# Automatic Gathering of Labeled Images Using a Web Crawler to Facilitate Automotive Machine Learning

**Chuho Yi, Jungwon Cho**

***Abstract*: Data preparation for machine learning is time consuming. We built an efficient machine-learning system and automated data preparation. Using a Web crawler, multiple keywords were used to download images to facilitate machine learning of automotive applications by local computers. A verification module then filtered the target images. Here, we explored weather while driving using only images containing vehicles. We found it simple to download images using a Web crawler without manually preparing the experimental learning data. We verified the utility of our approach using a computer vision algorithm. It is possible to eliminate most passive machine learning work when creating huge amounts of diverse data. Our method automatically generates a machine learning dataset for exploring weather features by capturing forward images taken from vehicles.***

***Index Terms*: Automatic gathering, Automotive, Labeled-Image, Machine learning, Web crawler**

## I. INTRODUCTION

Deep learning, an automatic process, has developed very rapidly over the past 10 years and is accepted as having great importance. Essential data collection remains labor intensive, as does identification of true and false data when creating training and test sets. Furthermore, if input is erroneous, output quality falls sharply. Here, we simplify the creation of learning data for a machine that explores driving weather. This is important information required by autonomous cars that explore their surroundings and use the data for appropriate operation. Kurihata detected rainy days using a template that detected raindrops falling on the windshield [1]. Gern exploited the optical flows of rainy and snowy environments to judge the current weather, and applied the data to lane recognition [2].

Here, we focus on vehicular applications that detect weather. First, as in all machine learning applications, it is necessary to define the target data. For example, in terms of images taken from cars, we create identifiers of sunny, rainy, and snowy weather. Next, it is necessary to engage in direct image capture and discrimination. During general machine learning, it is expensive in terms of time and human resources to install a camera in a vehicle, process the images, drive the vehicle while simultaneously capturing useful data, and engage in appropriate distinctions. Even in non-automotive fields, extensive preparation is generally required prior to data acquisition and machine learning. If such preparation is poor, the desired results will be inconclusive, and the process must be revisited.

To deal with these issues, we automatically acquired as much data as possible in a laboratory and purified the data with the aid of the Internet. Thus, acquisition of considerable data required minimal time and effort; if a problem arose, the settings could be changed to rapidly improve operation. Our method is shown in Fig. 1.

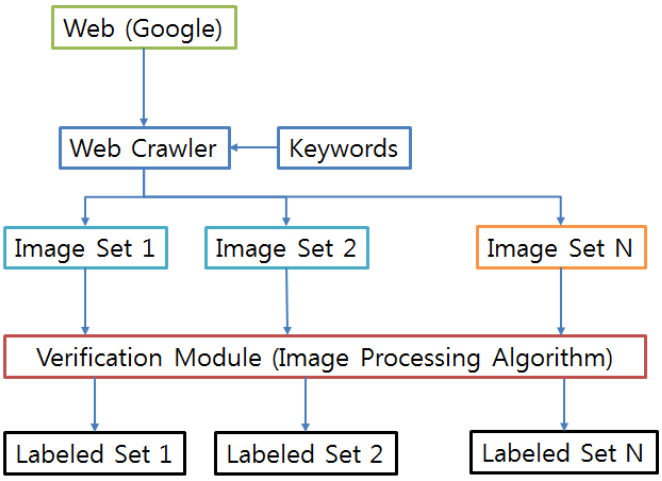



Figure 1. A system for automatically providing Internet datasets for machine learning.

## II. PROPOSED METHOD

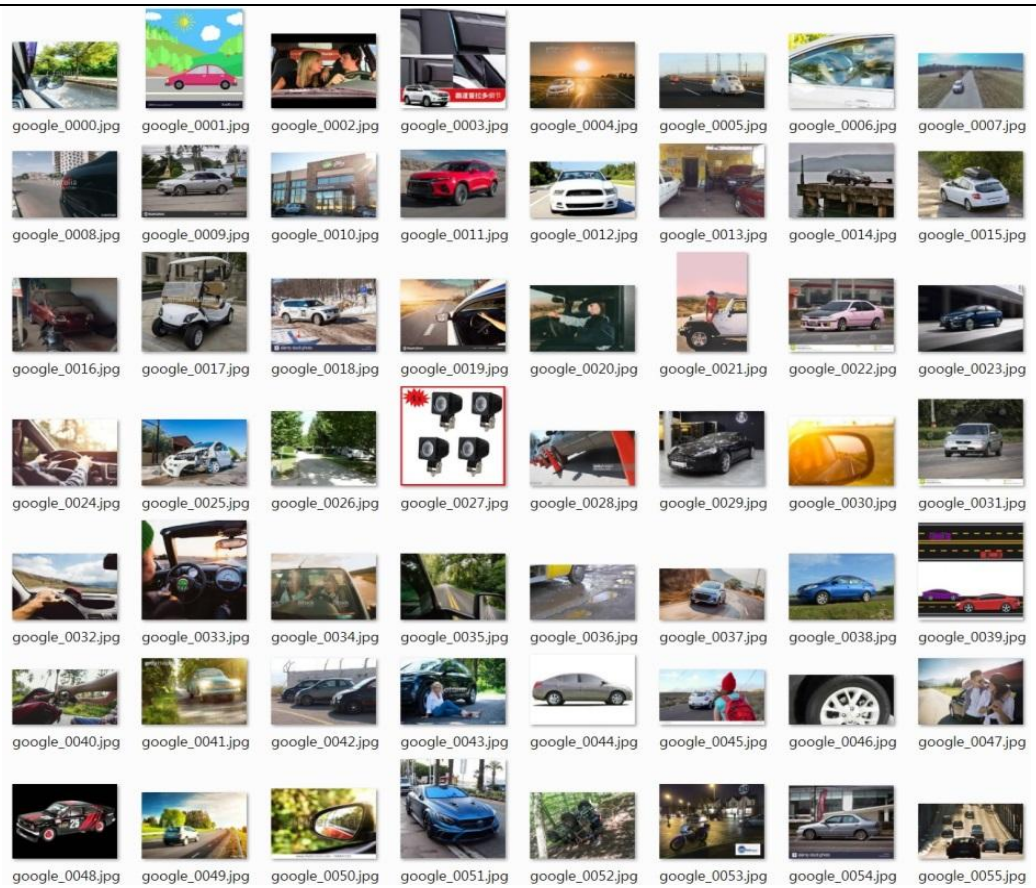
First, we used keywords to search (employing Google) for images on the internet. [3]. When a keyword is entered, an image crawler program automatically downloads the search results to a local computer [4]-[6]. The keywords differed by the target application. We assume that the user will be familiar with machine learning and intuitively use appropriate keywords. We used “automotive sunny road,” “automotive rainy road,” and “automotive snowy road” to target images taken by cars. Thus, various advertisements, images, icons, and products were collected and stored. If such data were not filtered, serious performance degradation would be possible. Only appropriate images must be included. We verified the

**Revised Manuscript Received on May 22, 2019.**

**Chuho Yi**, Dept. of Electronics, Dong Seoul University, Seongnam 13117, South Korea.

**Jungwon Cho**, Dept. of Computer Education, Jeju National University, Jeju 63243, South Korea.

Table Ⅰ. Collection of Google images automatically retrieved using the key phrase “automotive sunny road.”

| | |
|---|---|
| Sunny |  |



data using an existing image processing algorithm. For example, when an image taken from a vehicle was found, that image was required to show the front or the back of a vehicle. To automate this process, we used the Cascade classifier of the Harr feature described during the image-processing OpenCV rally [7], [8]. This recognizes image edges and target corners white and black, respectively, by passing target areas through the Harr kernel. When multiple detectors are sequentially applied, the algorithm is termed a “cascade classifier.” Given the choice between initial application of a simple or complex detector, we reserved the complex detector for candidate images that passed the initial simple detection; this significantly improved detection speed. Here, we apply this approach to vehicle detection.

Only images of interest (those approved by the verification module) were entered into the machine learning dataset. Keyword searching followed by verification facilitates the accumulation of large volumes of image data and can be repeated until the dataset is of adequate size.

## III. RESULTS AND DISCUSSION

We used AutoCrawler to search for and download images [9]. AutoCrawler runs on Python; when a search term is entered, Google automatically creates and saves a local folder. Tables Ⅰ, Ⅱ, and Ⅲ show road images taken on sunny, rainy, and snowy days, respectively. The Tables contain many unhelpful 2D images, pictures of car products, and images that are too zoomed in to be useful. Table Ⅰ features 693 images, Table shows 2,681, and Table displays 3,712.

As downloaded data often cannot be used for the machine learning, we used a verification module to filter out all inappropriate images. We confirmed that the vehicle size was appropriate in all images employing the Harr cascade classifier of OpenCV, which is compatible with Python [10]. Table Ⅳ below shows the appropriately sized vehicles in all image sets.

As shown in Tables Ⅳ and Ⅴ, our method efficiently and automatically stored image data from unmanned vehicles, facilitating rapid dataset construction.



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table II . Collection of Google images automatically retrieved using the key phrase “automotive rainy road.”

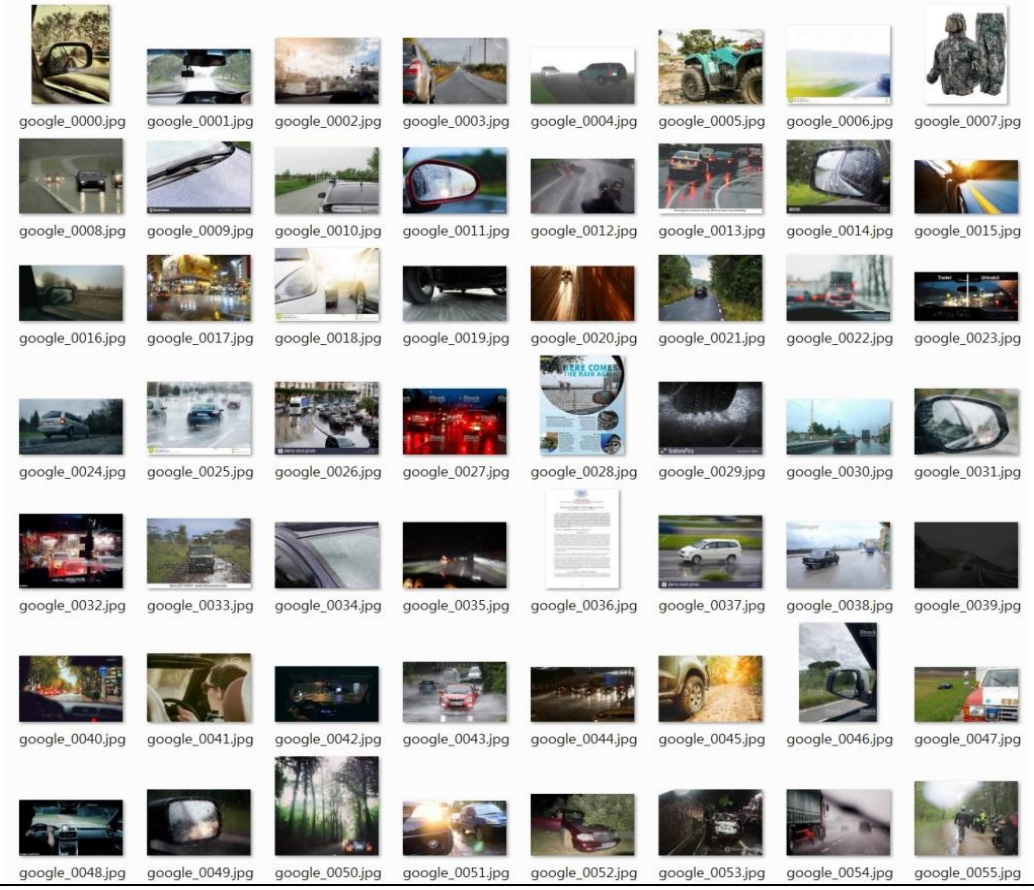
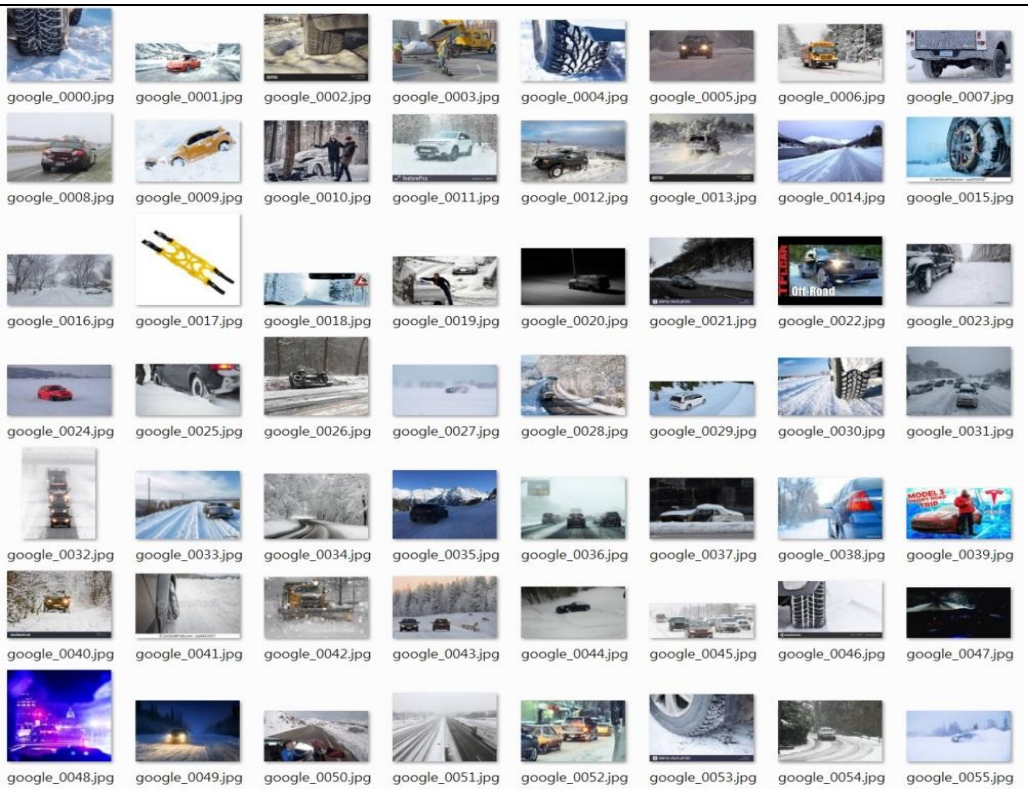
| | |
|---|---|
| Rainy |  |



Table III . Collection of Google images automatically retrieved using the key phrase “automotive snowy road.”

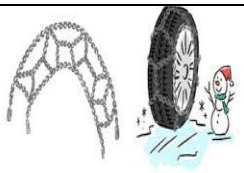
| | |
|---|---|
| Snowy |  |

promotion program funded by Jeju National University

Table Ⅳ. The results afforded by a verification module detecting vehicles using the Harr cascade classifier. The rows show images taken on sunny, rainy, and snowy days.

| | | | | |
|---|---|---|---|---|
| Sunny |  |  |  |  |
| Rainy |  |  |  |  |
| Snowy |  |  |  |  |

Table Ⅴ . Images excluded by the verification module. The rows show images taken on sunny, rainy, and snowy days.

| | | | | |
|---|---|---|---|---|
| Sunny |  |  |  |  |
| Rainy |  |  |  |  |
| Snowy |  |  |  |  |

## IV. CONCLUSION

We created a large, automatically generated, image training dataset. First, we searched the Internet to find appropriate images based on keywords and stored these locally. We then developed a method that automatically selected only appropriate images by reference to prior user-specified verification features. The method indeed downloaded appropriate images, as confirmed by the verification module. In the future, we will configure the training image sets to deal with more complex and diverse situations, using deep-learning classifiers to increase the accuracy of the verification module.

## ACKNOWLEDGMENT

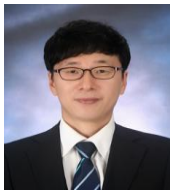
This research was supported by the 2019 scientific

## REFERENCES

1. H. Kurihata, T. Takahashi, I. Ide, Y. Mekada, H. Murase, Y. Tamatsu, and T. Miyahara, “Rainy weather recognition from in-vehicle camera images for driver assistance,” In IEEE Proceedings. Intelligent Vehicles Symposium, 2005. pp. 205-210.
2. A. Gern, R. Moebus, and U. Franke, “Vision-based lane recognition under adverse weather conditions using optical flow,” IEEE of Intelligent Vehicle Symposium, Vol. 2, 2002, pp. 652-657.
3. Google, Available: https://www.google.com/
4. C. Frankel, M. J. Swain, and V. Athitsos, Webseer: An image search engine for the world wide web, 1996.
5. C. Castillo, “Effective web crawling,” In Acm sigir forum, Vol. 39, No. 1, 2005, pp. 55-56.
6. D. Khurana and S. Kumar, “An Improved Approach for Caption Based Image Web Crawler,” International Journal of Computer Science & Management Studies, 12(02), 2012, pp. 2231-5268.
7. G. Bradski and A. Kaehler, Learning OpenCV: Computer

vision with the OpenCV library, O'Reilly Media, Inc., 2008.
8. M. Oliveira and V. Santos, Automatic detection of cars in real roads using haar-like features, Department of Mechanical Engineering, University of Aveiro, 2008.
9. AutoCrawler, Available: https://github.com/YoongiKim/AutoCrawler
10. Vehicle Detection with Haar Cascades Available: https://github.com/andrewssobral/vehicle_detection_haarcascades

## AUTHORS PROFILE



**Chuho Yi** is an Assistance Professor at the Department of Electronics in the Dong Seoul University. He has worked for 4 years as a chief researcher in Future IT R&D Laboratory of LG electronics. He is studying about Auto-driving, Intelligent robot, Artificial Intelligence and so on.

**Jungwon Cho** is a Professor at the Department of Computer Education, Jeju National University. He is the director of 'Research Institute of Education Science' and 'Center of Intelligent Computing Education' in Jeju National University. He is studying about Computing education, Intelligent information ethics, Intelligent information system and so on.