WEB-SCALE VISUAL CONTENT MINING WITH COMPUTER VISION METHODS

Dissertation Proposal

ABSTRACT

Social media provides low-cost and easy access to a large number of images with a wide variety of visual content. With this new source of large-scale image data, there is a critical demand for novel, scalable methods that can understand the information behind the visual content. For example, photos shared on social media are usually accompanied with temporal and location metadata, user profiles and textual annotations. When we integrate the visual content by user profiles or location tags, we can estimate a large variety of information. For example, we can help research in Psychology when we are interested in the human beings who take these photos and in Ecology when we are interested in the natural scenes as the visual content. These attractive properties of this new data source also bring us new challenges. Being able to efficiently processing data with advanced computer vision techniques is an important requirement. To accommodate the graphical structure of metadata in social media, we also need to develop new models and systems.

This thesis introduces three projects: a cloud based robotic system testing the feasibility of a memory and computing consuming task to work in near real time; a data mining application supporting Psychology research in gender differences in preference of color; and a natural event tracking system on a continental scale. We design and implement a cloud based system to apply these algorithms for object detection and recognition with the best precision in near real time. We test our method on an aerial autonomous vehicle. We take advantage of this new data source for psychology research. Based on the color pixels of these photos, we analyze the gender difference in color preference. We study the distribution of color pixels with respect to genders of photographers across geographic locations and content. We find strong sex differences for the predominant reddish and bluish hues, with female users uploading more photographs containing more reddish pixels and male users uploading more photographs containing more bluish pixels. Furthermore, we take Google

Street View to represent the color distribution in the environment, and compare with Flickr photos in three popular outdoor locations. We observe the overall preference of saturated color and reddish color of human compared to the environment. By mining visual content on social media websites web-scale images further benefit studies about the state of nature. We develop and evaluate three models to study ecology phenomena such as snow and vegetation coverage. First, we learn a binomial distribution to model the probability of the appearance of a natural phenomenon. Then, we compute the histogram of the confidence of each image being a positive evidence. Based on this histogram, we learn a classification model for the confidence of each user, and similarly apply this method to predict for each time and location. Finally, inspired by the study of multiple instance learning, we propose an end to end system taking a sequence of images as input and giving predictions as the final output of this holistic system. This process will happen twice, once for users and once for each day and location. The two aggregating processes are optimized simultaneously.

CONTENTS

CHAPTER 1

INTRODUCTION

Social media encourages the development of online social network, connecting billions of users sharing status and comments in text and multimedia format. These increasingly large number of multimedia files such as photographs become a new source of web-scale data with high variety and publicly available. Along with the image files and text annotations, social networking websites and mobile applications also record user profiles as well as time and location metadata of users activity. This new data source is quite noisy but is also easy accessible. It collects data constantly from a broad selection of locations on this earth. To better understand the information behind this large collection of data, there is a critical demand for novel, scalable methods to navigate the visual content together with the textual annotation and user activity. Researchers from computer vision, multimedia, data mining and machine learning become more and more interested in this interdisciplinary area, especially with the progress in computer vision and with the development of computing capacity.

When we integrate the visual content by the metadata of user activities, we will provide a very low-cost, large scale data source for researchers in many areas of social and natural science. For example, we will help research in social science when we are interested in the human behavior of those who take photos and share on social media, and in nature science

when we are interested in the natural scenes and the associated location, time and user profile. These attractive properties of this new data source also bring us new challenges. Being able to efficiently processing data with advanced computer vision techniques is an important requirement. To accommodate the graphical structure of metadata in social media, we also need to develop new models and systems.

This thesis presents our works of solving problems in three aspects to facilitate research in different areas with web-scale image data.

– In **Chapter 2** we design and implement a cloud based system to apply these algorithms for object detection and recognition with the best precision in near real time. We test our method on an aerial autonomous vehicle.
– In **Chapter 3** we take advantage of this new data source for psychology research. Based on the color pixels of these photos, we analyze the gender difference in color preference. We study the distribution of color pixels with respect to genders of photographers across geographic locations and content. We find strong sex differences for the predominant reddish and bluish hues, with female users uploading more photographs containing more reddish pixels and male users uploading more photographs containing more bluish pixels. Furthermore, we take Google Street View to represent the color distribution in the environment, and compare with Flickr photos in three popular outdoor locations. We observe the overall preference of saturated color and reddish color of human compared to the environment.
– In **Chapter 4** we present an example of web-scale images further benefit studies about the state of nature. We develop and evaluate three models to study ecology phenomena such as snow and vegetation coverage. First, we learn a binomial distribution to model the probability of the appearance of a natural phenomenon based on the number of users uploading images with and without the target phenomenon. Then, we compute

the histogram of the confidence of each image being a positive evidence. Based on this histogram, we learn a classification model for the confidence of each user, and similarly apply this method to build a histogram of user confidence in order to predict for each time and location. Finally, inspired by the study of multiple instance learning, we propose an end to end system taking a sequence of images as input and giving predictions as the final output of this holistic system. For the main challenge of multiple instance learning, we use a network to aggregate a bag of evidence. This process will happen twice, once for users and once for each day and location. The two aggregating processes are optimized simultaneously.

CHAPTER 2

REAL-TIME, CLOUD-BASED OJBECT DETECTION FOR UNMANNED AERIAL VEHICLES

2.1 INTRODUCTION

Web-scale visual data provides massive amount of information while requiring more complex models to take this advantage, therefore, whether we can apply computer vision techniques efficiently enough becomes a bottleneck question. Robotics is a typical area in need of near real-time performance. Unmanned Aerial Vehicles (UAVs), as a type of the autonomous robotics systems, are increasingly interesting to researchers in recent years. They are widely used in applications including reconnaissance and surveillance, search-and-rescue, and infrastructure inspection. Visual object detection is an important component of such UAV applications. Moreover, it's also very challenging because of noisy image quality with complex scene and most importantly, the conflict of the near real-time performance requirement and the relatively long running time of advanced object recognition techniques. Solving this problem not only improves the computer vision component on robotics systems, but also

extend the object recognition models based on very large dataset to near real-time applications.

Object recognition performance is rapidly improving mostly based on Deep Learning techniques with Convolutional Neural Networks. These deep models are trained with very large datasets and are typically consist of millions or billions of parameters. These computational demanding techniques require the support of hardware including gigabytes of memory and high-end Graphics Processing Units (GPUs). Even though, the techniques with best accuracy are still far away from near real-time. According to these computation demand and running time problems, it's infeasible to apply these deep models to drones featuring low-cost and light-weight. Numerous studies have explored the benefits of "Cloud Robotics". Cloud computing allows on-demand access to nearly unlimited computational resources, which is especially useful for customized hardware combination and periodical requirement of huge amounts of computation. In this project, we build our Deep Learning facilitated cloud system to fulfill the hardware requirement and solve the efficiency problem. In fact, because of the unpredictable network delay due to the communication with a remote cloud, we build a hybrid system especially with onboard processing for critical tasks requiring immediate reaction such as stability control.

2.2 APPROACH

We use a Parrot AR.Drone 2.0 as a low-cost hardware platform to test our cloud-based recognition system. It is small and lightweight, and can be operated both indoors and outdoors. The AR.Drone 2.0 is equipped with two cameras. The bottom-facing camera with lower resolution of 320×240 while the front-facing camera has a higher resolution of 1280×720. To allow this drone to see objects on the ground, we mount a mirror at a $45°$ angle to the front camera as in **??**.

Our approach consists of four main components shown at top **??**. Each component is implemented as a node in the Robot Operating System (ROS), allowing it to communicate with others using the ROS transport protocol. The controlling components and objectness estimation component are running on a laptop connected to the drone through the AR.Drone device driver package of ROS, over a WiFi link. On the other hand, the most computationally demanding component, the CNN based object detection node, runs on a remote cloud computing server that the laptop connects to via Internet. The bottom of **??** shows the pipeline of image processing in our hybrid approach. The drone takes off and starts to search with the downward-facing camera. Given input video taken from this downward-facing camera, the objectness estimator node runs the BING algorithm to detect generic objects of every frame, and then takes a high resolution image with the front-facing camera if it detects candidate objects in the frame. Therefore, only the "interested" images that have a high likelihood to contain objects are sent to the cloud server, where the CNN based object detection node is running to recognize the target objects.

2.3 EXPERIMENTSAL RESULTS

We conducted three sets of experiments to demonstrate that our approach performs successfully in a realistic but controlled environment. In the first set of experiments, we focus on testing the accuracy of recent deep network based object detectors with aerial images taken by the drone. Secondly, we evaluate the speed of our cloud based object detection approach, comparing with running time of the fastest deep learning based object detector on a local laptop. Finally, we verify our approach with the scenario of a drone searching for a target object in an indoor environment, as a simple simulation of a search-and-rescue or surveillance application. The first two sets of experiments were conducted on our aerial image dataset and the last experiment was conducted in an indoor room of about $3m \times 3m$.

2.3.1 OBJECT DETECTION ACCURACY

We first compared the ability of Faster R-CNNs with two recent state-of-the-art object detectors(YOLO and SSD) to recognize aerial images taken by the drone. YOLO and SSD are approaches that achieving real-time performance (faster than 30 FPS) on GPU by eliminating the most computationally demanding part(generating region proposals and computing CNN features for each region).

We collected 294 aerial images of 20 object classes and annotated 578 objects in the images. These images have the same object classes as the Pascal VOC 2007 dataset and are collected from two sources (some of them are taken by ourselves and the others are collected from 31 publicly available Youtube videos taken by the same drone as ours). Table **??** shows average precision of each algorithm on htis dataset. Here the SSD300 model and SSD500 model have the same architecture and the only difference is the input image size(300×300 pixels vs. 500×500 pixels). YOLO and Fast YOLO also use similar architectures except Fast YOLO uses fewer convolutional layers (24 convolutional layers vs. 9 convolutional layers for Fast YOLO).

According to our experiment, Faster R-CNN achieved 83.9% mean average precision (mAP) compared to YOLO models (78.3% and 79.4%) and two SSD models (81.6% and 82.6%).

2.3.2 RECOGNITION SPEED ON CLOUD SYSTEM

Our second set of experiments evaluate the running time performance of the CNN-based object recognition testing the extent to which cloud computing could improve recognition times, and the variability of the unpredictable communication times. We use the same dataset as in the previous section and compare the speed of each algorithm using GPU on a simulated cloud machine. We measure the running time including image loading,

pre-processing, and output parsing (post-processing) time.

Fig. **??** shows the running time of each algorithm as a function of its accuracy. The result shows detection speed and accuracy are inverse related. Fast YOLO showed te highest speed (57.4 FPS) with the lowest accuracy (mAP 78.3%), while Faster R-CNN has the lowest speed (3.48 FPS) with the highest accuracy (mAP 83.9%).

Then we compare Fast YOLO on a local laptop versus Faster R-CNN on the simulated cloud. A comparison of these computing facilities are showed in Table **??**. Fig. **??** shows the average running time of Fast YOLO on local machine is 7.31 seconds per image while for Faster R-CNN running on remote cloud is 1.29 seconds including latencies for sending each image to the cloud (which averaged about 600 ms) and for exchanging detected results and other command messages (which averaged 0.41 ms). Thus the cloud-based recognition performed about 5.7 times faster than the local Fast YOLO on average.

2.3.3 TARGET SEARCH WITH A DRONE

In the test scenario, we use a screwdriver as a target object and scattered various distractor objects on the floor in the indoor test room. The drone started this object searching mission with lower-resolution downward-facing camera, and run the BING algorithm for finding generic objects given the input video. When the drone finds any "interesting" objects on the floor, it switches to the front-facing camera to capture a photo at a higher resolution, then take picture of the candidate area and sends it to the cloud system ($t = 3$s and $t = 8$s). After this, the drone switch the camera back to the downward-facing camera and proceeds to the other candidate positions. The cloud system performs recognition in the meantime. The drone performs the same steps until it finds a target object, at which point the mission is completed ($t = 17$s).

BIBLIOGRAPHY