Data mining a functional neuroimaging database for functional segregation in brain regions

Finn Årup Nielsen*†‡, Daniela Balslev§, Lars Kai Hansen†*

July 3, 2006

Abstract

We describe a specialized neuroinformatic data mining technique in connection with a meta-analytic functional neuroimaging database: We mine for functional segregation within brain regions by identifying journal articles that report brain activations within the regions and clustering the abstract of the articles using non-negative matrix factorization on the bag-of-words matrix. We divide the brain activations reported in the articles according to the cluster assignment and test for difference between the spatial distribution of the sets of activations. Among our findings is that the memory and pain functions are spatially segregated within the cingulate gyrus.

1 Introduction

Meta-analytic-oriented databases in functional neuroimaging, such as the BrainMap [1] and Brede [2] databases, allow for automated data mining [3, 4, 5, 6]. These databases record so-called Talairach coordinates ("locations") [7] from published human brain mapping studies made with, e.g., positron emission tomography and functional magnetic resonance imaging. The locations represent focal brain activations and are each represented by a 3-dimensional coordinate referenced with respect to a "Talairach" brain atlas [7]. Typically a neuroanatomical term is also associated with the location. Apart from the locations the databases contain description of the experiments in the article that can be correlated to the spatial location information: For the Brede database we have used the words from the abstracts [5] and the linkage to a taxonomy of brain functions [6] as the basis for automated meta-analysis.

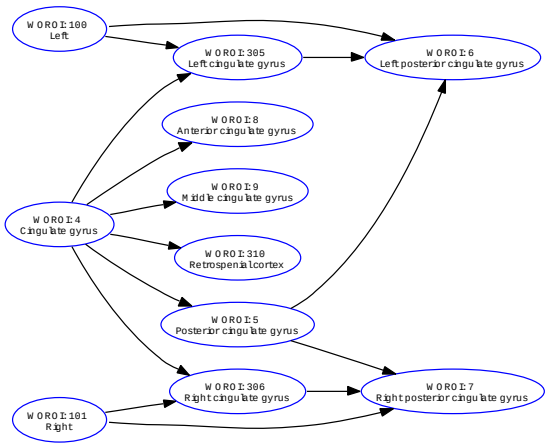

Figure 1: Part of the brain region taxonomy around "cingulate gyrus".

Besides the information from published human brain mapping studies the Brede database has a taxonomy for brain regions, see Fig. 1 for a part of it. It records, e.g., that the "cingulate gyrus" is a subregion of the "cerebral cortex" and that it is a super-region of the "left posterior cingulate gyrus". This hierarchy is partially built from information in the NeuroNames database [8] and the Mai Atlas [9]. It also maintains the variations in the naming, for, e.g., cingulate gyrus they are "cingulate gyri", "gyrus cinguli", "gyrus cingularis" and "cingulate cortex". The taxonomy does probably not capture all relevant variations for many brain regions.

We have previously made a focused data mining on the posterior cingulate brain region using textual data from the PubMed database [10]. In that work we

*Lundbeck Foundation Center for Integrated Molecular Brain Imaging

†Informatics and Mathematical Modelling, Technical University of Denmark

‡Neurobiology Research Unit, Copenhagen University Hospital Rigshospitalet

§Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre

found that memory and pain (processing) were two prominent functions for posterior cingulate, and that their locations were not equally distributed within this area. Below we will make a similar data mining restricting the analysis to data taken from the Brede database, but expanding the data mining to incorporate the many areas defined in the brain region taxonomy.

2 Method

Our method involves a number of steps that each relies on specific information in our database as well as statistical modeling of relations between the items:

1. Robust kernel density modeling in 3-dimensional brain space for identification of Talairach locations of interest
   (a) Select a brain region.
   (b) Get naming variations for the brain region and all its subregions.
   (c) Get locations that matches one or more of the names.
   (d) Model the distribution with kernel density modeling and discard outliers.
   (e) Include locations that did not match any name but lies in the region.
2. Text mining of abstracts
   (a) Get all abstracts that are associated with the locations.
   (b) Construct a bag-of-words matrix from words in the abstract excluding non-important words.
   (c) Cluster the abstracts
3. Robust multivariate test between sets of Talairach locations.
   (a) Extract locations based on cluster assignment.
   (b) Compare the distribution of set of locations.

We use the data from the Brede database [2] which recorded information from 166 journal articles with a total of 3389 locations. The taxonomy of brain regions contained 313 items. Some of these regions are functional and cytoarchitectonic defined areas and these were ignored. For the rest of the areas steps 1–3 listed above are independently carried out. A final fourth step involves the sorting and intertwining of results from all the brain regions.

After selection of a brain region r (step 1a) we obtain the variations of names from the brain region taxonomy (step 1b). This includes variation of names for the brain region itself as well as all its subregions. For, e.g., "cingulate gyrus" this amounted to 48 different names. We query the database for locations where the neuroanatomical name matches any of the variations (step 1c), and obtain a set of L_r 3-dimensional coordinates that can be represented in a matrix $\mathbf{L}(L_r \times 3)$. We model this data with a kernel density estimator and excluded the 5% most extreme locations in terms of probability density to get rid of outliers (step 1d) [3] giving a smaller set of coordinates. We can augment this smaller set by including coordinates associated with high probability density (step 1e), adding the extra locations that did not match any of the variations in the neuroanatomical names. Some initial tests were made with the inclusion of these locations. Often this would lead to inclusion of location with the label of the neighboring region. There are variation in the application of the Talairach atlas: The locations in the so-called MNI-space are converted by a Brett's piecewise affine transformation [11]. This will exclude some of the variation. However, there still is some overlap between regions when locations are collected across studies. The result presented below are did not include this step.

When we have identified all relevant locations for the brain region r we obtain the abstract of all the articles that contain the locations (step 2a). A bag-of-words matrix $\mathbf{X}(N_r \times P)$ is constructed by counting the frequency of each of the P words in the N_r abstracts (step 2b). A very large stop list is used to exclude ordinary stop words as well as words for neuroanatomy and frequent words not associated with human brain function. To avoid that abstracts where some words occur with high frequency will dominate, the element-wise square root $\sqrt{x_{np}}$ is used in the further processing.

For "clustering" the abstract (step 2c) we perform non-negative matrix factorization (NMF) [12] with an algorithm for updating an "Euclidean" cost function [13]. We factorize the bag-of-words matrix into three matrices plus a residual matrix $\mathbf{U}$ whose Frobenius norm is to be minimized

$$\mathbf{WSH} + \mathbf{U} = \mathbf{X}, \tag{1}$$

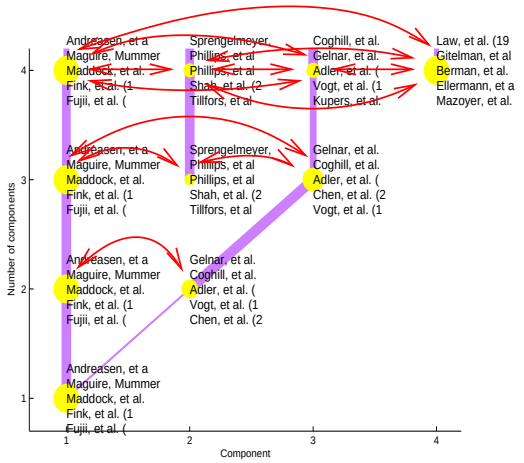

Figure 2: Illustration showing with arrows all the possible comparisons performed between locations in the NMF components (clustered articles). y-axis is size of the NMF K and x-axis the kth component.

where $\mathbf{W}(N_r \times K_r)$ and $\mathbf{H}(K_r \times P_r)$ are non-negative matrices $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$ and normalized [14], e.g., with the vectorial 2-norm $\|\mathbf{w}_k\|_2 = \|\mathbf{h}_k\| = 1$. $\mathbf{S}(K_r \times K_r)$ is a non-negative diagonal matrix. K_r is the size of the subspace, i.e., the number of components/topics. We distribute the scaling contained in $\mathbf{S}$ equally over the two matrices $\mathbf{W}$ and $\mathbf{H}$

$$\tilde{\mathbf{w}}_k = \mathbf{w}_k \sqrt{s_k} \tag{2}$$

$$\tilde{\mathbf{h}}_k = \mathbf{h}_k \sqrt{s_k}. \tag{3}$$

A winner-take-all function is invoked for exclusive assignment of each abstract n to a component k

$$\breve{\mathbf{w}}_{nk} = \begin{cases} \tilde{w}_{nk} & \text{if } \forall k' \neq k : \tilde{w}_{nk} \geq \tilde{w}_{nk'} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

A row vector $\mathbf{h}_k$ in the $\mathbf{H}$ matrix contains loading for words on the kth component. The words associated with the highest load are used to label the component.

We vary K_r between 2 and $\tilde{K}_r = \left\lceil \sqrt{\min(N_r, P_r)} \right\rceil$ and thus generate a set of factorized matrices for each brain region r: $\breve{\mathbf{W}}_{K=2,r} \ldots \breve{\mathbf{W}}_{K=\tilde{K},r}$. Each of the matrices contains an assignment of each of the N_r articles to a specific component/topic, and we construct sets of articles for the kth component in the K-sized NMF for the rth brain region:

$$\mathcal{A}_{k,K,r} = \{n : \breve{w}_{n,k,K,r} > 0\} \tag{5}$$

All locations from these articles are extracted. All combination of two sets of location within each brain region and within each of the K-sized NMF are compared, $\mathcal{A}_{k,K,r} \leftrightarrow \mathcal{A}_{k',K,r}$, e.g., for an NMF with five components ($K = 5$) that gives $K!/(2(K-2)!) = 10$ comparisons. We perform this procedure for all the different sizes of NMF subspaces, see Fig. 2 for an illustration with $\tilde{K} = 4$.

For comparison of the distributions between two sets of locations we perform multivariate statistical tests in 3-dimensional Talairach space. We apply the Hotelling's T^2 test [15] and a Monte Carlo permutation test on the "peeling mean" [15, p. 111–112]. The peeling mean provides a robust estimate of the mode by successively deleting the convex hull layers of the data points and taking the mean of the points associated with the last and innermost convex hull [16], see Fig. 3. The permutation test on the peeling mean is performed by randomizing the locations between the two sets. This test is performed since the Hotelling's T^2 is not reliable with non-Gaussian distributed locations.

The Hotelling's T^2 test is applied in two different ways: The first computes the test statistics from the original two sets of locations and the second computes it by first finding the average within each article and then making the test statistics based on the two sets of averages. The latter way is to ensure that an article containing many coordinates in a specific area will not dominate the test statistics. Neither of the three tests allows us to say in which way two sets of coordinates differ.

We use an Internet search engine reporting style, where results are reported in a sorted list with the most relevant information on the top. The results we will present are ordered according to a conjunction P-value, where the resulting P-value is the maximum across the P-values from the three different statistical

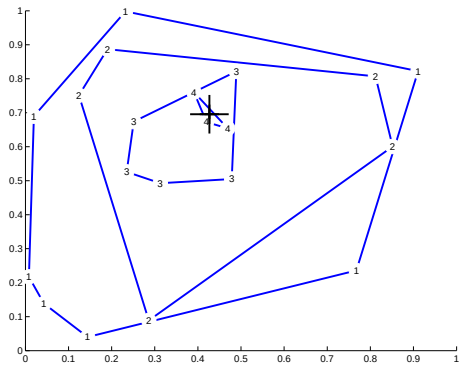

Figure 3: Convex hull peeling mean in 2 dimensions.

#	P-values			(First set) - (Second set) - Brain region
1	0.000	0.000	0.000	(pain, painful, 211) - (visual, eye, 565) - Cerebral Cortex (14)
2	0.000	0.000	0.000	(pain, painful, 230) - (visual, eye, 587) - Telencephalon (13)
3	0.000	0.000	0.002	(pain, painful, 97) - (memory, retrieval, 141) - Cingulate gyrus (4)
4	0.000	0.002	0.003	(pain, painful, 269) - (visual, eye, 607) - Forebrain (12)
5	0.000	0.005	0.000	(expressions, facial, 15) - (recognition, humans, 10) - Amygdala and Hippocampus (202)
6	0.000	0.004	0.005	(memory, retrieval, 22) - (pain, painful, 5) - Anterior cingulate gyrus (8)
7	0.000	0.004	0.005	(memory, retrieval, 22) - (pain, painful, 5) - Posterior medial prefrontal cortex (204)
8	0.000	0.006	0.000	(ear, musical, 5) - (retrieval, faces, 13) - Right frontal lobe (82)
9	0.000	0.000	0.006	(pain, painful, 100) - (memory, retrieval, 159) - Limbic gyrus (125)
10	0.009	0.002	0.000	(memory, episodic, 27) - (motor, sensorimotor, 20) - Cerebellum (32)
11	0.001	0.004	0.011	(artefacts, categorization, 2) - (memory, word, 28) - Precentral gyrus (68)
12	0.000	0.001	0.015	(pain, painful, 71) - (words, memory, 45) - Limbic lobe (2)
13	0.000	0.000	0.016	(pain, painful, 79) - (memory, episodic, 72) - Prefrontal cortex (22)
14	0.000	0.000	0.024	(artefacts, categorization, 7) - (verbal, visual, 16) - Middle frontal gyrus (148)
15	0.000	0.002	0.029	(memory, episodic, 26) - (pain, painful, 5) - Medial prefrontal cortex (55)
16	0.000	0.031	0.002	(musical, ear, 6) - (artefacts, decision, 10) - Right temporal lobe (86)
17	0.002	0.037	0.009	(pain, noxious, 25) - (motor, visual, 20) - Insula (67)
18	0.000	0.042	0.000	(memory, retrieval, 34) - (pain, painful, 25) - Posterior cingulate gyrus (5)
19	0.006	0.006	0.044	(memory, episodic, 15) - (sensory, visual, 6) - Right fusiform gyrus (134)
20	0.000	0.003	0.047	(visual, emotional, 13) - (faces, familiar, 7) - Left superior temporal gyrus (129)
21	0.000	0.049	0.027	(retrieval, memory, 10) - (rest, memory, 6) - Left anterior cingulate gyrus (94)
22	0.000	0.056	0.006	(memory, episodic, 165) - (artefacts, categorization, 24) - Frontal lobe (18)
23	0.000	0.056	0.042	(facial, faces, 12) - (memory, words, 28) - Left cingulate gyrus (305)
24	0.001	0.039	0.063	(ear, musical, 5) - (artefacts, decision, 10) - Right inferior frontal gyrus (296)
25	0.003	0.070	0.027	(recognition, word, 9) - (eye, attention, 15) - Precuneus (171)

Table 1: Automatically generated list of the 25 most relevant functional segregations in brain regions. Columns 2–4 are P-values for the Hotelling's T^2 test with the original coordinates of the locations (column 2) and with the averaged-within-article coordinates used for the test (column 3). Column 4 is the P-value for the peeling permutation test. The words in parentheses are the words associated with the highest load on the components and the number in the parentheses are the number of locations in the component. To the right of the name of the brain region is shown the Brede database identifier for the region.

tests [17].

The data processing uses the Brede toolbox [18], and once the data are entered in the Brede database the entire processing pipeline runs automatically.

3 Results and discussion

Table 1 lists the most relevant functional segregations — one for each brain region. The top entries are in a sense trivial since they indicate a high-level segregation for areas such as "cerebral cortex", "telencephalon" and "forebrain", and the most relevant functional segregation our method reports is between "pain" and "visual". The most frequent words in the abstracts of the Brede database are "visual", "memory", "motor" and "perception", and many of the studies in the Brede database are pain studies. So it is not surprising that we with the Brede database find that major high-level segregation in the brain is between "pain" and "visual". "Pain" locations are mostly distributed in the anterior part of the brain, while "visual" locations have their major share in the posterior part.

Apart from this high-level segregation the most prominent functional segregation appears in the "cingulate gyrus" between pain and memory. Fig. 4 shows the locations for this area colored according to component. A number of subregions and super-region to this area also appear with a segregation between these two functions: "anterior cingulate gyrus", "posterior medial prefrontal cortex", "limbic gyrus" and "posterior cingulate gyrus". Many of the studies that make up these areas were included in connection with our previous study of posterior cingulate [5], where this segregation was identified, and it is thus not surprising that this is refound.

The compound region "amygdala and hippocampus" is segregated into "expressions" and "recognition", and corresponds to a well known functional division in the medial temporal lobe where the hippocampus area is mainly associated with memory whereas the amygdala is involved in the processing of emotional stimuli such as facial expression [20].

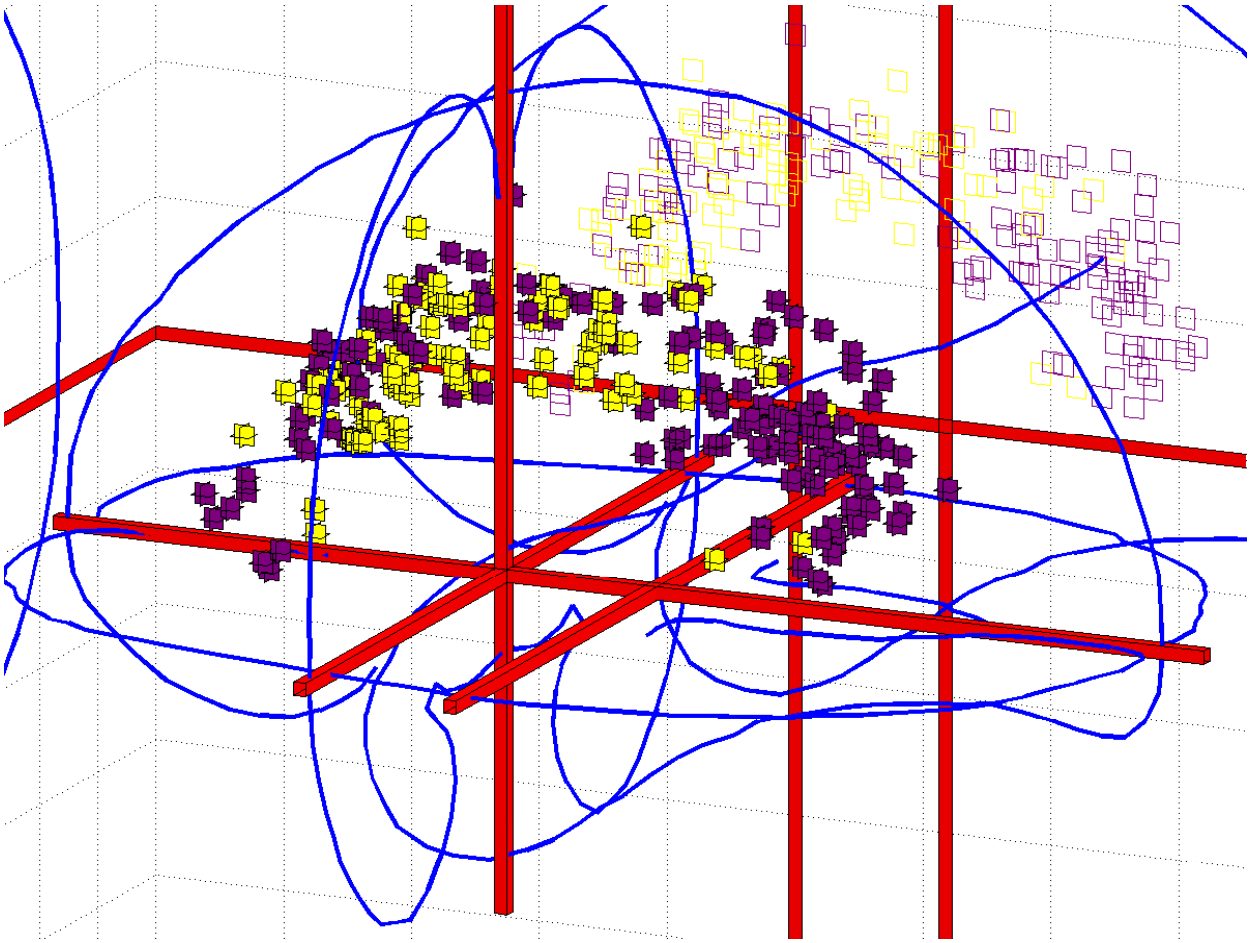

Figure 4: The most relevant functional segregation for "cingulate gyrus" within the Brede database: memory (dark/magenta) and pain (light/yellow). The two sets of locations are plotted in a Corner Cube Environment where each location is represented by a glyph in 3-dimensional space and projected onto "walls" [19] (In this plot only the sagittal projections are visible). The view is from back upper left.

Our method has shortcomings, e.g., some of the results are affected by a number of studies from a single group that investigates artefacts and categorization and reports many activations in specific parts of the brain across articles. Since approximately the same wording is used the abstracts are clustered together, and when the locations from the associated articles are extracted these are spatially clustered often giving rise to segregation when tested against other sets of locations. Furthermore, all the words in a specific article will be modeled together with all locations in that article, e.g., for "cerebellum" a segregation between "memory" and "motor" is found. Actually the "memory" studies have some kind of movement response — overt speech or button pressing — and this is probably why the memory studies activate in cerebellum.

4 Conclusion

We have devised a method that mines a neuroimaging database to extract the main functional modules within a brain region. Such a tool would allow the individual researcher to access the growing base of knowledge generated by the functional imaging studies.

References

[1] Peter T. Fox and Jack L. Lancaster. Neuroscience on the net. *Science*, 266(5187):994–996, November 1994.

[2] Finn Årup Nielsen. The Brede database: a small database for functional neuroimaging. *NeuroImage*, 19(2), June 2003. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY. Available on CD-Rom.

[3] Finn Årup Nielsen and Lars Kai Hansen. Modeling of activation data in the BrainMap^{TM} database: Detection of outliers. *Human Brain Mapping*, 15(3):146–156, March 2002.

[4] Finn Årup Nielsen and Lars Kai Hansen. Finding related functional neuroimaging volumes. *Artificial Intelligence in Medicine*, 30(2):141–151, February 2004.

[5] Finn Årup Nielsen, Lars Kai Hansen, and Daniela Balslev. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics*, 2(4):369–380, Winter 2004.

[6] Finn Årup Nielsen. Mass meta-analysis in Talairach space. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 985–992, Cambridge, MA, 2005. MIT Press.

[7] Jean Talairach and Pierre Tournoux. *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publisher Inc, New York, January 1988.

[8] Douglas M. Bowden and Richard F. Martin. NeuroNames brain hierarchy. *NeuroImage*, 2(1):63–84, 1995.

[9] Jürgen K. Mai, Joseph Assheuer, and George Paxinos. *Atlas of the Human Brain*. Academic Press, San Diego, California, 1997.

[10] Finn Årup Nielsen, Daniela Balslev, and Lars Kai Hansen. Mining the posterior cingulate: Segregation between memory and pain component. *NeuroImage*, 27(3):520–532, 2005.

[11] Matthew Brett. The MNI brain and the Talairach atlas. http://www.mrc-cbu.cam.ac.uk-/Imaging/Common/mnispace.shtml, February 2002. Accessed 2005 April 9.

[12] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[13] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 556–562, Cambridge, Massachusetts, 2001. MIT Press.

[14] Wie Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA, 2003. ACM Press.

[15] Kantilal Vardichand Mardia, John T. Kent, and John M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London, 1979.

[16] C. Bradford Barber, David P. Dobkin, and Hannu T. Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, December 1996.

[17] Matthrew Brett, Tom Nichols, Jesper Andersson, Tor Wagner, and Jean-Baptiste Poline. When is a conjunction not a conjunction? *NeuroImage*, 22, 2004. Presented at the 10th Annual Meeting of the Organization for Human Brain Mapping, June 14–17, 2004, Budapest, Hungary. Available on CD-ROM.

[18] Finn Årup Nielsen and Lars Kai Hansen. Experiences with Matlab and VRML in functional neuroimaging visualizations. In Scott Klasky and Steve Thorpe, editors, *VDE2000 - Visualization Development Environments, Workshop Proceedings, Princeton, New Jersey, USA, April 27–28, 2000*, pages 76–81, Princeton, New Jersey, April 2000. Princeton Plasma Physics Laboratory.

[19] Kelly Rehm, Kamakshi Lakshminarayan, Sally A. Frutiger, Kirt A. Schaper, De Witt L. Sumners, Stephen C. Strother, Jon R. Anderson, and David A. Rottenberg. A symbolic environment for visualizing activated foci in

functional neuroimaging datasets. *Medical Image Analysis*, 2(3):215–226, September 1998.

[20] Ian Q. Whishaw and Bryan Kolb. *Fundamentals of human neuropsychology.* Worth Publishers, 4th edition, July 1995.