Computers and Geosciences 109 (2017) 67–74

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Classification of radiolarian images with hand-crafted and deep features

Ali Seydi Keçeli [a,*,1], Aydõn Kaya [a,1], Seda Uzunçimen Keçeli [b,1]

[a] Hacettepe University, Computer Engineering Department, 06800, Ankara, Turkey
[b] Hacettepe University, Geological Engineering Department, 06800, Ankara, Turkey

ARTICLE INFO

Keywords:
Deep learning
Image features
Radiolarian image
Pattern classification

ABSTRACT

Radiolarians are planktonic protozoa and are important biostratigraphic and paleoenvironmental indicators for paleogeographic reconstructions. Radiolarian paleontology still remains as a low cost and the one of the most convenient way to obtain dating of deep ocean sediments. Traditional methods for identifying radiolarians are time-consuming and cannot scale to the granularity or scope necessary for large-scale studies. Automated image classification will allow making these analyses promptly. In this study, a method for automatic radiolarian image classification is proposed on Scanning Electron Microscope (SEM) images of radiolarians to ease species identification of fossilized radiolarians. The proposed method uses both hand-crafted features like invariant moments, wavelet moments, Gabor features, basic morphological features and deep features obtained from a pre-trained Convolutional Neural Network (CNN). Feature selection is applied over deep features to reduce high dimensionality. Classification outcomes are analyzed to compare hand-crafted features, deep features, and their combinations. Results show that the deep features obtained from a pre-trained CNN are more discriminative comparing to hand-crafted ones. Additionally, feature selection utilizes to the computational cost of classification algorithms and have no negative effect on classification accuracy.

1. Introduction

Radiolarians are planktonic protozoa widely distributed in the oceans, throughout the water column adapted for a drifting existence. The skeleton of radiolarians is made up of pure amorphous silica and is the most characteristic morphological feature of the organism. The intricacy and the architectural diversity of the skeletons have long given rise to amazement (De Wever et al., 2002). They are important biostratigraphic and paleoenvironmental indicators for paleogeographic reconstructions. Radiolarians sensitivity makes them useful in the biostratigraphical correlation of oceanic sediments. Radiolarians have an increasing value as depth, paleoclimate, and paleotemperature indicators (Casey et al., 1990). Radiolarian paleontology still remains the cheapest and comparatively rapid way to obtain dating of deep ocean sediments (O'Dogherty et al., 2009a).

Studying fossilized radiolarians became possible after the etching technique was established and described by Dumitrica (1978) and Pessagno Jr and Newport (1972) and validated by the radiolarian paleontologists (De Wever et al., 2002). Processing patterns change with different types, but there are classical preparation procedures. The radiolarian skeletons are extracted from broken rock samples by treating them with a mixture of concentrated acids (10%) (hydrofluoric, hydrochloric, nitric, and acetic) and water (90%). After a lapsed reaction time of 24 h, washing and sieving should be performed with 63 μm and 280 μm meshes and the residue on the 63 μm must be collected. Finally, radiolarian skeletons are collected from the residue with the help of a slender brush under a binocular microscope. Those picked out are examined under Scanning Electron Microscope (SEM) for detailed taxonomic studies and illustration (De Wever et al., 2002). To identify the radiolarian to the corresponding genera and species one would need to exhaust former studies up until today. Not only morphological characters of the skeleton but also geometry and property of the shells including, the number of segments, horns, arms, feet, spicules and pores on the surface of the skeleton and the apertures are important to identify genera and species of the radiolarians. Subsequent to taxonomic studies paleontologists work on the radiolarian fauna they get and decide which time interval the sample fits in geological time scale, then paleontologist can model the biostratigraphy of the study area and can make paleogeographic interpretations. Although species classification method described above is an old-fashioned approach, it is still the only method that is used in this process. In the literature, there is not a newer approach for taxonomic classification.

* Corresponding author.
E-mail addresses: aliseydi@cs.hacettepe.edu.tr, aliseydikeceli@gmail.com (A.S. Keçeli), aydinkaya@cs.hacettepe.edu.tr (A. Kaya), sedaujeo@hacettepe.edu.tr (S.U. Keçeli).
[1] All authors have equal contribution.

http://dx.doi.org/10.1016/j.cageo.2017.08.011
Received 23 February 2017; Received in revised form 2 August 2017; Accepted 10 August 2017
Available online 16 August 2017

The automatic classification problem has been explained in taxonomy studies previously. Apostol et al. (2016) proposed a method for classification of Radiolarian fossil images. A translation, rotation and scale invariant method that utilizes Fourier transform is proposed by Fimbres-Castro et al. (2013) to identify radiolarians.

Also there are a variety of methods in automatic classification and retrieval of plankton images. Segmenting the region of interest (ROI), feature extraction and finding a similarity or classifying image are the main steps of these systems. There are many image segmentation algorithms available e.g. thresholding-based, color-based, texture-based, model-based, etc. (Shapiro and Stockman, 2001). Wavelet descriptors (Arai, 2013; Landre and Truchetet, 2001) and morphological features like textural features (Hu and Davis, 2005), granulometric features (Luo et al., 2004; Zhao et al., 2010), the contour and boundary shape features (Arai and Rahmad, 2013; Rahmad and Arai, 2015), histogram of oriented gradients (HOG) (Bi et al., 2015), etc. are used commonly in the classification and image retrieval approaches. Fourier transform and Fourier masks are used in diatom and phytoplanktons image classification (Barajas-Garcia et al., 2016; Solorza and Alvarez-Borrego, 2015; Ventura et al., 2015).

Content-based image retrieval studies mainly focus on unsupervised matching unknown image to the known/labeled set of images and determining the similarity. Arai (2013), Arai and Rahmad (2013), Rahmad and Arai (2015) have used Euclidean distance, Landre and Truchetet (2001) have used iterative clustering as similarity measure. Compared to CBIR systems, classification approaches have studied more. SVM classifiers (Bi et al., 2015; Hu and Davis, 2005; Luo et al., 2004; Sosik and Olson, 2007); ensemble tree classifiers (Bell and Hopcroft, 2008; Zhao et al., 2010); discriminant classifiers (Li et al., 2014); statistical methods (Pearl et al., 2013); active learning approaches (Luo et al., 2005) are presented in several studies. There are semi-automatic classification studies, besides the automatic classification ones. Ye et al. (2011) proposed a Bayesian probabilistic model and Mike et al. (2016) proposed computational geometric model-based approach as semi-automatic methods. Deep learning is a new approach to machine learning that is being used in many research areas (Yu and Deng, 2011). The deep learning models must be trained with a large amount of data, and there are datasets with many instances available in plankton image studies (Sosik and Olson, 2007). Orenstein et al. (2015) proposed two convolutional neural network (CNN) approaches on plankton images. Al-Barazanchi et al. (2015) proposed hybrid CNN approaches for data collected with the shadowed imaged particle profiling and evaluation recorder (SIPPER) (Samson et al., 2002). Lee et al. (2016) presented a transfer learning approach and compared the results with CNN with/without transfer learning.

The genera and species of the radiolarians are hard to determine because there are hundreds of species that vary in morphological features. Utilization of machine learning methods could ease this process as in the plankton image studies. The aim of this study is to help radiolarian paleontologists to be quicker in systematic studies and to help paleontologists to learn about the radiolarian systematics in a comprehensible way. In the proposed method, hand-crafted features (morphological and textural) and deep features are extracted and, feature selection is applied to select more discriminative features. Extracted features are grouped as hand-crafted, deep and combined sets. Classification algorithms are trained with these feature sets and results are analyzed.

In later sections, we first give details on experiment dataset, feature extraction and feature selection methods. In the experiment section, we present the classification methods and results. Finally, we discuss the results of our work and future plans.

2. Method

The flow of the proposed system can be divided into three main parts. First part is preprocessing step that consists segmentation of radiolarians. Pre-processing is made to select radiolarian body and remove unwanted structures. The second part is feature extraction. Hand-crafted and deep features are extracted from the extracted radiolarian silhouette. A pre-trained CNN is used to extract deep features. We use AlexNet (Krizhevsky et al., 2012) which is trained on Imagenet dataset. The activation values from the fully connected layers (Fc6, Fc7, and Fc8) of Alexnet are used as deep features. Each fully connected layer describes the image with a different abstraction. Deep features are high dimensional data, hence Relieff (Kononenko et al., 1997) feature selection algorithm is applied to select the strong features. The last step is the classification step and various classifiers are trained with selected deep and hand-crafted features for prediction. The trained models are tested with k-fold cross validation. In Fig. 1 the workflow of the proposed method is shown.

2.1. Dataset

There are 381 known Triassic radiolarian genera and most have been published since the 1970's following the introduction of Scanning Electron Microscopy (SEM). Seventy-three genera as synonyms and 11 genera as homonym were declared and 282 genera validated as Triassic in the organized image catalog of Mesozoic Radiolarians (O'Dogherty et al., 2009b), the revision made by the InterRad Mesozoic Working Group. In this study, the sample image data set is composed of SEM photographs of 27 species of Triassic radiolarians. Selected species have at least 10 sample images with good preservation and have high-quality SEM images through literature up until today and are chosen randomly from the studies carried around the world such as USA, Austria, Switzerland, Italy, Slovenia, Croatia, Slovenia, Serbia, Greece, Cyprus, Turkey, Japan, China, Thailand, Philippines and New Zealand which are cited in Table 1. In the literature, there are some hand-drawn radiolarian images from outdated literature or SEM images with bad preservation, some images of broken species or aff. and cf. of the species (samples without integrity). Most of the species does not have enough SEM images or images with good preservation. They are not taken into consideration. Sample radiolarian images of 4 different species are shown in Fig. 2.

The species in the dataset belong to three orders of radiolarians which are Entactinaria Kozur and Mostler (1982); Spumellaria Ehrenberg (1875) and Nassellaria Ehrenberg (1875). Mentioned species in the sample image data set are represented in Table 1.

2.2. Pre-processing

Due to varying image quality, a pre-processing step is applied to standardizing the images. Before feature extraction, non-image structures like tag characters or other text structures printed by the SEM imaging system should first be removed. In order to achieve this, a segmentation step is applied on images in the dataset. First, Otsu's thresholding method (Otsu, 1975) is applied on images, and binary maps are generated. Then, connected component analysis is applied to the binary image and the largest connected component is selected as radiolarian. Next, in the original image pixels that are not in the segmented binary mask is set to zero (0). In Fig. 3, original image, binary mask and segmented image are shown.

2.3. Features

After obtaining the mask, feature extraction is performed on the segmented image. Morphological, texture-based and deep features are extracted from each sample. The morphological features are eccentricity, solidity circularity, aspect ratio, area bounding box, moments and granulometric features. The texture features are Gabor features and wavelet moments. Some of these features like granulometric and GLCM are generally used in studies about microbiological studies (Arai, 2013; Arai and Rahmad, 2013; Bi et al., 2015; Rahmad and Arai, 2015). We utilize these features in our study too. The other morphological features and moment invariants are utilized however pure amorphous silica skeleton is the most distinct morphological feature of the radiolarians (De