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

February 15, 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 summarization 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 Timeline

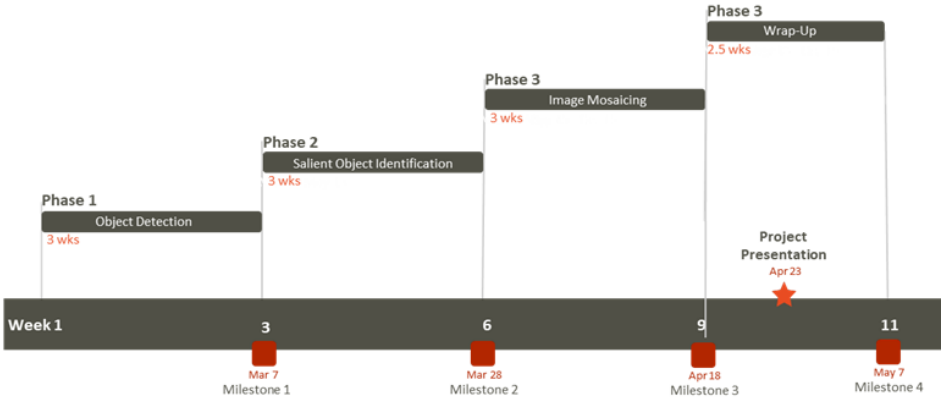

Figure 1: Anticipated Project Timeline

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.

[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.

[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.