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)
Kyle Bradbury (Energy Initiative), Leslie Collins (Pratt), Timothy Johnson (Nicholas), Jordan Malof (Pratt), Richard Newell (Nicholas, Energy Initiative)

Abstract

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 PV panel arrays 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.

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 of 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 capacity can also be measured. This information will help system operators better manage the grid system, aid energy planners in decision-making as solar energy continues to expand, and help policymakers understand why some specific areas in a region experience more installations than others. The potential valuable analyses are numerous.

To tackle this task, the project team developed a graphical user interface (GUI) which eased the process of manually creating an extensive database of training data. Several cities in California were chosen based on availability of high resolution orthoimagery (less than 0.5 m^2), recency of imagery (imagery from within the past three years), and density of solar installations (higher density was more desirable and CSI data were available to determine density estimates). The ground-truthed training data can now be fed into a solar panel identification algorithm, an initial version of which was developed separately using a smaller data set.

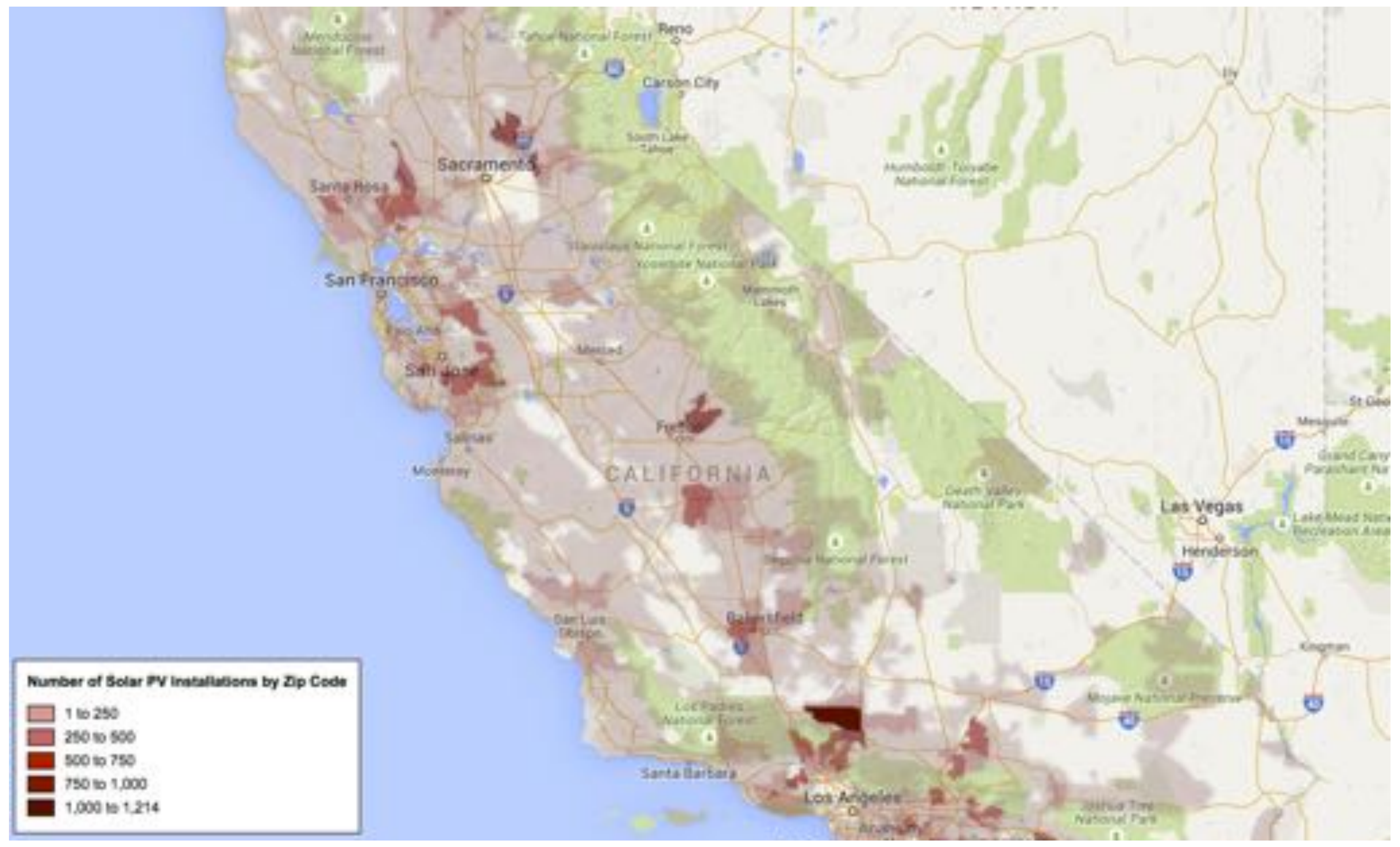

Figure 1: Choropleth map of solar panel installation density by zip code in California using data from the California Solar Initiative and created with Google Fusion Tables

Creation of Data Set

The data set constructed by our team contains details about more than 13,000 rooftop solar panel arrays found in the California cities of Fresno, Stockton, Modesto, and Oxnard. These cities have recent, high-resolution orthoimagery available to download from the U.S. Geological Survey (USGS) website [2]. We built a graphical user interface from scratch to allow us to work through hundreds of large images (typically showing 1.5 km^2 each) and to precisely mark the areas containing rooftop solar arrays by drawing polygons around these regions. Figure 2 demonstrates our GUI in use.

Figure 2: Snapshot of GUI developed for ground-truthing location of solar arrays

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 about each continuous solar array polygon is contained: the latitude and longitude of the centroid and vertices of the polygon, the area of the polygon, the pixel positions of the centroid and vertices relative to the USGS image containing the polygon, the filename of that image, and the city in which the polygon is located.

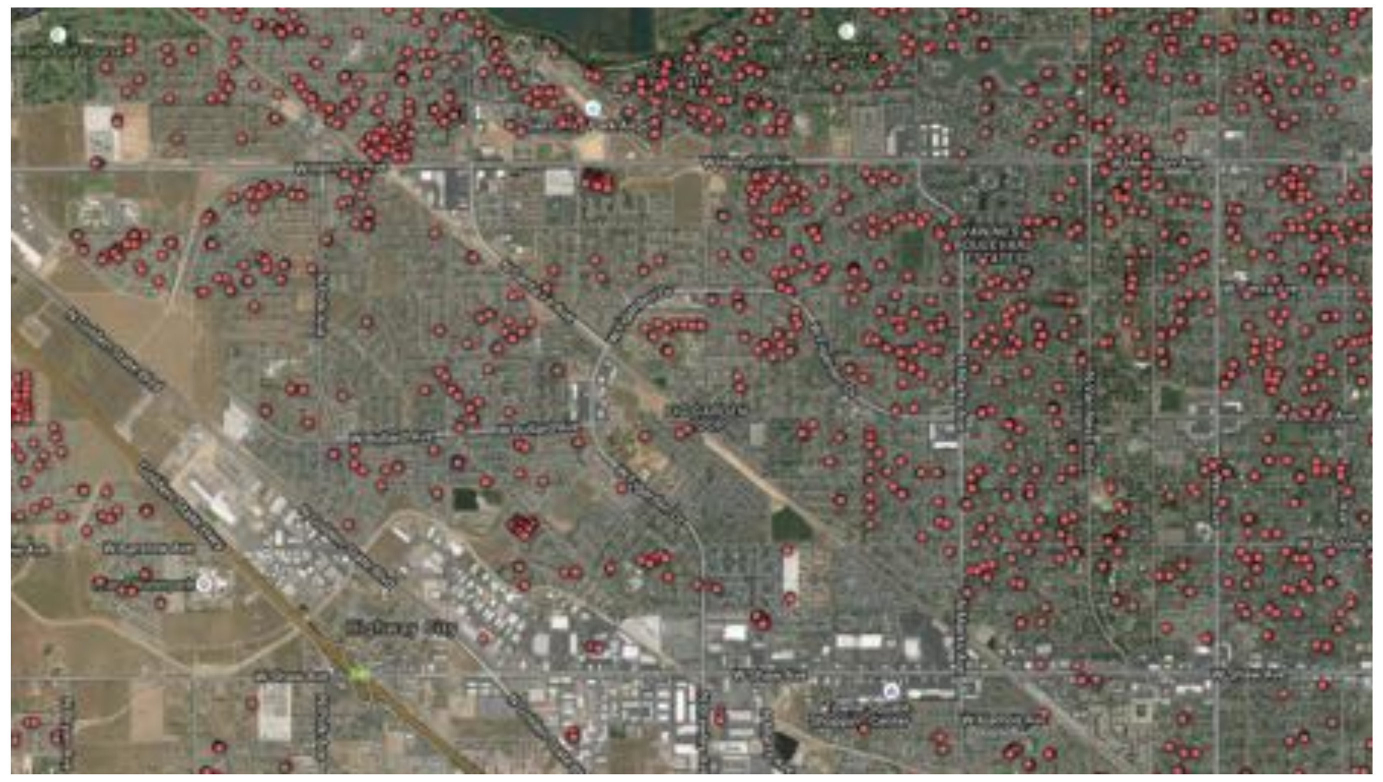

Figure 3: Close-up of an area of Fresno, California with coordinates of centroids of ground-truthed solar arrays plotted as red points created with Google Fusion Tables

The cities used have varied geographies and building styles, but further ground-truthing in different states would likely form a somewhat stronger training set. However, as it stands, our current data set can provide valuable insight to researchers, and we are confident it can be used to further develop a strong, scalable solar PV identification algorithm.

Automation of Solar Photovoltaic Detection

In addition to the creation of the data set, significant progress has also been made towards developing the machine learning algorithm to detect solar photovoltaic arrays from satellite imagery. As this work was conducted in parallel, a smaller set of 100 ground-truthed satellite images was used to create and test the algorithm with the vision of ultimately training and implementing the algorithm on the larger data set discussed previously. The algorithm itself can be broadly defined as two main steps: (1) prescreening each image to select regions of high interest, and (2) classification of each region based on its unique features.

The prescreening step is an essential component of the algorithm as it helps narrow down regions investigated as potential solar arrays within large high resolution satellite orthoimages. This has the advantage of reducing the processing and classification time significantly and of allowing the algorithm to scale up easily to work on a larger data set. The prescreener has three main steps: (1) run the Maximally 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.

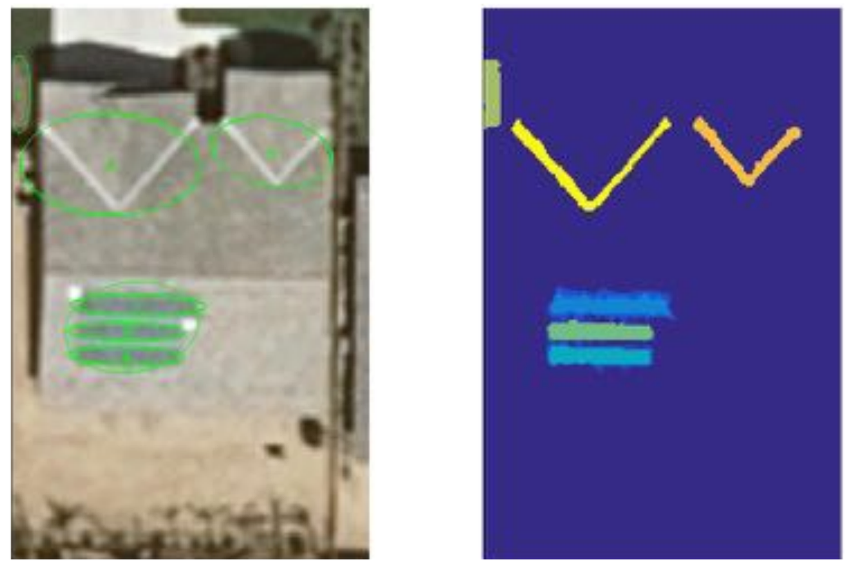

Figure 4: Initial regions of interest selected by the prescreener. In the right image, all shapes not colored dark blue are initial regions

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 (RBG)
3. Texture based features: Gabor filters with 5 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.

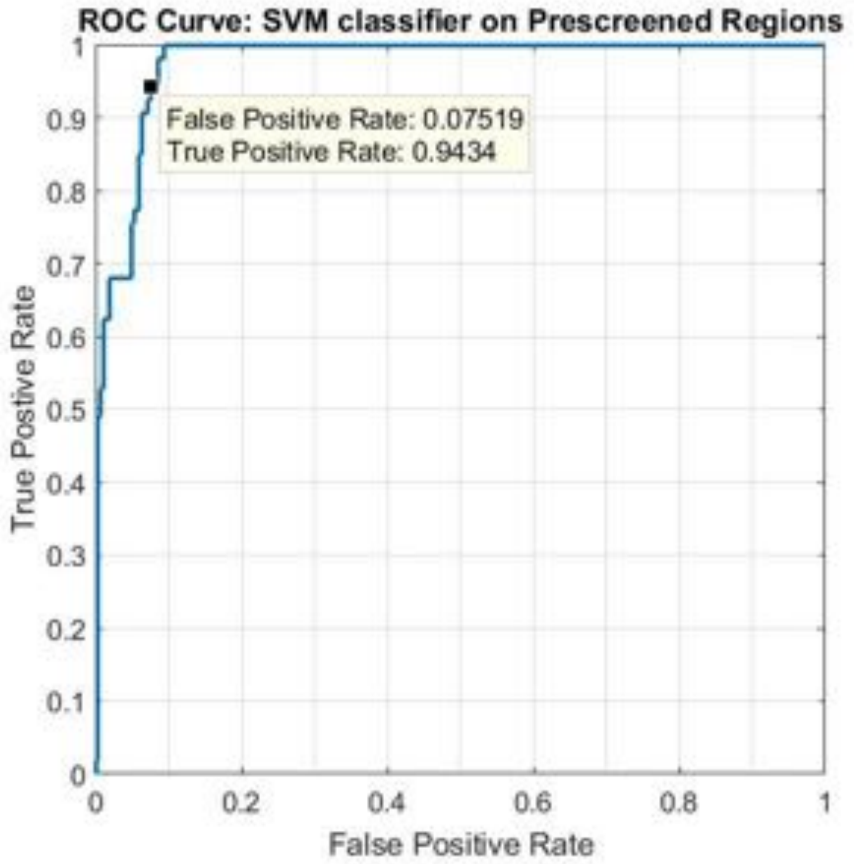

Figure 5: ROC curve showing the trade-off of true positives and false positives in identifying solar panel regions in the test data set.

Acknowledgements
Dr. Michael Gustafson, Maggie Booz, Adam Caves, David Clifton, Maydha Devarajan, Rose Newell, Natalia Odnoletkova, Matthew Seong, and Joseph Stalin

References
[1] "Photovoltaic (Solar Electric)." SEIA.org. Solar Energy Industries Association, 2014. Web. 17 July 2015.
[2] "EarthExplorer." USGS.gov. U.S. Department of the Interior, 7 July 2015. Web. 19 July 2015.
[3] Vijayaraj, Veeraraghavan, Eddie A. Bright, and Budhendra L. Bhaduri. "High resolution urban feature extraction for global population mapping using high performance computing." Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International. IEEE, 2007.