Reviewing Brain Image Databases and Connectomes

Southern Illinois University Edwardsville

CS 434 - Database Management Systems

David Shimkus

dashimk@siue.edu

June 30, 2022

Abstract

One of the most complex structures known to man is the human brain. Emerson Pugh, former president of the Institute of Electrical and Electronics Engineers (IEEE), is credited with saying, "If the human brain were so simple that we could understand it, we would be so simple that we couldn't." Nevertheless, studying the brain not only has esoteric benefits but also has a variety of tangible medical applications. Using non-intrusive *in vivo* methods for studying this organ, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scanning, has allowed scientists and medical professionals insights into this complex organ. As these types of technologies and modalities improve the granularity of their data aggregation, their storage requirements increase [3]. This review aims to summarize different database architectures and data types used for brain analyses. It concludes by presenting associated tradeoffs and limitations of the different systems as an opening to further research, and a call for more granular imaging techniques and larger and more efficient database systems.

Introduction - Brain Image Databases

There are many ways to represent a brain with data. The most straight forward method is to store image scans of the brain into a pictorial database optimized for images. Medical imaging and diagnostics in this field have been improving rapidly in the last few decades [8]. Many processes nowadays generate hundreds if not thousands of images for each subject instance [3]. These data sets are often collections from different angles forming 3D structures, as well as incorporating a time series as a fourth dimension [3]. Taking small sections or patches of the

brain at a time can be treated the same as voxel morphometry and follow relational principles associated with this discipline [2].

One cannot treat the human brain as simple data though, as tweaking voxel size and other parameters requires a somewhat thorough understanding of an average brain's physical makeup. The human brain has been widely studied for centuries and different means of labelling and classification have arisen long before modern imaging and mapping techniques were implemented [10]. One such method of cataloguing the brain is known as creating an "atlas" which is most commonly a series of parallel cross sections taken in a 3-dimensional coordinate system (sometimes referred to as Talairach Space) and then listed together in a set as shown in Figure 1 [5, 10]. Brain atlases form the foundation of the means of classification of different sections, and all subsections are defined in these types of buckets (such as certain sections belonging to one hemisphere). By structuring a database with these types of existing classifications in mind, scientists and database engineers can enforce logical and real-world constraints that neurologists and others have already defined.

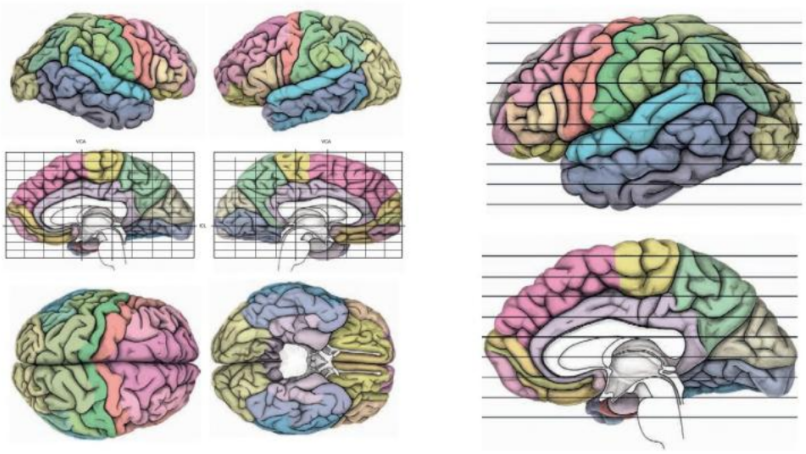

Fig. 1 - Different views of the brain illustrating the atlas concept [10]

These atlas layers are not all scanned at the same time, but happen over a certain period, with a time delta applied between each layer. Aggregating these layers together with respect to time can help illustrate how fluids move through tracts and portions of the brain which can help with analysis [2]. When storing these types of scans into a database it is important to bear in mind this spatial-temporal nature of the data.

The Digital Imaging and Communications in Medicine (DICOM) international standard is a good baseline and benchmark for granularity and storage requirements for modern medical images [9]. For example, following this standard a single instance of a Computed Tomography Perfusion (CTP) scan of the brain can result in around 5 - 10 GB of images [3]. The CTP scan is one of many types of brain imaging techniques. Other well-known procedures include MRI, PET scans, and Electroencephalography (EEG). High-resolution microscopic scans of a single brain can require 10's of terabytes of storage [5]. Implementing any Brain Image Database (BRAID) with appropriate spatial and temporal resolutions can require unique database implementations utilizing compression, delineation, data mining, and medical specific algorithms [1, 3, 6, 8].

Introduction - Connectomes

Another common way to attempt to represent the brain is with a mapping of the neuronal connections, known as a connectome [8]. Connectomes are conceptually different than image databases in that they attempt to capture each neuron and surrounding cells as individual units and illustrate connections with each other in a graph-like structure as illustrated in Figure 2 [5, 8]. In this context if neurons are described as nodes, then functional connectivity can be

described as mapping signal activation frequency or strength to edge weights between the connected neurons [5]. Note that electric activations generally flow in one direction along edges, forming a type of directed graph. Although there is indeed a type of backpropagation that occurs to strengthen the edge weights that is beyond the scope of this review [14].

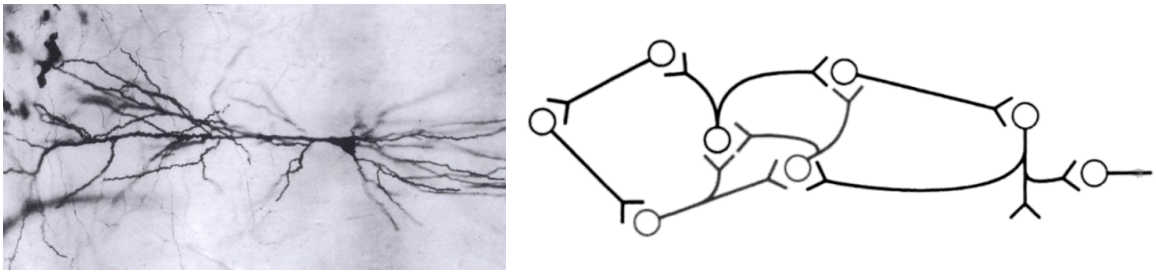

Fig. 2 - Micrograph of a neuron and directed connected graph [14]

The core issue with connectome data is that it is almost always derived from imaging data as its source [8]. One might be able to build a complete data descriptive model of a neuron and connectome, but how they are organized must be based in real-world scenarios. Scanning techniques for this input such as MRI modalities can have a spatial resolution of around 0.7 mm isotropic voxels [7]. A cubic millimeter of human or primate brain tissue can contain roughly 50,000 neurons of which each could have around 6,000 connections to neighboring cells [13]. This roughly equates to a single MRI voxel containing around 35,000 neurons. The average adult human brain is approximately 1,400 cubic centimeters [7]. These estimations very roughly align to total a product of 70 billion neurons, where the reported average of neurons in adult human brains is between 85 billion to 100 billion [7]. However, neuronal density has many intricacies and does not necessarily scale directly with size as is evident with primates and

humans having much higher neuronal counts than that of larger brained animals such as the whale or elephant [7].

Problem Description

Depending on what type of question one is trying to answer, a connectome may be considered too much information and a normal image database may suffice. Ultimately, a need for a complete connectome representing an average adult human brain is necessary for the advancement of neuroscience and humanitarian fields. **Today's neuroimaging captures data at around a cubic millimeter voxel scale while a full connectome representation requires a scale of 10's of nanometers. Even if such a granular imaging technique were implemented, current database systems would be unable to store or query such an exabyte-scale data set effectively.** Using compression, estimation, aggregation, and other workaround techniques show promise but have drawbacks as well.

Related Work - Brain Image Databases

Because it is not currently feasible to fully map every neuron and connection in an individual human subject's brain, compromises and alternative approaches are used depending on what one is trying to accomplish. For example, diagnosis and treatment of various diseases often do not require neuron-level granularity and an image database of scans is sufficient [4]. Because of this, and the fact that connectomes are built from these types of images, it is still prudent to implement and optimize these types of BRAIDs.

It is critical to note and emphasize that every human brain is affected by various factors such as genetics, age, disease, and more. There are no two brains exactly alike, and the brain changes over time [7]. Studying averages and other types of combinations of multiple subjects can give rise to a way to abstract high level functions of different portions of the brain and how they communicate with each other [5]. Reducing bias by studying a widely diverse range of subjects may assist in producing better aggregate data but may risk over-normalizing data that may be particular to certain demographics [11].

There are generally three different database-management system (DBMS) architectures that could be used to implement such a BRAID: relational (RDBMS), object-oriented (OODBMS), and object-relational (ORDBMS) [1]. It is suggested that an ORDBMS is the best choice when managing complex data such as images and processing complex queries. Choosing to implement a BRAID with an ORDBMS has been shown to afford more extensibility over a traditional RDBMS, yet with apparently less overhead than that of a OODBMS [1]. This type of model not only supports traditional Standard Query Language (SQL) queries but provides object level functions as well which is potentially useful when dealing with large binary data such as RGB images. One of the key benefits of storing the data in this type of way is for data mining and machine learning.

An effective way to store MRI and other modalities' image data is combining the atlas and voxel concept with pre-existing RGB storage concepts. For example, suppose a specific atlas is being used that defines the brain regions into the traditional Talairach Space dimensions. These dimensions are referenced in the traditional anatomical planes of Coronal, Sagittal, and Axial (sometimes known as transverse) dimensions. These provide the coordinate system to be used

as indexes alongside the RGB data in the database [1, 15]. Figure 3 illustrates that these axes are no different from normal Euclidean space. An example "object-tuple" in an ORDBMS could be described as in the format of (zyxx, raw) - indicating the zyxx format represents a line segment (z, y, x_{begin}, x_{end}) with a corresponding raw RGB value or signal strength [1]. These object-tuples would be in one of the anatomical dimensions and have a corresponding voxel size. Lastly, the image data itself can finally be stored against this tuple. Image resolution does not necessarily correlate to voxel size, and this can be an important distinction to make when discussing compression techniques.

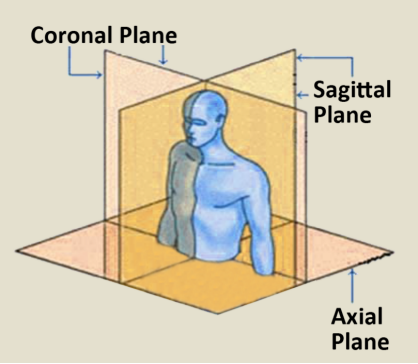

Fig. 3 - anatomical planes [15]

Once all the voxel, image, and signal strength tuples have been defined against the appropriate anatomical planes, the fourth dimensional time series relation can be implemented. This concept is important not only due to the nature of how a single subject's image scanning takes place over time but provides an opportunity for an additional dimension to compress the data against. Utilizing redundancy concepts in compression methods against this time dimension can result in an average compression rate of 0.53 and space saving of more than 47% [3]. This can be quite significant considering a single subject's scan can be on the order of gigabytes magnitude depending on the experiment protocols' time duration and modality [3].

With a framework and schema for a single subject's BRAID now laid out, it is prudent to expand this concept for multiple subjects. Three main entities - STUDY, SUBJECT, and SUBJECT_IMAGE can form the backbone of the multi-subject BRAID, upon which the more granular object-tuples of the scans themselves can be stored against [1]. In this way, a hybrid RDBMS and ORDBMS may prove to be the most effective. Figure 4 provides an example diagram for such a schema, and ties in the concepts discussed previously.

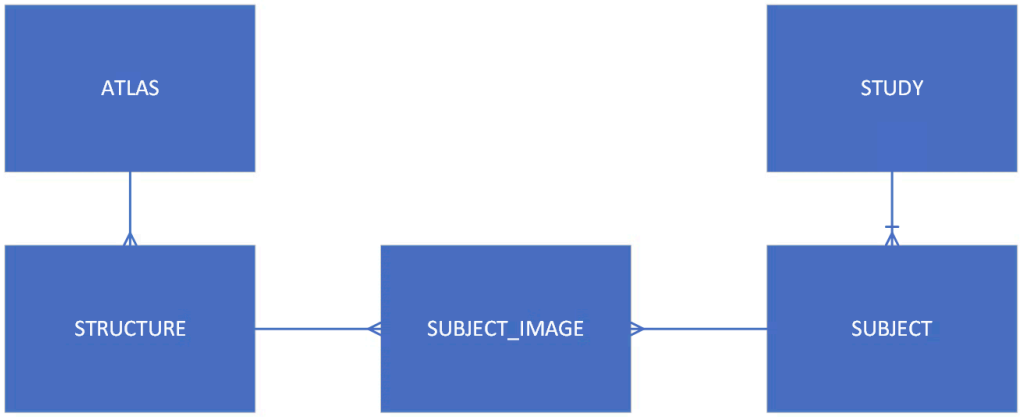

Fig. 4 - Multi-subject BRAID entity-relationship diagram

Content-based medical image retrieval (CBMIR) is an entire field of study and methodologies to appropriately handle large amounts of medical image data. As different studies can provide wildly different images based on modalities, resolution, voxel size, and more, the need for intelligent methods of sifting through the data arises. Again focusing on MRI data, researchers have demonstrated feature extraction methods to apply tags to images in a deep learning type of application [4]. This allows a user to not have to sift through many images but rather have the algorithms classify images to make the database more useable. An example use case for such a scenario would be flagging different scans for deleterious patterns such as concussions.

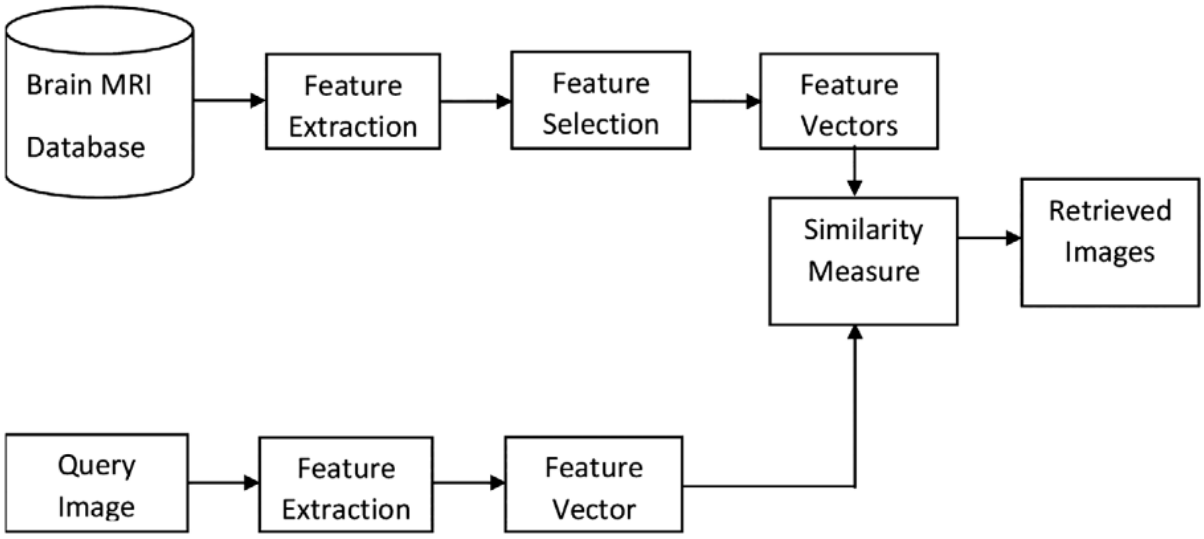

Fig. 5 - CBMIR feature extraction [4]

Figure 5 shows such a feature extraction type of approach, where the classification of subjects' images happens via algorithmic processes as well as classification of the user's query. A similarity comparison is used against the extrapolated vectors to arrive at a proposed set of relevant images with an associated confidence interval [4]. Because of the very large datasets obtained during MRI scans, by focusing on attributes such as color, shape, and texture the feature extraction methods can provide average precisions of more than 95% for a given input image and desired classification such as Alzheimer's disease and stroke [4].

Related Work - Connectomes

Bridging the gap between BRAIDs and connectomes involves imaging the "tracts" and connections in the brain on top of the voxel-based atlas models. Diffusion-weighted (DW) MRI and functional MRI are two common types of MRI used for this input. While both methods measure activity differently, they each can be used for tractography algorithms [1, 2]. DW MRI

measures proton displacement in water as it flows through portions of the brain and is a relatively newer technology. Even with challenges such as the inevitable presence of imaging noise in the data, DW has a low signal-to-noise ratio (SNR) and helps provide more precise orientations of brain fibers [2].

In contrast to DW MRI, blood oxygenation level dependent (BOLD) based functional MRI data provides another means to illustrate "pathways" in the brain [1]. Tweaking parameters used to measure oxygenation levels on images with a color gradient indicating signal strength can reveal excellent visualizations of these connections.

Regardless of methods used to acquire the data, they must ultimately be fed into an imaging program for display. Figure 6 illustrates a popular implementation of a tractography imaging software MRtrix. This software is freely available under the GNU General Public License and written exclusively in C++, again lending itself well to the ORDBS type model [2]. It natively supports common MRI file formats such as the DICOM, NIfTI and others [2]. Using these types of software against images in BRAIDs lets users begin to understand how certain portions of the brain interact with each other, and networks and subnetworks begin to emerge. For example, certain areas appear more active during memory recall like PET scans.

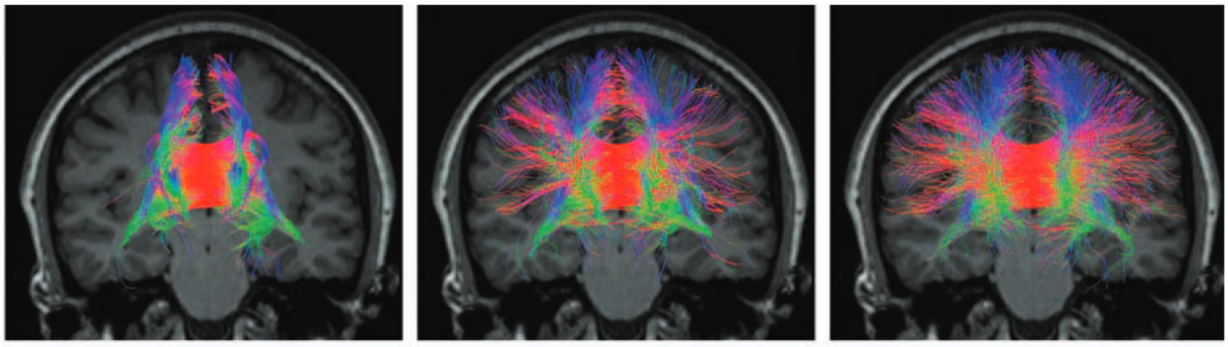

Fig. 6 - MRtrix displaying a DW image with different parameters [2]

Perhaps the most famous recent project to attempt to create a meaningful impact in the neuroscience field is the Human Connectome Project (HCP). This project began in 2010 and was an effort to advance data quality and availability. Over 27 petabytes of data have been shared from this project, and "HCP-style" neuroimaging has become a set of best-practice strategies for analysis [8, 12]. The HCP infrastructure is based on two independent XNAT databases (an imaging informatics platform), one for private use and one for public use [12]. Many preprocessing and storage pipelines were developed specifically for this project and warrants a separate review entirely. ConnectomeDB is the primary backend component to the public use database and can be accessed with custom client applications such as Connectome Workbench [12].

Analysis

While neuroimaging has advanced in leaps and bounds in the past decades, connectomics is still in its relative infancy. The advancement of imaging techniques and associated BRAIDs is the vital, living steppingstone required for the completion of an average adult human connectome. Only by bringing together the key ideas and concepts of both BRAIDs and connectomes can a secure foundation be laid. Imaging techniques seem likely to continue to advance rapidly, so staying prepared to accept and analyze larger and more granular data sets is an immediate challenge and focus.

The HCP really shined with sharing data where other systems struggled by focusing on providing consistently high download rates (~300 Mbps maximum) as well as providing physical hard drives shipped to recipients at no cost [12]. The lesson here is that there is a strong

humanitarian element to this type of data, and the advancement of these sciences benefits all. Developers producing high quality open-source tools for this data such as MRtrix further illustrates the ethical drivers behind this type of research. Another ethical topic to consider is the sensitivity of medical data amongst subjects, as well as the hard requirement to perform much of the analysis as minimally invasive as possible. Evidence shows that PET and MRI scans of animal brain regions is an increasingly popular method for tuning scanning parameters and while it shows promise to alleviate some of the ethical concerns, the variability amongst animal subjects can be greater than that of humans [6]. DICOM and other standards for file formats and data quality help unify efforts and prevent unnecessary roadblocks related to design conflicts and variability issues.

Ultimately, any brain exists in time and space and must be treated as such when forming storage schema. Data trends such as decreasing voxel size and increasing image resolution will also trend towards increasing granularity as technology advances. This does not prevent hardening the theory and methodologies to be used to store and access brain data.

Future Work

There seems to be a larger advancement in the machine learning approaches used to analyze image data as compared to the connectome data. There is indeed work being done in this field, and it suggests the structural connectivity of the brain (i.e. derived from images) only has a modest correlation to the functional connectivity of the brain (i.e. connectomics) [11]. There are challenges associated with how diverse individual brains can be, which affects functional predictions more so than structural predictions [11]. The only known fully connected

connectome in existence is that of a nematode, with the common fruit fly following closely behind [12]. A logical next step would be completing a full connectome of a mouse and primate brain before moving onto humans [7]. There will undoubtedly be many procedural and storage challenges identified and addressed in these early stages, and many opportunities for research and advancements in the immediate future.

Citations

[1] Herskovits, E., & Chen, R. (2008). Integrating data-mining support into a brain-image database using open-source components. Advances in Medical Sciences, 53(2). https://doi.org/10.2478/v10039-008-0009-9

[2] Tournier, J.-D., Calamante, F., & Connelly, A. (2012). MRtrix: Diffusion tractography in crossing fiber regions. International Journal of Imaging Systems and Technology, 22(1), 53–66. https://doi.org/10.1002/ima.22005

[3] Fahmi, F., Sagala, M. A., Nasution, T. H., & Anggraeny. (2016). Sequential — storage of differences approach in medical image data compression for Brain Image Dataset. 2016 International Seminar on Application for Technology of Information and Communication (ISemantic). https://doi.org/10.1109/isemantic.2016.7873822

[4] Sampathila, N., Pavithra, & Martis, R. J. (2020). Computational approach for content-based image retrieval of K-similar images from brain MR image database. Expert Systems. https://doi.org/10.1111/exsy.12652

[5] Varoquaux, G., & Craddock, R. C. (2013). Learning and comparing functional connectomes across subjects. NeuroImage, 80, 405–415. https://doi.org/10.1016/j.neuroimage.2013.04.007

[6] Villadsen, J., Hansen, H. D., Jørgensen, L. M., Keller, S. H., Andersen, F. L., Petersen, I. N., Knudsen, G. M., & Svarer, C. (2018). Automatic delineation of brain regions on MRI and PET images from the pig. Journal of Neuroscience Methods, 294, 51–58. https://doi.org/10.1016/j.jneumeth.2017.11.008

[7] Herculano-Houzel, Suzana. “The Human Brain in Numbers: A Linearly Scaled-up Primate Brain.” Frontiers in Human Neuroscience, vol. 3, 2009, https://doi.org/10.3389/neuro.09.031.2009.

[8] Van Essen, David C., et al. “The Wu-Minn Human Connectome Project: An Overview.” NeuroImage, vol. 80, 2013, pp. 62–79., https://doi.org/10.1016/j.neuroimage.2013.05.041.

[9] “DICOM - Digital Imaging and Communications in Medicine.” https://www.dicomstandard.org/about

[10] Mai Jürgen K., et al. Atlas of the Human Brain. Academic Press, 2016.

[11] Sarwar, T., et al. “Structure-Function Coupling in the Human Connectome: A Machine Learning Approach.” NeuroImage, vol. 226, 2021, p. 117609., https://doi.org/10.1016/j.neuroimage.2020.117609.

[12] Elam, Jennifer Stine, et al. “The Human Connectome Project: A Retrospective.” NeuroImage, vol. 244, 2021, p. 118543., https://doi.org/10.1016/j.neuroimage.2021.118543.

[13] https://www.rc.fas.harvard.edu/case-studies/connections-in-the-brain/

[14] Levitan, Irwin B., and Leonard K. Kaczmarek. The Neuron: Cell and Molecular Biology. Oxford University Press, 2015.

[15] Radiology Rounds. Image Reconstruction Planes. https://www.ipfradiologyrounds.com/hrct-primer/image-reconstruction/