Since 2007, solar photovoltaic (PV) residential and commercial installations in the United States have increased by over 1300%, and solar energy has become a significant portion of the overall U.S. energy system [1]. Currently, the solar industry lacks information about energy capacity at a granular spatial scale. Solar producers, urban planners, energy policymakers, and the research community require a ground-truthed, publicly available, nationwide, and granular PV installation database for improved decision-making and energy system design.

To aid the development of such a database, we created a data set of over 13,000 rooftop solar panel array images using high-resolution orthoimagery. The unique, groundbreaking data set will serve as a ground-truth training model for future machine learning algorithms that can automatically identify rooftop solar PV. Some preliminary algorithm development for identifying solar panel regions with a small training set has yielded encouraging results. With a highly accurate algorithm based on our large data set, a database of rooftop solar PV can be created by an automated process for the entire U.S. and beyond.

Abstract

Solar Power Estimation Through Remote Sensing

Arjun Devarajan (BS’18), Sharrin Manor (BSE’16), Jeffrey Perkins (NCSU BSE’18), Raghav Saboo (G’16), Wuming Zhang (BSE’18)

Introduction

Currently, information on solar capacity, locations of installations, and energy generated is gathered by groups like the Energy Information Administration (EIA) via a variety of methods – self-reported surveys, tax rebate applications, reports from utility companies, etc. Despite these efforts, the information that exists is incomplete at a disaggregated level for the nation as a whole. It is difficult to find up-to-date data with granularity finer than the county or utility level. The California Solar Initiative (CSI), for instance, is accompanied by a public record that lists all the applications received for California’s tax rebate program with granularity at the zip code level. This can be used to identify which general areas in California have higher solar installation densities than others (Figure 1). However, this database is limited because not everyone who installed solar in California applied for a tax rebate or even made their installation during the years completed by the CSI.

A machine learning solution that analyzes orthoimagery (satellite imagery or aerial photography taken orthogonal to the surface of the earth) to identify rooftops with solar will allow researchers to accurately and precisely map solar energy generation in the United States. The true solar capacity of any location in the U.S. and the exact distribution of the energy within the grid system, aid energy planners in decision-making as solar energy continues to experience more installations than others. The potential valuable analyses are numerous.

The result of this ground-truthing process is an immense data set. Figure 3 shows an example visualization of all the solar arrays marked as points in a section of the city of Fresno. These thousands of data points are in a table format where valuable information can be extracted from each detected region. Through this step, we are able to obtain a set of separate regions for removing these regions, a Mean Shift algorithm was implemented utilizing the centroid of each detected region.

The prescreener has three main steps: (1) run the Maximaly Stable Extremal Regions algorithm to extract continuous “blobs” of similar pixels, (2) for each of these regions, compute a likelihood ratio based on two color models, and (3) keep the top 8-10 regions by likelihood ratio.

An example of the result of the prescreener is shown in Figure 4 above. However, it was noticed that the prescreener retains duplicate or overlapping regions. In order to remove these regions, a Mean Shift algorithm was implemented utilizing the centroid of each detected region. Through this step, we are able to obtain a set of separate regions for each image without any duplicates. After this step, with each potential region containing the identified solar panels, the algorithm looks to classify the regions based on three sets of features extracted from each of the regions:

1. A shape based feature: perimeter/area ratio
2. Coloration based features: mean pixel intensity within each channel (RGB)
3. A prescreener feature: Gabor filters with 8 different scales and 8 different orientations [3]

With these features implemented along with a simple support vector machine classifier, we ran a 100 k-fold cross-validation to obtain the receiver operating characteristic (ROC) curve, as pictured in Figure 5. The test data included 53 regions of solar panels and 266 regions without solar panels and the algorithm was able to obtain a False-Positive Rate of 7.5% and True-Positive Rate of 94.34%.

These results are promising, and with the basic algorithm now in place, future work will revolve around improving performance by fine-tuning the prescreener, improving the features used for classification, and utilizing the larger ground-truth data set to train the classifier.

Acknowledgements

Dr. Michael Gustafson, Maggie Booz, Adam Caves, David Clifton, Maydra Devarajan, Rose Nowell, Natalia Odineltkova, Matthew Seong, and Joseph Stalin

References

