Pattern Recognition xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Pattern Recognition

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

Indexing and mining large-scale neuron databases using maximum inner product search

Zhongyu Li[a], Ruogu Fang[b], Fumin Shen[c], Amin Katouzian[d], Shaoting Zhang[a,*]

[a] Department of Computer Science, University of North Carolina at Charlotte, USA
[b] School of Computing and Information Sciences, Florida International University, USA
[c] University of Electronic Science and Technology of China, China
[d] IBM, Almaden Research Center, San Jose, CA, USA

ARTICLE INFO

Keywords:
Neuron morphology
Large-scale retrieval
Binary coding

ABSTRACT

Morphological retrieval is an effective approach to explore large-scale neuronal databases, as the morphology is correlated with neuronal types, regions, functions, etc. In this paper, we focus on the neuron identification and analysis via morphological retrieval. In our proposed framework, multiple features are extracted to represent 3D neuron data. Because each feature reflects different levels of similarity between neurons, we group features into different hierarchies to compute the similarity matrix. Then, compact binary codes are generated from hierarchical features for efficient similarity search. Since neuronal cells usually have tree-topology structure, it is hard to distinguish different types of neurons simply via traditional binary coding or hashing methods based on Euclidean distance metric and/or linear hyperplanes. Therefore, we employ an asymmetric binary coding strategy based on the maximum inner product search (MIPS), which not only makes it easier to learn the binary coding functions, but also preserves the non-linear characteristics of the neuron morphological data. We evaluate the proposed method on more than 17,000 neurons, by validating the retrieved neurons with associated cell types and brain regions. Experimental results show the superiority of our approach in neuron morphological retrieval compared with other state-of-the-art methods. Moreover, we demonstrate its potential use cases in the identification and analysis of neuron characteristics from large neuron databases.

1. Introduction

How the brain works is one of the most challenging issues in neuroscience. As neurons are the basic elements of the brain, understanding their properties and network connectivity is a key step to tackle this challenging problem. There are approximately 86 billion neurons in the human brain and no two neurons are exactly the same. Figuring out each neuron's properties is difficult. Generally, neurons tend to express distinct morphologies based on their cell types, brain regions and functions. Therefore, it is reasonable to explore the neuronal properties through their morphologies. Recent development in visualization and image processing techniques [12,22,26,31,32] enabled accurate segmentation, tracing and reconstruction of 3D neuronal models from microscopic images. Meanwhile, the fast growing 3D neuron image databases such as NeuroMorpho [1,28] provide a public platform to associate neuronal properties with morphologies. Therefore, morphology-based neuron retrieval becomes an effective way to assist neuroscientists to identify unknown neurons and discover the relationship between the neuronal morphology and the property.

Morphology-based neuron retrieval is made possible because of the recent rapid advancements in neuron tracing techniques [4,10,25,45]. Costa et al. [5] proposed the concept of neuromorphological space, which analyzed the tree-like shape and designed quantified measurements of neuron cell. Wan et al. [40] designed *BlastNeuron* for automated comparison, retrieval and clustering of 3D neuron morphologies. In the retrieval stage, *BlastNeuron* searches for similar neurons via the normalization of rank scores in terms of the similarity of feature vectors. Despite its high accuracy, this method could be inefficient when handling a large-scale neuron database. Therefore, Mesbah et al. [24] proposed a data-driven hashing scheme, i.e., hashing forest, to search among large neuron databases. By establishing multiple unsupervised random forests, 128 or more binary bits are generated to represent morphological features. Hashing forest algorithm has achieved efficient and accurate results in neuron retrieval. Nonetheless, it usually needs a large number of bits (e.g., larger than 128), so its efficiency can be further improved with shorter binary codes. More importantly, the encoding process relies on the embedding of the Euclidean distance, which may not be a suitable similarity

* Corresponding author at: UNC Charlotte, Computer Science, 28223 Charlotte, NC, United States.
E-mail address: szhang16@uncc.edu (S. Zhang).

http://dx.doi.org/10.1016/j.patcog.2016.09.041
Received 30 December 2015; Received in revised form 3 June 2016; Accepted 21 September 2016
Available online xxxx
0031-3203/

measure for neuron retrieval issue, as features of neuron data usually lay in complex feature spaces that may not be linearly separable. Therefore, advanced hashing algorithms are important to solve these challenges for efficient and precise retrieval.

As described in Ref. [24], binary coding and hashing techniques have achieved great success in efficient retrieval among large-scale databases, with many methods proposed in recent years, including, but not limited to, Spectral Hashing (SH) [42], Anchor Graph Hashing (AGH) [20], Iterative Quantization (ITQ) [9], and others [11,36,37,50]. However, they may not be directly applicable to the neuron retrieval problem, as the features of 3D neuron morphological data are dramatically different from 2D natural images, and different features usually reflect different level of neuron similarity. For example, the tree-like structure imposes a challenge to differentiate neuron types, since treating all features with equal importance may lead to inaccurate search results. In addition, although supervised binary coding and hashing methods have already been investigated in medical image analysis [19,48,49], it is preferred to employ unsupervised methods for neuron retrieval, since the annotations for neurons may be incomplete.

Although neuromorphology and binary coding are both well-studied in recent years, how to combine them for neuron retrieval remains a challenging problem. Specifically, there are three challenges in binary coding based neuron morphological retrieval:

1. The feature vectors of each neuron are much shorter 30–50 dimensions) compared with traditional 2D images' feature vectors 100–10000 dimensions). Therefore, only much shorter bits of binary codes can guarantee the retrieval efficiency. Employing much shorter binary codes to represent large-scale neuron data is a great challenge;
2. Despite the limited length of neuron feature vectors, each type of feature has their specific meaning, e.g., branch number reflects the connection of neuron cell. Bipolar neurons have two branches, while multipolar neurons have three or more branches connected with other neurons. Currently, most image retrieval methods are either addressing the single feature's binary coding or fusing multiple features in different retrieval stage [3,21,44,46,47]. However, in neuron retrieval problem, each single feature is too short to obtain reliable retrieval results, and fusing multiple features is usually time-consuming. Thus, the specific biological indication and the computational complexity in neuronal feature representation need to be considered.
3. As each neuromorphological feature is extracted based on the tree-like structure, this limitation of feature extraction may cause a tough question, in which the tree-like structure will lead to similar features extracted from different types of neurons, e.g., some unrelated neurons express similar feature vectors. How to differentiate them in non-linear space is a hard problem.

In this paper, we design a binary coding framework to effectively and efficiently analyze large neuron databases. This framework is based on the recent progress of the maximum inner product search, which was proposed for image retrieval [35]. Specifically, we employ the method in Ref. [35] as the baseline, and then adapt it to handle multiple features or feature hierarchies, which is necessary to achieve high precision in this neuron retrieval. We validate the efficacy of the proposed method in the neuron retrieval problem with a large-scale database, and it outperforms several other binary coding or hashing methods. In addition, according to the neuron information provided by NeuroMorpho [28], our proposed method can retrieve similar neurons in terms of the morphology, cell types and brain regions.

The remaining paper is organized as follows: Section 2 briefly reviews the work related to 3D neuron morphology and binary coding methods. Section 3 provides the details of the proposed MIPS based binary coding with feature hierarchy for neuron retrieval system, followed by experiment results and discuss its potential use case in Section 4. Finally, Section 5 concludes the paper and presents future work.

2. Related work

2.1. Neuron tracing for 3d neuron morphology

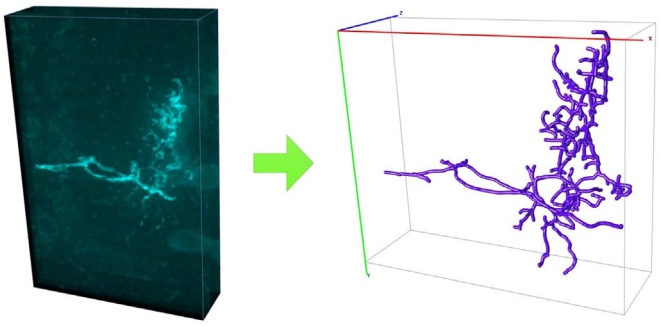
Neuron tracing aims to manually or automatically reconstruct 3D neuron morphology from fluorescence or electron microscopy images. Compared with 2D neuron image, 3D morphological data reflect spatial structure of the 3D neuron cell with more comprehensive information [51]. From the original microscopy images to the 3D neuron morphological data, neuron tracing consists a number of processing step, including image preprocessing (e.g., noise reduction, deconvolution, mosaicking), segmentation (e.g., soma, dendritic trees, spines, axons segmentation), reconstruction and connection [6,7,23,39,52]. In recent years, there are many tracing and reconstruction software released which make the 3D neuron morphological data easier to acquire. Fig. 1 illustrates a microscopy image from neuron slices [26] and its corresponding 3D morphological data through *Vaa3D* [32]. As shown in Fig. 1(b), morphological data provides more precise and quantitative measurements for neuron cells which facilitate features extraction for further retrieval and analysis..

Benefited from algorithms and software for neuron tracing, more and more 3D morphological databases are released in recent years. Unlike 2D medical images which can extract features with many well-studied algorithms, how to extract features from 3D neuron data is still an unsolved problem. For neuron cells, from axon to soma and then to dendrite, they usually express a tree-like structure. In Refs. [5,43], the authors introduced many quantitative measurements to analysis the tree-like structure of neuron cells.

Therefore, we can utilize these quantitative measurements as neuron morphological features. Specifically, we calculate three levels of measurements in order to reflect neuron morphology more comprehensive:

1. Global measurements, such as neuron's total height, depth, volume, etc. This level of features can express the holistic information of neuron cells;
2. Branch measurements, such as the Euclidean distance from compartments to somas, branch length, etc. This level of features denotes the information of neuron branches that are directly connected to the soma;
3. Bifurcation measurements, such as the angle between two terminal branches, etc. This level of features reflects the bifurcation's information of branches not directly connected to the soma.

In this paper, we calculate in total 38 measurements at the above three levels. Then we assemble them as morphological features to represent each neuron cell for further retrieval and analysis.



(a) (b)

Fig. 1. From original microscopy neuron image to 3D morphological data: (a) original microscopy slices; (b) 3D neuron morphology with quantitative measures.