Creating Cloud Segmentation Data Set Using Sky Images of Afyonkarahisar Region

Ardan Hüseyin Eşlik[1,*], *Emre* Akarslan[1], and *Fatih Onur* Hocaoğlu[1]

[1] Solar and Wind Application and Research Center Afyon Kocatepe University Afyonkarahisar, Turkey

Abstract. The use of sky images in solar radiation intensity estimation has been one of the most studied topics in the literature since it improves the estimation results. The first step in processing sky images with image processing methods is to separate the pixels in the images as clouds or sky. This process is known as cloud segmentation in the literature. In this study, the sky is photographed using the sky imaging system installed at Afyon Kocatepe University Solar and Wind Energy Application and Research Center at times with different clouding characteristics and cloudiness rates in Afyonkarahisar Region. The photographs are divided into 25 parts, and small sky patterns are obtained. The pixels in the obtained sky patterns are manually segmented, and a cloud segmentation dataset is created for future studies. Since the resulting dataset contains high-resolution images and pre-labeled data, it can be used to obtain more accurate results for the segmentation process and allows learning algorithms to learn faster. The dataset can be used by researchers in studies such as solar energy forecasting, meteorology, and weather forecasting, and the dataset in this paper will be shared with researchers upon request.

1 Introduction

In image processing, segmentation involves dividing a picture or image into specific parts and identifying these parts. This can be used to analyze the image's content and detect specific objects or backgrounds. Cloud segmentation is important in applications such as solar energy and weather forecasting. This process involves detecting clouds that reflect or block sunlight. This helps to obtain more accurate sunlight data needed for solar energy forecasting. Cloud segmentation is usually performed using image processing techniques or learning algorithms. These techniques include basic image processing techniques, classification algorithms, or learning-based techniques. For example, image processing techniques may include histogram equalization, filtering, or morphological operations. Classification algorithms usually include algorithms such as support vector machines, k-nearest neighbors, or neural networks. Among learning-based techniques, deep learning techniques are often used.

Cloud segmentation is performed using image processing or learning algorithms, and cameras, radars, and satellites usually collect data. This data is used to detect clouds that

* Corresponding author: ardanhuseyineslik@gmail.com

reflect or block sunlight. This process helps obtain more accurate sunlight data for solar energy forecasting.

In image processing, masks are a tool used for segmentation. A mask is a mathematical function determining whether a particular image region should be selected. Masks usually contain binary values, i.e., "1" for the selected region or "0" for the unselected region. Within an image, the selected region is marked by the mask with a 1, while other regions are marked with a 0. Datasets with masks form a subset of the datasets used for segmentation. These datasets may contain data that has already been labeled or classified, so they can be given to learning algorithms to be used for segmentation. Data sets with masks allow learning algorithms to obtain more accurate results and learn faster. They also allow learning algorithms to perform better because they contain the data needed for the segmentation process, and this data has been pre-processed. Therefore, algorithms spend less time and get more accurate results. Using data sets with masks in applications such as solar energy prediction and satellite image analysis is especially important.

Many studies on cloud segmentation have been carried out, and different datasets have been used. Drönner et al. proposed a method for cloud segmentation based on Convolutional Neural Network (CNN) architecture. The success of the proposed approach is tested using Meteosat Second Generation (MSG) satellite data. As a result of the evaluations, they achieved a 94% success rate [1]. Soumyabrata et al. proposed a new method for segmenting sky/cloud images from sky imagers based on a systematic analysis of different color spaces and their components using partial least squares regression. Their study used HYTA (Hybrid Thresholding Algorithm) and SWIMSEG datasets. They also created the Singapore sky imaging segmentation dataset from their collected sky images. The proposed method has shown better results than other methods in the literature according to many error evaluation criteria. [2]. Soumyabrata et al. presented a systematic approach for selecting color spaces and components to segment sky/cloud images. Using mainly principal component analysis (PCA) and fuzzy clustering for evaluation, they aimed to identify the most suitable color components for this task. They utilized the HYTA dataset in their study. The evaluation results showed that the proposed approach could be successfully used in cloud segmentation [3]. Wanyi et al. proposed a new method for cloud segmentation utilizing convolutional neural network architecture. The performance of the method is compared with traditional methods in the literature. The results show that the proposed method gives better results than traditional methods and can be successfully used in cloud segmentation. [4]. Hasenbalg et al. compared six different cloud segmentation algorithms in the literature and analyzed their performance. They used 829 manually segmented sky images to compare these algorithms. The results showed that three of the six segmentation methods (CSL, HYTA+, and FCN) achieved overall accuracy above 90% [5]. Ye et al. proposed a supervised approach for cloud detection and recognition in whole-sky images. They utilize advanced image processing techniques and machine learning algorithms to identify and categorize clouds with high precision. For this purpose, they manually labeled and classified sky images obtained from meteorological stations. The research contributes to improved cloud monitoring in remote sensing applications, such as weather forecasting and climate studies [6]. Zhang et al. created CloudNet, a ground-based system for classifying clouds, with a deep convolutional neural network. The authors used this method to accurately analyze cloud properties from ground observations. Furthermore, they established an extensive database of 11 cloud categories, including contrails, and utilized CNNs to classify them. This approach notably enhanced cloud classification accuracy, contributing to meteorological research. [7] Nie et al. introduced SKIPP'D, a dataset containing sky images and photovoltaic power generation data for short-term solar forecasting. They collected and curated this dataset to aid research in solar energy forecasting. SKIPP'D provides valuable resources for improving the accuracy of solar power generation predictions [8]. Terrén-Serrano et al. presented Girasol, a dataset

comprising sky images and global solar irradiance data. They assembled this dataset to support research in solar energy modeling and forecasting. Girasol offers valuable insights into solar energy potential and variability, benefiting the renewable energy sector [9].

Section 2 of this study introduces the system and its components installed in the Afyon Karahisar Region Solar and Wind Energy Application and Research Center (GÜRAM) for imaging the sky. Section 3 explains how the imaging process is performed using the system, and the sky images obtained are presented by giving examples.

2 Materials and methods

2.1 Sky imaging system and components

In order to obtain the data set aimed to be created within the scope of the study, a sky imaging system is first created to collect sky images. The sky imaging system consists of a digital camera, a lens and a panel to protect the system from external factors. Canon EOS 80D model is used as a digital camera. The camera has a Digic 6 image processor, CMOS sensor type, 24 MP effective pixels, and a maximum resolution of 6000x4000. Detailed specifications of the digital camera are given in Table 1.

Table 1. Detailed specifications of the digital camera.

FEATURE	VALUE
Camera Segment	D-SLR
Megapixel	24 megapixels
Sensor	CMOS Sensor with 1.6x Multiplier
Processor	Digic 6
Light Sensitivity (ISO)	100-16.000 (25.600)
Photo Capture Speed	7.0 frames per second
Focus System	Dual Pixel CMOS Focus System
Maximum Curtain Speed	1/8000

In the sky imaging system, a fisheye lens is used in addition to the digital camera to photograph the sky at a wide angle. In this way, it aims to view all the clouds in the sky when the photo is taken. Samyang brand 8mm fisheye lens with a 167-degree viewing angle is preferred as the fisheye lens. Detailed specifications of the fisheye lens used are presented in Table 2.

Table 2. Detailed specifications of the fisheye lens.

FEATURE	Value
Focal length	8mm
Diaphragm	Maximum: f/3.5 Minimum: f/22
Perspective	167 °
Lens Type	Fisheye

FEATURE	Value
Diaphragm	f/3.5
Minimum Focus Distance	12" (30.48 cm)

Finally, the system is placed and fixed in a panel so that it would not be affected by adverse weather conditions such as rain and snow. After the panel is placed on the roof of GÜRAM located on the Afyon Kocatepe University campus, the system is ready for image collection. The sky imaging system is shown in Figure 1.

Fig. 1. Sky imaging system placed on the roof of Afyon Kocatepe University Solar and Wind Application and Research Center (GÜRAM).

2.2 Performing the imaging process and selecting the images to be masked

With the camera's timer in the sky imaging system, sky photographs are obtained at desired time intervals (10sec interval is used in this study). The camera settings are set to ISO100, 1/2500s exposure time, RGB, and 4000x6000 pixel resolution and kept constant throughout the shooting. Images are taken regularly every day between 10:30 and 16:30 and stored in JPEG format.

After acquiring thousands of images captured at 10-second intervals on various days, the process of choosing images for the dataset has commenced. At this stage, images are carefully chosen by considering the distinct cloud attributes aimed to be encompassed in the dataset. Specifically, clear skies, low cloud skies, high cloud skies and overcast skies are distinguished and sky images are selected equally for each sky type. After conducting examinations and evaluations, we identified 33 sky images with distinct forms of cloud coverage, which are separated to form the dataset. The remaining images are not evaluated and will be used for future research studies. Examples of selected images are provided in Figure 2.

The images presented are specially selected, with the top right image corresponding to a mostly clear sky and the bottom right image to a mostly overcast sky. On the other hand, the upper left and lower left images are examples of images with different cloudiness characteristics. All images used in the study are available for sharing with researchers upon request.

Fig. 2. Some examples of the sky collected using the sky imaging system.

3 Data generation

After collecting the sky images and selecting the images for the dataset, the process of marking (labelling) the sections of the images with either clouds or sky is initiated. The objective at this stage is for the cloudy parts of the images to be coloured a specific colour and the sky sections to be labelled with a different colour. Consequently, a labelled dataset for use in cloud/sky segmentation problems can be obtained. In this study, Adobe Photoshop 2020 software is utilized to annotate images. The program's colour and marking-based selection methods are employed to select clouds and sky. Once selected, the cloud pixels are coloured white and the sky pixels are coloured black for labelling purposes. Subsequently, an algorithm written in Python language is used to convert all images into binary format to enable binary labelling, with the assigned values of 0 or 1. In this context, the cloud parts in the sky images recorded with a resolution of 4000x6000 pixels are labeled with the value one, and the sky parts are labeled with the value zero in the binary system. The pixels containing the sun in the images are considered the sky for this study and are labeled with a value of zero. In future studies, it aimed to accept the sun as a separate class and create a third label.

After completing the labeling process for all selected sky images, the images are divided into 25 equal parts. In this way, the number of data is increased for studies that require more data, such as deep learning, and the image size is reduced to reduce the processing load that may occur during the studies. As a result, after the segmentation process, 825 different images with 800x1200 resolution are obtained from 33 sky images with 4000x6000 pixel resolution. In this way, a cloud segmentation dataset containing different sky cloudiness conditions is created. Figure 3 shows some examples from the cloud segmentation dataset.

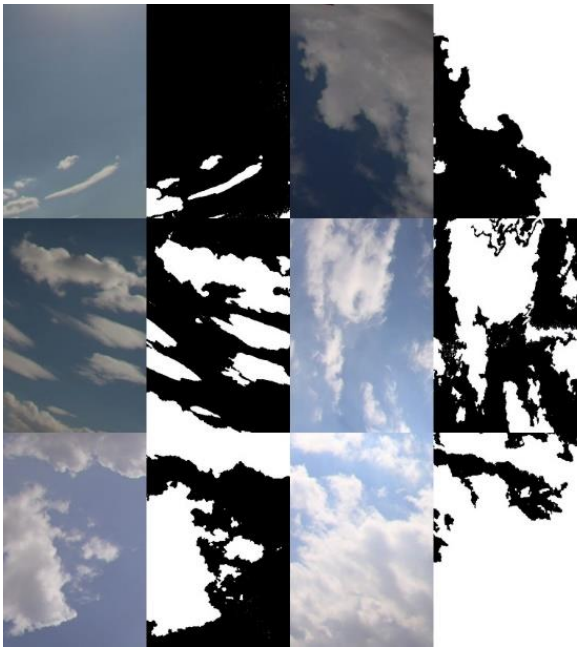

Fig. 3. Some examples from the segmentation dataset created within the scope of the study.

Examining the images of the sky and corresponding masks in Figure 3 reveals that cloud and sky separation is achievable with high accuracy. This is due to the independent handling and manual segmentation of each image. However, although the accuracy is high, the manual segmentation is considered time-consuming and difficult. Furthermore, due to variations in the transparency and colour of clouds, separation of the sky and clouds can be challenging in some areas. For instance, the determination of where highly transparent clouds that may resemble fog start and finish is an intricate issue. Thus, it is unfeasible to establish cloud/sky distinctions that generalise as precise for every individual. Future studies aim to achieve full autonomy of the sky imaging system, classify the sun as a separate entity, increase the amount of data by segmenting additional sky images, and integrate labeling.

4 Results

In this research, we generated a cloud segmentation dataset comprising of 825 sky images and corresponding masks, employing data provided by Afyon Kocatepe University GÜRAM. Firstly, we assembled a sky imaging system composed of a digital camera, lens, and a panel. Next, we analyzed the images captured via the sky imaging system and selected images presenting distinct cloud covers following evaluations. Finally, we manually labelled the selected images into two classes, cloud or sky, and transformed them into binary image format for ease of use with machine learning and deep learning techniques. This dataset will assist researchers in more accurately predicting and analysing the location and shape of clouds. Specifically, it will enable testing of forecasting methods used in studies such as solar energy forecasting, meteorology and weather forecasting. Additionally, the dataset can be utilised for power generation and related industries. The dataset provided in this article will be made available to researchers upon request.

This study is supported by Afyon Kocatepe University Scientific Research Project Coordination Unit with the number 20.FEN.BIL.25.

References

1. J. Drönner, N. Korfhage, S. Egli, M. Mühling, B. Thies, J. Bendix, B. Freisleben and B. Seeger, Fast cloud segmentation using convolutional neural networks. Remote Sensing, 10, 11 (2018), 1782.
2. S. Dev, Y. H. Lee and S. Winkler, Color-based segmentation of sky/cloud images from ground-based cameras. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 1 (2016), 231-242.
3. S. Dev, Y. H. Lee and S. Winkler, Systematic study of color spaces and components for the segmentation of sky/cloud images. IEEE, City, (2014).
4. W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, Z. Wang, Y. Xia, Y. Liu, Y. Wang and C. Zhang, SegCloud: A novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation. Atmospheric Measurement Techniques, 13, 4 (2020), 1953-1961.
5. M. Hasenbalg, P. Kuhn, S. Wilbert, B. Nouri and A. Kazantzidis, Benchmarking of six cloud segmentation algorithms for ground-based all-sky imagers. Solar Energy, 201 (2020), 596-614.
6. L. Ye, Z. Cao, Y. Xiao and Z. Yang, Supervised fine-grained cloud detection and recognition in whole-sky images. IEEE Transactions on Geoscience and Remote Sensing, 57, 10 (2019), 7972-7985.
7. J. Zhang, P. Liu, F. Zhang and Q. Song, CloudNet: Ground-based cloud classification with deep convolutional neural network. Geophysical Research Letters, 45, 16 (2018), 8665-8672.
8. Y. Nie, X. Li, A. Scott, Y. Sun , V. Venugopal and A. Brandt, SKIPP'D: A Sky Images and Photovoltaic Power Generation Dataset for short-term solar forecasting. Solar Energy, 255 (2023), 171-179.
9. G. Terrén-Serrano, A. Bashir, T. Estrada and M. Martínez-Ramón, Girasol, a sky imaging and global solar irradiance dataset. Data in Brief, 35 (2021), 106914.