise object distribution graph of top 5 most frequently detected objects

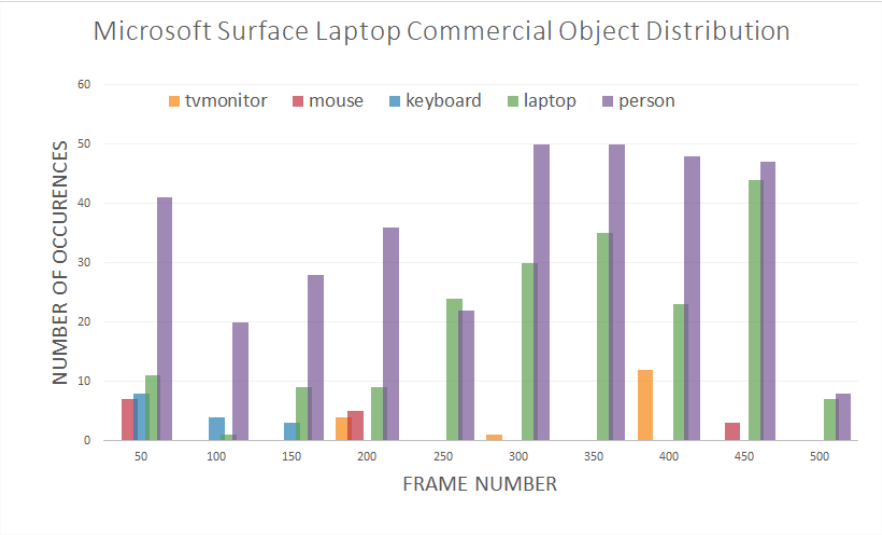

Figure 6: Results on the Microsoft Surface laptop commercial: Binned frame-wise object distribution graph of top 5 most frequently detected objects

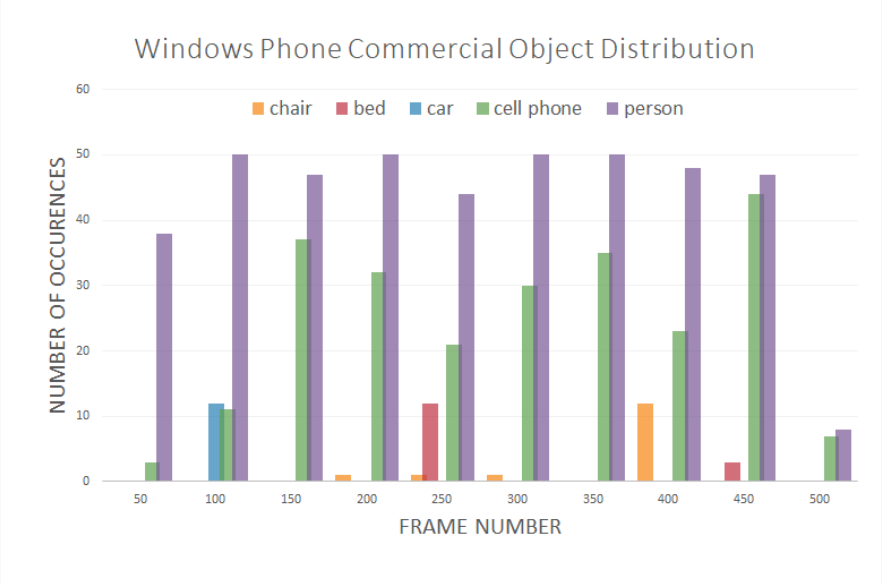

Figure 7: Results on the Microsoft Windows phone commercial: Binned frame-wise object distribution graph of top 5 most frequently detected objects

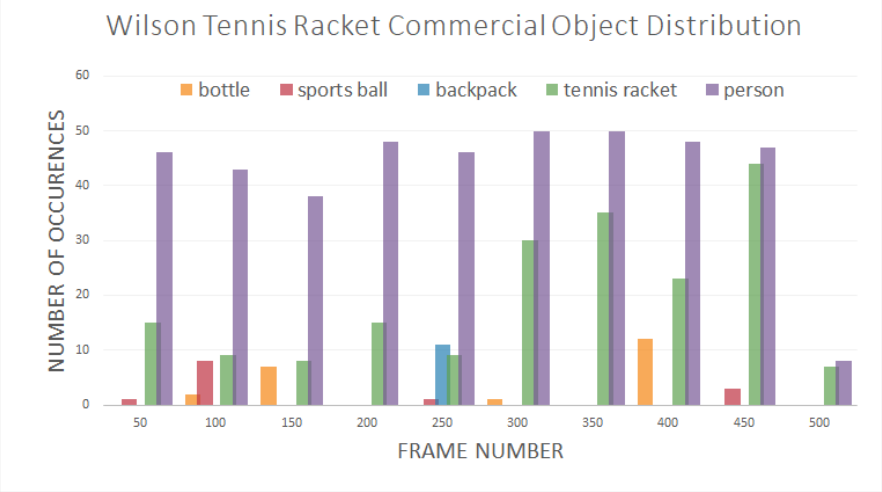

Figure 8: Results on the Wilson tennis racket commercial: Binned frame-wise object distribution graph of top 5 most frequently detected objects

Figures 5, 6, 7, and 8 illustrate the binned detected-object-distribution by time (frame number) in 4 commercials advertising a car (Mercedes Benz), a laptop (Microsoft Surface), a phone (Windows), and a tennis racket (Wilson) respectively. The car advertisement (Fig 5) was a basic case where the most frequently detected object was a car itself, and, hence, did not require any further estimation for determining the salient object. However, in each of the other commercials the most frequent object category was *person*. A naive most-frequent estimation would erroneously label these advertisements as being people-centric. However, using our heuristics, each of these commercials was correctly labeled as being about the actual object class to which it belonged.

5 Discussion of Challenges

The following subsections include a discussion of some of the challenges we faced during the project thus far.

5.1 Incorrect object detection

Despite having remarkable accuracy, the object detection module was by no means perfect in its classification of objects. A major cause of these mis-classifications was found to be motion-blur and scene-transition blurring in videos that possessed slow or smooth fade-in/fade-out transitions between different scenes. This transition-blur caused the image to possess artifacts that were a combination of artifacts from separate scenes. Figure 9 illustrates misclassification due to motion-blur and scene-transition blur.

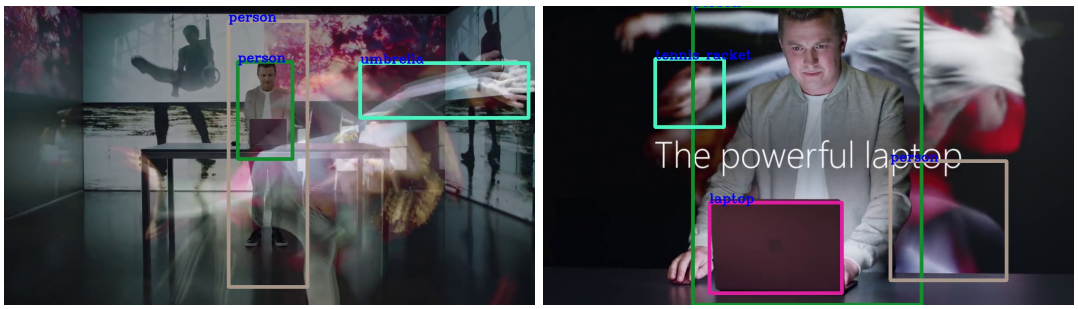

Figure 9: Object misclassification due to motion-blur and scene-transition blur

Another classification error was found to be occur in peculiar circumstances wherein a flexible object was found to metamorphose into a shape that represents a different entity. For instance, Figure 10 illustrates misclassification of a human arm (shaped like elephant's ears and trunk) as an elephant, the Mercedes logo mistaken for a clock, and a bag (shaped like a horse's snout) labelled as a horse.

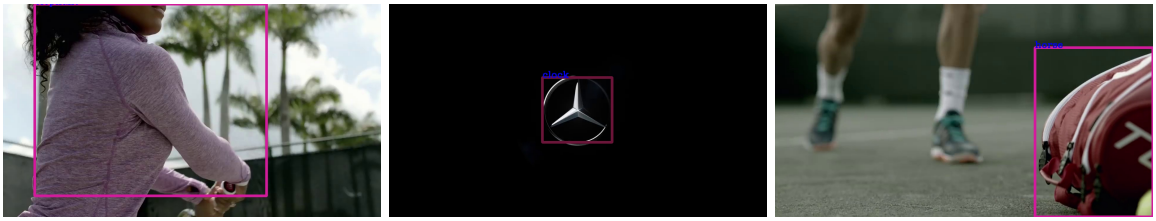

Figure 10: Object misclassification due to motion-blur and scene-transition blur

In order to handle each of the aforementioned imperfections in classification, we suppressed objects detected during the salient object identification phase using a composite of frequency and confidence thresholding (described in 4.2). Detected objects falling short of these thresholds were no longer considered candidates for being the salient object of focus.

5.2 Salient Object Heuristics

Since most advertisements feature humans in some capacity or the other, people were often the most frequently detected object in commercials. Heuristically, we decided to disqualify people from being candidates for the object of focus. More generally, this heuristic can be extended to most living creatures who are seldom advertised in commercials. It is worth noting that this strategy is applicable only in the presence of an object that has a significant frequency even if that frequency is lower than that of *person* class. This would ensure that mobile phone commercials, despite featuring more humans than phones, would still be classified correctly, while political campaign advertisements would be accurately identified as being people-centric.

5.3 Limited Computation Power

Lack of computational power due to absence of access to GPU-equipped machines was identified to be a common factor that limited our ability to improve certain undesirable outcomes. The classification time for videos currently takes longer than the length of the video, which is undesirable for scalability reasons. This could be avoided if machines with GPUs were accessible to us since the object identification task naturally lends itself to parallelization. Lack of computing power is also the primary reason for avoiding repeated iterative training or tweaking of network architecture due to the huge training and testing time overheads.

6 Timeline

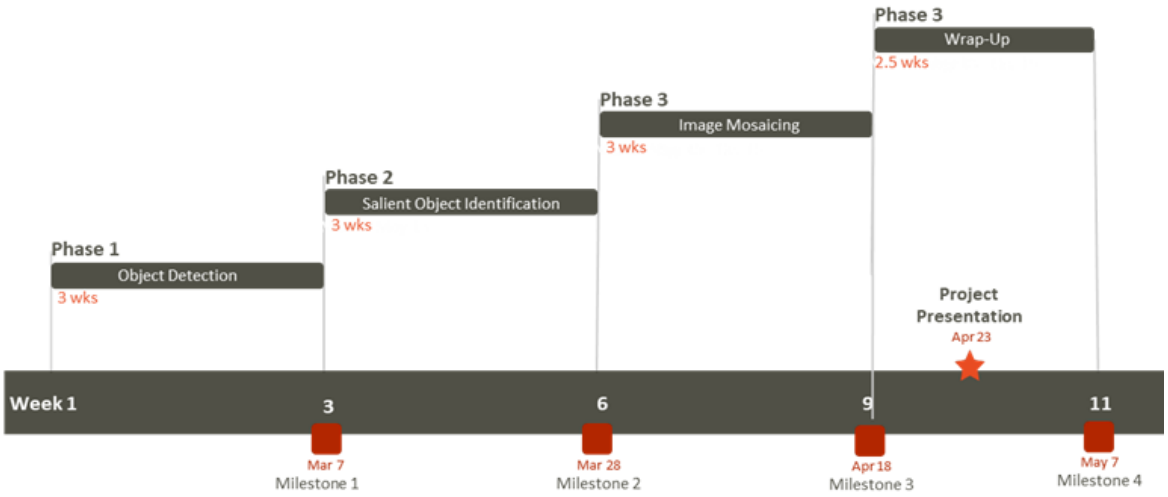

Figure 11: Anticipated Project Timeline

References

[DLHS16] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[Gir15] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[LAE+15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[RDGF15] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[RF16] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[RHGS15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[VN14] Dushyant Vaghela and Kapildev Naina. A review of image mosaicing techniques. *CoRR*, abs/1405.2539, 2014.