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 2.1.
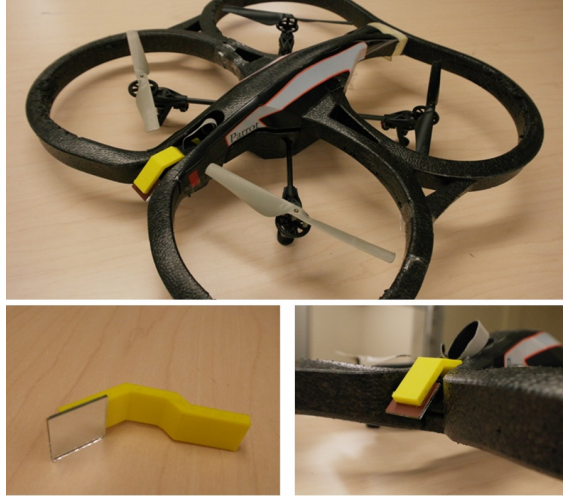
Figure 2.1: We use the Parrot AR.Drone2.0 as our hardware platform (top), adding a mirror to the front-facing camera in order to detect objects on the ground (bottom).

Our approach consists of four main components shown at top 2.2. 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 2.2 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.
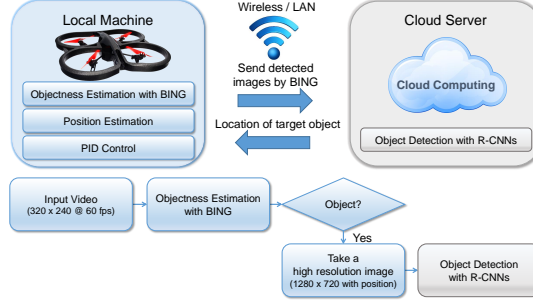
Figure 2.2: System Overview: Our approach consists of four main components: BING based objectness estimation, a position estimation for localization, PID control for navigation, and R-CNNs based object detection. All components are implemented under the ROS framework, so each component can communicate with every other via the ROS network protocol (top). Given input video, the local machine detects generic objects in every frame with BING, then takes a high resolution image and sends it to the cloud server if the frame contains generic objects. The cloud server then runs R-CNNs based object detection to find a target object (bottom).

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%).

Table 2.1: Object detection results on our aerial images collected by the drone.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Fast YOLO	**87.5**	84.6	0.0	50.0	65.5	**100.0**	87.9	80.0	**92.3**	47.1	60.0	75.0	88.9	**100.0**	76.6	**100.0**	54.5	**100.0**	66.7	84.6	78.3
YOLO	60.9	88.2	80.0	80.0	92.3	**100.0**	87.2	**100.0**	70.4	50.0	40.0	**81.0**	77.8	93.4	88.7	**100.0**	45.2	**100.0**	81.8	79.2	79.4
SSD300	60.0	**94.1**	20.0	**100.0**	90.0	**100.0**	**100.0**	**100.0**	75.0	47.8	50.0	66.7	81.3	92.9	**92.9**	**100.0**	66.7	**100.0**	85.7	82.6	81.6
SSD500	66.7	88.2	50.0	88.9	**100.0**	92.9	93.2	**100.0**	72.1	**65**	**85.7**	69.6	88.9	93.8	81.7	**100.0**	66.7	89.5	66.7	**100.0**	82.6
Faster R-CNN	70.6	93.8	**83.3**	85.7	91.9	92.9	89.7	**100.0**	87.2	62.5	-	77.3	**100.0**	93.8	81.7	66.7	**72.7**	**100.0**	**100.0**	62.5	**83.9**

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.
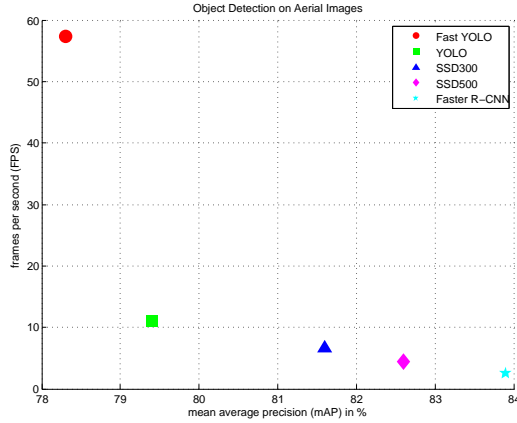
Figure 2.3: A running time comparison of recent state-of-the-art object detectors on our aerial images.

Fig. 2.3 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. 2.4 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
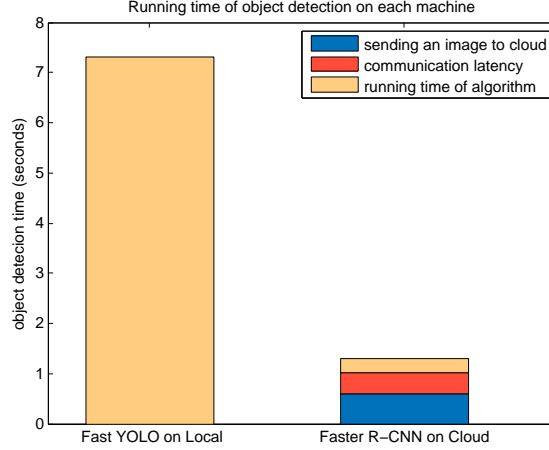
9

Figure 2.4: Running time of object detection on each machine.

finding generic objects given the input video. Fig. 2.5 shows a sequence of images taken during this test. 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).

Table 2.2: Hardware Comparison between Local and Cloud Machine

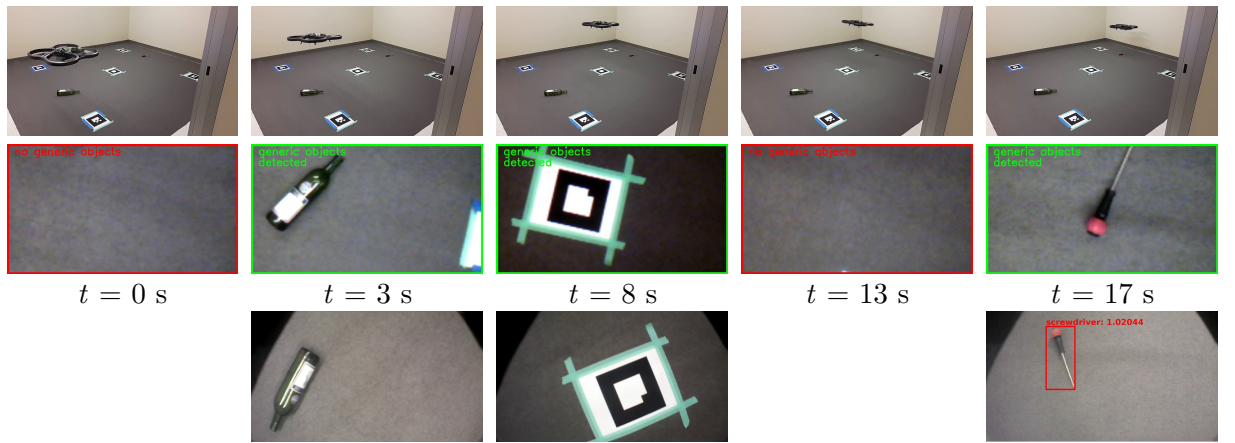|	local computer	cloud computer
CPUs	one Intel Core i7-4700HQ @ 2.4GHz	two Intel Xeon E5-2680 v3 @ 2.5GHz
GPUs	one Nvidia GeForce GTX 770M	two Nvidia Tesla K40
RAM	16 GB	128 GB

10

Figure 2.5: Target Search with a Drone: First rows show movements of the drone during the experiment, and second and third rows indicate detection results from BING and R-CNNs respectively. At $t = 0$ s the drone started to search for a target object and did not find generic objects with BING. At $t = 3$ s, $t = 8$ s, the drone found generic objects with BING, thus took high resolution pictures and sent them to cloud server. However, R-CNNs did not detect a target object in those images. At $t = 17$ s, the drone found generic objects again, thus it took the high resolution picture and sent it to cloud server. Then, finally R-CNNs based object detector found a target object.

CHAPTER 3

ASSESSING COLOR PREFERENCES AND SEX DIFFERENCES USING MASSIVE ONLINE DATA

3.1 INTRODUCTION

Variation in human color preferences across cultures and sexes has not been well characterized. The majority of past studies have worked with relatively small samples of participants asked to make explicit decisions with limited stimuli consisting of blocks of single colors. Using this method, Hurlbert and Ling [?] found a universal preference for bluish hues, along with a greater cross-cultural female preference for reddish hues. They gave an evolutionary explanation related to the adaptive benefit to foraging females of seeking out a red object on a green background (e.g., fruit against leaves). Palmer and Schloss [?] argued that color preferences can be explained in terms of object preferences and associations between objects and colors, which can be influenced by evolution, culture, and individual experience. The question of sex differences was left open in this paper.

We propose that large-scale novel data-mining of images available on the Web can over-

come the limitation of study sizes. Moreover, the metadata associated with the web-scale images provides more interesting information for our complimentary studies. When people upload their photos to social networks, they are sharing information about which scenes and types of photos they prefer. By analyzing the color spectra over 15 million photographs on Flickr, an online photo-sharing network, we measure male and female preferences in an implicit (behavior-based) rather than explicit (ratings-based) manner and on a much larger scale than can be done in a lab experiment. 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. To take a step towards exploring the reason for this difference and the relation of content preference and color preference, we analyze the least entropy textual tags used by photographers taking photos with most amount of color pixels in each hue interval. 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 study the color distribution and observe the overall preference of saturated color and reddish color of human compared to the environment.

3.2   DIFFERENCES IN COLOR DISTRIBUTIONS ON FLICKR

To characterize each photo's color distribution, we consider color histograms in CIE L*C*h* (LCh) color space representing each pixel in a perceptually-uniform space according to Lightness, Chroma, and Hue dimensions. Here we focus only on the Hue dimension, dividing it evenly into 65 discrete bins. Each photo is first converted into a histogram of the distribution of pixels over hue angle. We then normalize the histogram by user, and take the histograms of all users from a particular population (e.g., men) and combine them into a single aggregated histogram.

We first analyze a general dataset with a larger number of images and without any constraint on content, photographer or location where the photos are taken. We collected data for about 150 million publicly-available images from Flickr using the public API, as well as public profile information (including gender) for each photographer. We then downloaded a random subset of 15 million photographs taken by 2.8 million photographers. Preliminary results indicate strong overall sex differences for the predominant reddish and bluish hues, with women uploading more photos with more reddish pixels and men uploading more photos with more bluish pixels (see Fig 3.1).

To allow more careful examination of how preferences vary across content and geo-locations, we also downloaded: (1) photos of people uploaded to the Flickr Allpeople group [?] with at least one human figure on it from 22141 female and 39025 male photographers, and (2) photos geo-tagged within three specific places: Death Valley National Park (with 376 female and 2435 male users), San Diego Balboa Park (with 765 female and 2646 male users), and Disneyland (with 1184 female and 3279 male users). In Fig 3.4, we can find the results from all of the four datasets generally support the preliminary results. That is, the difference in reddish and bluish color preference between genders is consistent across content and geo-locations.

In addition to directly analyzing the hue distribution of pixels, another experiment presenting in Fig 3.6 also supports our results by examining the distribution of users with "preference" of colors along the hue axis. According to Random dataset, we present the number of users with respect to genders sharing photos with more than 10% pixels in each of the 65 color bins showing the consistent result of gender preference on reddish and bluish color.
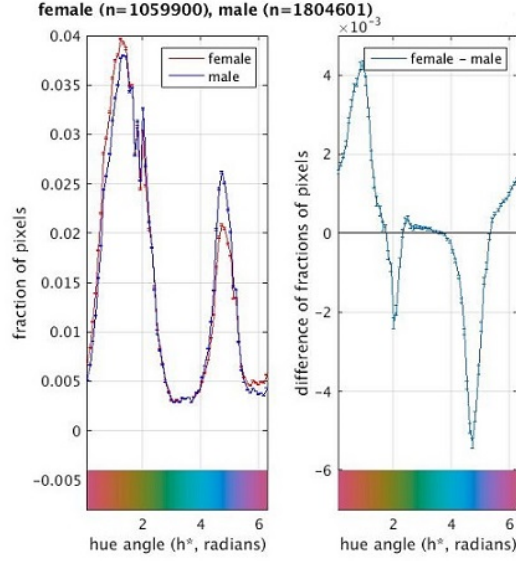
Figure 3.1: Results on Random dataset. Left: Distribution across hue for male and female photographers. Right: Difference between female and male hue distributions, indicating more reddish hues for women and more bluish hues for men.
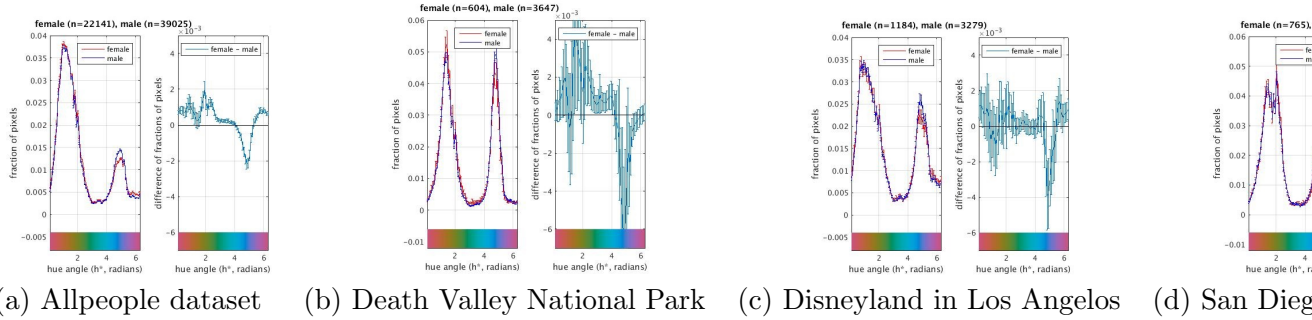

(a) Allpeople dataset   (b) Death Valley National Park   (c) Disneyland in Los Angelos   (d) San Dieg

Figure 3.2: Hue distributions and hue difference distributions of 4 different dataset with constraint on content in (a) and locations in (b)-(d).

3.3 GENDERS PREFERENCE ON CONTENTS

To give a general idea of what content people take pictures of when they prefer certain colors, we perform a further experiment. Generally people would like to tag their images with the occasion, names of people in the picture and their most interested objects. Therefore, we consider the image tags to represent the content of images.

We rank all the tags appearing in 15 million images in the Random dataset, normalized by users, to find the 500 most frequently used tags. To better visualize the tag distribution,
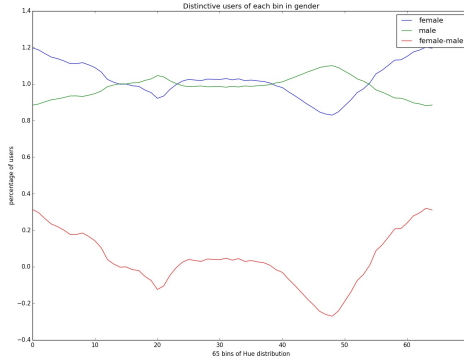
Figure 3.3: Number of users with more than 10% color pixels in each color bin (according to Allpeople dataset) showing the consistaent result of gender preference on redish and bluish color.

we adopt the hue space of Palmer 32 colors [?]. These are 32 evenly-speed color points in LCh space, including eight hues in each of two lightness and two chroma values (giving Light, Muted, Dark, and Saturated variants of each hue).

We find images containing more than 10% pixels of each of the 8 hues. In these images, we aggregate them to users. So each tag used multiple times in images from one user, is only counted once. In this way, we compute the tag frequency and entropy across 8 hue centers, for all users, and for male and female users respectively. We compose the tag distribution maps in two ways. Firstly, for each hue, find the top 10 most frequently used tags and sort them by the least entropy, presented in Fig. ?? with female user tags on top, male users in the middle and all users at the bottom. In the second way, the top 500 tags are sorted by entropy value first. Along this list, we find 10 most frequently used tags for each hue. This plot is also presented in Fig. ?? in the same way as the first plot.
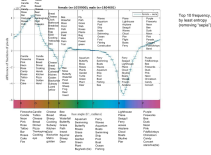
(a) Frequency tag distribution sorted by entropy ascendingly.
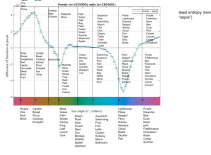
(b) Frequency tag distribution sorted by overall tag entropy.

Figure 3.4: Most concentrated tags on each Palmer color hue for female and male users.

3.4 COLOR DISTRIBUTION OF FLICKR VS. STREET VIEW

The next research question we address is the difference between Flickr photos and the environment where the photographers could have taken pictures. We plot the color distribution in histogram of Palmer 32 colors.

To estimate color distributions in the environment, we collect data from Google Street View. For each geotagged Flickr photo, we query the Google Street View Image API with its GPS information, and download Street View images in 4 view angles (0, 90, 180, 270 degree). If there are no Street View images available for a GPS location, we discard the corresponding Flickr photo.

The result turns out to be very interesting in respect to locations. In Fig. 3.5, we present the results of three locations of Death Valley National Park, Disneyland at Los Angeles and San Diego Balboa Park. In general, we observe the preference of saturated color and reddish color of human compared to the environment.
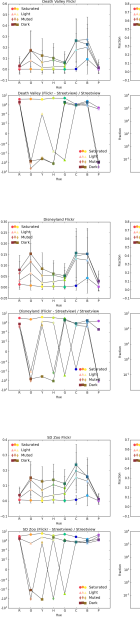
(a) Color of Flickr and Street View in Deathvalley National Park. It shows photographers
(b) Color of Flickr and Street View in DisneyLand at Los Angeles. There is clearly more warm color and
(c) Color of Flickr and Street View in San Diego Balboa Park. In this very

Figure 3.5: Comparison of hue distribution in Palmer's 32 colors for flickr photos and the
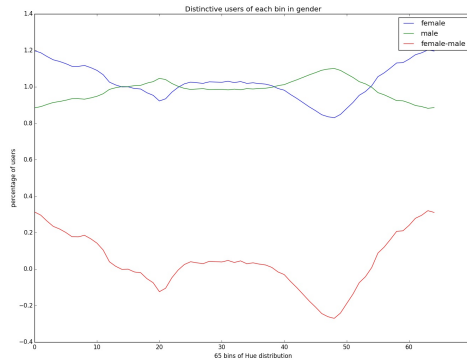Google Street View images at the same GPS lcoation.

Figure 3.6: Number of users with more than 10% color pixels in each color bin (according to
Allpeople dataset) showing the consistaent result of gender preference on redish and bluish
color.

BIBLIOGRAPHY

[1] https://www.flickr.com/groups/allpeople/.

[2] Aesthetic preference for color combinations. http://socrates.berkeley.edu/ plab/pdf/SchlossPalmer.pdf

[3] Malavika Bhaskaranand and Jerry D Gibson. Low-complexity video encoding for uav reconnaissance and surveillance. In *Military Communications Conference (MILCOM)*, pages 1633–1638. IEEE, 2011.

[4] Pierre-Jean Bristeau, François Callou, David Vissiere, and Nicolas Petit. The navigation and control technology inside the ar. drone micro uav. *IFAC Proceedings Volumes*, 44(1):1477–1484, 2011.

[5] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3286–3293, 2014.

[6] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convo-

lutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[9] Ken Goldberg and Ben Kehoe. Cloud robotics and automation: A survey of related work. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2013-5*, 2013.

[10] Anya C Hurlbert and Yazhu Ling. Biological components of sex differences in color preference. *Current Biology*, 17(16):R623–R625, 2007.

[11] Ben Kehoe, Sachin Patil, Pieter Abbeel, and Ken Goldberg. A survey of research on cloud robotics and automation. *IEEE Transactions on Automation Science and Engineering*, 12(2):398–409, 2015.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[13] Stephen E Palmer and Karen B Schloss. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882, 2010.

[14] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, 2009.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Inkyu Sa, Stefan Hrabar, and Peter Corke. Outdoor flight testing of a pole inspection uav incorporating high-speed vision. In *Field and Service Robotics*, pages 107–121. Springer, 2015.

[17] Teodor Tomic, Korbinian Schmid, Philipp Lutz, Andreas Domel, Michael Kassecker, Elmar Mair, Iris Lynne Grixa, Felix Ruess, Michael Suppa, and Darius Burschka. Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue. *IEEE robotics & automation magazine*, 19(3):46–56, 2012.