Project mini-proposals

@rno klein, September 2007

Informatics

1. Journal Cannibalism
2. Heuristic Brain Atlas
3. The Oracle
4. PubRev
5. PDF Citation Extractor
6. Neuroimaging WikiDB

Image processing, feature matching, and optimization

7. Medial axis construction
8. Bayesian feature matching
9. Optimal evolutionary feature matching
10. Nonlinear deformation evaluation
11. Decision fusion and performance bias estimation
12. Cortical surface representation

Journal Cannibalism

Goal

Create an application and database that will store and inter-relate the disparate facts, speculations, and opinions expressed in scientific articles.

Background

Often when one reads a scientific paper, one runs across multiple facts or opinions that may have little bearing on the article's title, abstract, or conclusions but are nonetheless valuable pieces of information. If one has an assertion that needs corroboration, such as "two thirds of the brain's surface is buried within the folds of the brain," it would be highly desirable to be able to find articles and particular statements that corroborate or refute such a statement. It would also be extremely useful to trace the lineage of such statements through citations and chronology to derive the source of such ideas or data.

NLP relevance

From a natural language processing perspective, a worthy challenge would be to automatically summarize and cross-reference disparate facts extracted from semi-structured text (journal abstracts/articles) for corroborating or refuting statements.

Heuristic Brain Atlas

Goal
Create the ultimate reference of hints, tips, and wisdom regarding brain morphology.

Background
Information about covarying shapes within a brain are tucked away in disparate publications and databases. This information needs to be assembled and organized for use by and for the brain imaging community, to enable rule-based or statistically-weighted approaches to brain parcellation and clinical diagnosis.

Description
Various sources of information regarding anatomical statistical variability, covarying morphological patterns, well-conserved feature arrangements, etc. will be assembled and incorporated into a database of facts and figures about brain morphology. In conjunction with patient or subject data, this database will help direct clinical attention to morphological indicators of possible neurological conditions.

NLP relevance
A worthy challenge would be to automatically summarize and cross-reference disparate facts extracted from semi-structured text (journal abstracts/articles) for answering questions or corroborating statements.

The Oracle

Goal
Create an online resource where scientific researchers and lay people alike will be able to ask questions regarding biology and receive appropriate, up-to-date answers. Not a search engine that returns multiple hits, but a lucid, logical answer to a question posed in natural language.

Background
The Oracle would combine three movements in contemporary internet development:

- the wiki, in particular "knowledge collection from volunteer contributors" such as is used in the OpenMind project (http://www.openmind.org)
- ontologies/topic maps for organizing and integrating vast, disparate data sources
- machine learning approaches applied to implicit feedback measures (user data such as mouse click and query histories).

We can consider here the wiki to be used for data input, the ontology for data organization, and learning algorithms for analyzing data for inferring meaning. Considerable effort is required in natural language processing to do the second and third steps effectively. Organization also demands an effective data model, system of categorization, or "ontology," and inferring meaning demands an underpinning of commonsense.

Ontologies under development

- NLM's Unified Medical Language System (http://www.nlm.nih.gov/research/umls/about_umls.html)
- NeuroNames (http://braininfo.rprc.washington.edu/indexabout.html)
- BIRNlex (http://132.239.132.249:8080/xwiki/bin/view/+BIRN-OTF-Public/About+BIRNLex)
- Ontology for Biomedical Investigations (http://obi.sourceforge.net/)

A few working commonsense examples

START (InfoLab Group, MIT)
Natural Language Question Answering System
http://start.csail.mit.edu

ConceptNet (Media Lab, MIT)
Commonsense knowledgebase and natural-language-processing toolkit
http://pedia.media.mit.edu/wiki/ConceptNet

Montylingua (Hugo Liu, Media Lab, MIT)
Commonsense-enriched natural language understander
http://web.media.mit.edu/~hugo/montylingua/

NLP relevance
This is a natural language question-answer system targeted first at medical students, then to clinical researchers, and finally to clinicians and the lay public.

PubRev

Goal
Create a browser-based interface and backend application for reviews of articles on PubMed.

Background
The only place to read reviews for scientific publications are for the few publications that provide editorial reviews, for open-access online journals such as those on biomedcentral.com, or informal comments posted on a disparate array of social bookmarking sites, such as citeulike.org. What is desperately needed to understand and use reactions to the scientific literature is a cohesive, central venue for finding and writing reviews that are cross-referenced to the corresponding article(s). Either this could take the form of an add-on to PubMed itself, or it could be a separate website that performs PubMed searches and provides reviews of articles in addition to the articles themselves.

Description
The most useful features of popular social bookmarking websites like del.icio.us can be incorporated, such as searches by reviewer, reviewer tags, recommendations, etc. A few novel aspects may be addressed as well:

- When reading a review, one should be able to indicate one's dis/agreement with a review in addition to dis/approval of the article itself.
- Reviews for related articles should somehow be brought to the attention of the reader so that issues raised in the review of one article may have been addressed in the related article. The reviewer for the second article should not have to write a review from scratch as well.
- Web analytics (number of visitors and number of reviews/reviewers, etc.) should be displayed.

Requirements
A proof-of-concept implementation could take the form of a website that performs PubMed searches just for open-access online journals, such as those of biomedcentral.com, extracts the prepublication history of visited articles, and places the reviews in a visually compact and easily navigable form alongside the article upon request. To address concerns about the quality of wiki versus peer reviews (see Satra's comment below), publisher reviews could serve as the primary "official" reviews that are rated and commented on by visitors, and wiki entries could be moderated in automated ways (natural language processing for keyword relevance, emotional terms, etc.) and manual ways (moderators could consist of article authors, peer reviewers, publishers, or general editors). The interface for adding comments could also be highly structured, greatly reducing low-quality entries.

NLP relevance
1. Automated moderation of opinions would mitigate flaming and irrelevance.
2. Assigning affective type and valence (pro or con) to citations within journal articles could automate opinion gathering about a field.

PDF Citation Extractor

Goal
Extract citation information directly from PDF documents, to incorporate existing collections of documents into a document/citation management system, such as Zotero.

Background
The bulk of a researcher's time while conducting a literature search is (1) finding articles, (2) incorporating articles of interest into a document management system and their citations into a citation management system, and (3) formatting subcollections of citations when creating or sharing bibliographies. There are search engines for quickly finding relevant articles and there exist software programs for formatting bibliographies, but there does not seem to exist a satisfactory system for organizing documents such that they are addressable from the citation reference manager, unless those documents were initially received using that manager.

Description
We propose to create a program to extract the citation from a PDF, export the citation for management software (such as Zotero), store a copy of the PDF in a customized way, and cross-reference the PDF with its citation. The difficult step is extracting the citation, since publishers do not make good use of PDF metatags, and text parsing is never perfect for unstructured documents. One idea for getting the right citation for a given PDF is to feed a tentative citation as keywords in a PubMed search, and depending on the search results, expand or refine the query until a match is found.

> Jay Bohland (Cold Spring Harbor): "Store (in a database) an association between a unique hash created from some attributes of the pdf text and the citation. This begins to create a database against which other users can query their own pdf's more directly (I consider going through Pubmed, etc. "indirect"). The hash would need to be derived from document text and would have to be sufficiently unique. This is perfectly analogous to the situation with music metadata and CDDB - which uses a hash of the file lengths on a cd to find the proper metadata (which only has to be "solved for" or entered once). This also helps to solve problems related to non-standard pubs (book chapters, different scans of the same article, etc) and those not listed by databases we can query. It also eventually sidesteps having to query pubmed so many times, and with 3s delays between each query. Finally, I don't think copyright would be an issue - you would never be sending the document text across the net - just some hash derived from it."

Requirements
Initial code has just been written by Satrajit Ghosh at MIT for extracting text from a PDF, and automatically finding a likely citation through PubMed. Jay Bohland of Cold Spring Harbor Laboratory is evaluating the code and will be working with their scientific programmer to expand its capabilities. Both parties have expressed interest in creating a publicly available tool with full functionality. Discussions are ongoing.

NLP relevance
1. Extracting and parsing text could lead to gross characterization of content for automatically inferring relationships between articles.
2. Citation spidering could help categorize an article based on collaborations and opinionated statements about the cited articles within other articles.

Neuroimaging WikiDB

Goal
Create a community-editable web resource (wiki) with a database containing a careful review of brain atlases and parcellation protocols, and brain image registration techniques. An emphasis would be placed on evaluation methods to compare the different approaches in the literature and in software, and to recommend how different approaches lend themselves to particular applications. To navigate the information more effectively, corresponding interactive visualizations would form part of the interface to the website.

Background
The corresponding author for the April 2007 IEEE Transactions on Medical Imaging paper "Brain Functional Localization: A Survey of Image Registration Techniques," Ali Gholipour, kindly gave me permission to use his data taken from the 330 citations in his paper for such a website. David Kennedy and Christian Haselgrove of the soon-to-be-released Neuroimaging Informatics Tools and Resources Clearinghouse (http://www.nitrc.org) have expressed interest in incorporating results from the resource for inclusion in their NIH-funded website, poised to become the center for software discussion and dissemination by and for the neuroimaging community.

Requirements
A student would be required to expand the existing database considerably, and to give particular scrutiny to the methods sections of all relevant papers to make useful comparisons. The student would launch this database as a community-editable wiki so that it may continue to expand. A successful student who produces a detailed summary report in the form of a thesis will have demonstrated mastery of the field, and hence a Master's degree would be appropriate. If software evaluations were conducted and a novel, competitive method and evaluation scheme were to arise, or if a theoretical formulation were to address image correspondence or shape similarity, this would be worthy of a PhD thesis.

NLP relevance
1. Extracting and parsing text could lead to gross characterization of content for automatically inferring relationships between articles.
2. Citation spidering could help categorize and rate algorithms in articles based on collaborations and opinions expressed about those citations.

Medial axis construction

The Mindboggle software package (http://www.binarybottle.com/mindboggle.html), used for automating anatomical labeling of human brain MRI data, matches features across brains as one of its principal steps. These features consist of skeletons of non-white matter, primarily cortical folds of the brain.

Presently, skeletal construction is performed on each slice independently, and the resulting slice-stack is further processed to construct a 3-D skeleton. A major problem with this method is that skeletal construction is affected by how the image is sliced. To mitigate this problem, we have tested a new technique of creating a 3-D skeleton in a slice-wise manner along x, y, as well as z axes. Although we can then run Mindboggle three times on the same brain and combine matching results across all three runs, the axes themselves are still merely an orientation convention and will not adequately handle brain folds lying along any axial orientation.

Student project
Implement a new "exact medial axis" algorithm for creating a 3-D skeleton, based on Thiel and Remy (2005, C code available). This will avoid having to choose along which axis a skeletal slice-stack should be constructed, avoid skeletal artifacts such as those arising from folds lying along a slice, increase accuracy and reduce inconsistencies across acquisitions and orientations, and result in much faster processing times.

The student(s) will gain an appreciation of some significant challenges in biomedical image processing, as well as characterization of anatomically relevant shapes.

References

1. Remy, E., and E. Thiel. “Exact medial axis with euclidean distance.” *Image and Vision Computing* 23, no. 2 (February 1, 2005): 167-175.
2. Couprie, Michel, David Coeurjolly, and Rita Zrour. “Discrete bisector function and Euclidean skeleton in 2D and 3D.” *Image Vision Comput* 25, no. 10 (2007): 1543-1556.
3. Saúde, Couprie, and Lotufo. “Exact Euclidean Medial Axis in Higher Resolution.” *Discrete Geometry for Computer Imagery*, 2006.

Bayesian feature matching

The Mindboggle software package (http://www.binarybottle.com/mindboggle.html), used for automating anatomical labeling of human brain MRI data, matches features across brains as one of its principal steps. Mindboggle presently combines pieces (of folds) within a brain and matches these features across brains in a rule-based manner, without drawing upon earlier experience such as information gained from a population of manually parcellated brains.

Student project
Statistically describe the variations and covariations in the morphology of our population of a set of 40 manually parcellated ROIs, and use this information in a Bayesian framework to guide feature matching. Below is an example approach:

We will adopt the general approach and slightly modify the nomenclature of a successful Bayesian framework for face identification (, 2005). In that work, the authors formulate a probabilistic similarity measure which is based on the probability that the differences between two images, Δ, are within the variations (*e.g.*, expressions) of the same face. In their formulation, they are solving a binary pattern classification problem between intraclass (same-face) variation, Ω_I, and extraclass (different-face) variation, Ω_E. The similarity measure is the posterior probability that two images are within intraclass variation given a difference Δ between the images, according to Bayes' Theorem:

$S(I_1,I_2) = P(\Omega_I \mid \Delta) = P(\Delta \mid \Omega_I)\, P(\Omega_I)/P(\Delta)$
where $P(\Delta) = P(\Delta \mid \Omega_I)P(\Omega_I)+P(\Delta \mid \Omega_E)P(\Omega_E)$

The maximum *a posteriori* (MAP) rule is then used to solve the classification problem. Two images are considered to correspond to the same class if $P(\Omega_I \mid \Delta) > P(\Omega_E \mid \Delta)$, or equivalently, if $S(I_1,I_2) > 1/2$.

This is analogous to our feature-matching problem; just as they need to decide whether a candidate face corresponds to a given (identified) face, we need to decide whether a candidate feature (combination of pieces) corresponds to a given (labeled) feature (*e.g.*, the central sulcus). Our images I_1 and I_2 will correspond to images not of faces but of these features, and each class refers to the set of instances of a given feature (*e.g.*, central sulci in a set of brains). The difference Δ between two images will be computed by Mindboggle's cost function.

We will statistically describe intraclass variations Ω_I and extraclass variations Ω_E. We will estimate the likelihoods $P(\Delta \mid \Omega_I)$ and $P(\Delta \mid \Omega_E)$ and priors $P(\Omega_I)$ and $P(\Omega_E)$ from these variations (with the assumption that their distributions are Gaussian).

Calculating Bayesian priors should drastically reduce the number of likely combinations of pieces and the number of candidate matches to consider. This would make it computationally feasible to remove the stringent constraints in the present version of Mindboggle, namely, that only a maximum of three pieces are allowed to combine for matching with each single atlas piece.
-
Moghaddam, Baback, Tony Jebara, and Alex Pentland. “Bayesian Modeling of Facial Similarity.” *Advances in Neural Information Processing Systems (NIPS 11)*, 1998. http://citeseer.ist.psu.edu/moghaddam99bayesian.html

Optimal evolutionary feature matching

The Mindboggle software package (http://www.binarybottle.com/mindboggle.html), used for automating anatomical labeling of human brain MRI data, matches combinations of features across brains as one of its principal steps. Matching sulcus pieces is a significant combinatoric problem, and Mindboggle uses a simple, non-exhaustive, non-probabilistic, non-evolutionary strategy. The primary advantages of employing these characteristics are faster computation and elimination of intermediate results. An alternative would be to perform probabilistic matching to search the solution space more broadly. Genetic algorithms and simulated annealing are two such approaches; parallel recombinative simulated annealing (PRSA), a combination of the two, allows for parallelism and convergence to a global solution (Mahfoud, 1995).

Student project
Program a PRSA algorithm to run on brain feature matching problems.

Reference
Mahfoud, S. W, and D. E Goldberg. “Parallel recombinative simulated annealing: a genetic algorithm.” *Parallel Computing* 21 (1995): 1-28.

Nonlinear deformation evaluation

Atlas-based approaches to segmenting or labeling brain image data consist predominantly of warping algorithms. The primary problem with relying solely on warping to nonlinearly register one brain to another is that without sufficient constraints, there are many ways to reshape a brain to look like another without regard for anatomical borders. Indeed, image correspondence is often mistaken for anatomical correspondence. Because point correspondence from one cortex to another is ill-defined, some degree of manual intervention is often used to initially assign corresponding points or curves about which deformations are to be performed. It is also questionable to assume that intervening points between two well-defined landmarks correspond to intervening points between matching landmarks. The point correspondence problem is simply revisited at a smaller scale; if different anatomical structures happen to exist between the corresponding pairs of landmarks in two brains, there may be no point-to-point correspondence.

Evaluations that have been made of nonlinear deformation strategies are in general difficult to assess. They are usually demonstrated with restricted label sets or sparse landmarks, and evaluated under artificial conditions or most commonly by visual inspection where image correspondence can be mistaken for anatomic correspondence. We are aware of only a few studies that have compared different nonlinear registration algorithms (Hellier, *et.al.*, 2001, 2002, 2003).

Student project
Establish a reasonable set of evaluation measures and apply them to some of the most prominently used algorithms: SPM, ANIMAL, AIR, Rueckert's Image Registration Toolkit, and Thirion's Demons algorithm.

References

Hellier, P, C Barillot, I Corouge, B Gibaud, G Le Goualher, D L Collins, et al. "Retrospective evaluation of intersubject brain registration." *IEEE transactions on medical imaging* 22, no. 9 (September 2003): 1120-30.

Hellier, P., J. Ashburner, I. Corouge, C. Barillot, and K. J Friston. "Inter subject registration of functional and anatomical data using SPM." Edited by T. Dohi and R. Kikinis. *Proc. 5th Medical Image Computing and Computer-Assisted Intervention, Tokyo, Japan, Lecture Notes in Computer Science*, 2002.

Hellier, P., C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, L. Collins, et al. "Retrospective Evaluation of Inter-subject Brain Registration." *Lecture Notes in Computer Science* 2208 (2001): 258-??
http://citeseer.ist.psu.edu/hellier01retrospective.html

Decision fusion and performance bias estimation

The Mindboggle software package (http://www.binarybottle.com/mindboggle.html), used for automating anatomical labeling of human brain MRI data, uses multiple, individual atlases to label a single target brain image. This approach has been determined to be superior to using a probabilistic, average, or interim (propagated) atlas (Heckemann, 2006). However, presently Mindboggle uses a simple majority voting rule to establish a single label per voxel of the target brain image. We will use a more sophisticated multi-label classifier approach to decide on a single label for each voxel. Rather than giving equal weight to the label assigned by each atlas for a given voxel, we will use a weighting scheme, which has been determined to give more accurate results (Warfield, et.al., 2004).

As a single label is elected for each target brain image voxel, the other candidate labels and their weights needn't be thrown away. We want to use these labels and weights to provide a probabilistic labeling of a target brain. The importance of retaining this information lies in one's ability to have confidence measures for labels, to highlight regions or boundaries that should be checked more closely for errors or for morphological deviations from the norm. Low-confidence areas may be corrected through manual editing.

Student project
Implement a multi-label classifier approach to decide on a single label for each voxel. Rather than giving equal weight to the label assigned by each atlas for a given voxel, use a weighting scheme which has been determined to give more accurate results, the STAPLE algorithm (Warfield, et.al., 2004).

-

Heckemann, Rolf A., Joseph V. Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. “Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.” *NeuroImage* 33, no. 1 (October 15, 2006): 115-126.

Warfield, S.K., K.H. Zou, and W.M. Wells. “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.” *Medical Imaging, IEEE Transactions on* 23, no. 7 (2004): 903-921.

Cortical surface representation

There are distinct advantages to performing and viewing the results of brain image segmentation, anatomical labeling, and analysis on surface-based and on volume-based representations of the brain. Transforming data to and from either representation will increase the flexibility in visual inspection, manual editing, and anatomical/functional analysis.

Student project
Evaluate Freesurfer, BrainVisa, and Caret to determine whether one of their surface construction methods should be used with Mindboggle software (http://www.binarybottle.com/mindboggle.html) as its input or as a conversion utility for its output. Otherwise, create a novel method for surface representation of the human brain cortex that can be transformed to a volume-based representation.