Brain Struct Funct
DOI 10.1007/s00429-007-0160-2

ORIGINAL ARTICLE

Bottom-up and top-down brain functional connectivity underlying comprehension of everyday visual action

S. J. Hanson · C. Hanson · Y. Halchenko · T. Matsuka · A. Zaimi

Received: 3 March 2007 / Accepted: 5 September 2007

Abstract How can the components of visual comprehension be characterized as brain activity? Making sense of a dynamic visual world involves perceiving streams of activity as discrete units such as *eating breakfast* or *walking the dog*. In order to parse activity into distinct events, the brain relies on both the perceptual (bottom-up) data available in the stimulus as well as on expectations about the course of the activity based on previous experience with, or knowledge about, similar types of activity (top-down data). Using fMRI, we examined the contribution of bottom-up and top-down processing to the comprehension of action streams by contrasting familiar action sequences with those having exactly the same perceptual detection and motor responses (yoked control), but no visual action familiarity. New methods incorporating structural equation modeling of the data yielded distinct patterns of interactivity among brain areas as a function of the degree to which bottom-up and top-down data were available.

Introduction

A fundamental question in visual neuroscience is how information becomes integrated across visual, memory, and perceptual pathways to produce understanding of everyday action sequences. Previous work shows that in order to comprehend the meaning of action streams, people naturally and spontaneously parse activity into distinct units of meaning (e.g., eating dinner, going to the movies, etc). Moreover, observers who are asked to categorize action sequences in real time produce consistent judgments about action cluster start and end points (Hanson and Hirst 1989; Newtson 1973; Hanson and Hanson 1996; Heider and Simmel 1944; Zacks and Tversky 2001). Although parsing video action sequences into constituent structure is a complex visual task, it seems clear that at a basic level it involves a perceptual detection component that is dependent on comprehension of the action sequence and expectancies over time. Thus, the study of how people understand action sequences offers a unique opportunity to investigate how visual pathways in the brain that underlie cognitive processes such as feature encoding, motion detection, visual-spatial integration, visual search, visual attention, and working memory interact.

In a series of experiments we contrast subjects' viewing of three different kinds of actions. The first what we will call schema-rich sequences which consist of highly familiar action sequences, like "making coffee and drinking it", or "assembling a chair from a box" etc. (see typical movie frames of two movies in Fig. 1). There were three different schema-rich sequences: (1) a person assembling a chair with parts taken from a box, (2) a person making coffee and then sitting and drinking it and (3) a person coming into a room and using a computer while studying from a book. The second class of action sequences we call schema-free (see Fig. 2). These consisted of an oddball task in which a single rectangle jittered up and down randomly with fixed mean excursion from a baseline and fixed standard deviation. At seemingly random times the rectangle would jump 2 s.d. higher then at any other time point indicating a significant change point. These points were not random, but rather yoked to the response time points detected while viewing schema-rich videos and a third type

S. J. Hanson (✉) · C. Hanson · Y. Halchenko · T. Matsuka · A. Zaimi
RUMBA Laboratories, Psychology Department, Rutgers University, Newark, NJ 07102, USA
e-mail: jose@psychology.rutgers.edu

of action sequence (schema-poor), in which cartoon objects (circles, squares etc.) made arbitrary movements. By yoking the schema-free stimulus to response time points obtained for the schema-rich stimulus, we were able to control for simple perceptual or motor aspects of the schema-rich task. Consequently, the contrast between schema-rich and schema-free provides an index of the "top-down" neural correlates of cognitive control and visual spatial attention. Although not all low level visual processing (eye movements, spatial frequency, texture, motion etc.) are controlled for, our main goal in this study was differentiate between a simple detection task and an event parsing task with the same response time/detection demands. Moreover, we attempted to minimize any visual content effects by aggregating over the three different video action sequences that were yoked to their appropriate time point schema-free controls (see Fig. 2 for specific procedures).

Thus our use of "bottom-up" explicitly focuses on the difference between a visual expectancy based detection (independent of visual featural details) and a vigilance type process invoked by tasks similar to an oddball task. This type of contrast reflected our interest in dorsal and ventral visual streams and how they are modulated by prefrontal and parietal areas of the brain. For example attentional modulation is known to occur between parietal, and extrastriate areas in the monkey brain (Desimone and Duncan 1995; Kastner and Ungerleider 2000). We might therefore expect to see a human analogue to MT (MT+) and STS as well as parietal and prefrontal areas during an event perception task (Zacks et al. 2001; Hanson et al. 2001). Similarly, oddball tasks typically involve interactions between known parietal and visual areas as well as prefrontal areas such as anterior cingulate cortex that, according to some theories, may index change-point events (Posner and Gilbert 1999; Miller and Cohen 2001). Hence, despite the obvious differences in visual simulation between an event perception task and a time/response yoked oddball task, our interest lies in the functional overlap of the perceptual detection component, which we would argue is identical to that in an event perception change point detection task. In Fig. 3, we provide more detail about the theoretical processing stages of the event perception task and the specific functions we are associating with "top-down" or "bottom-up" processing. Note that in this figure we are showing a time line in which occasional event change points (squiggle in each parenthetical) are broken out while subjects are watching the familiar action sequence. Any given event change point clearly consists of many complex visual processing elements including low level visual processing, saccadic eye motion, and detailed featural content (e.g., "the man has a beard"). We argue that such an event change point is the key to differentiating the event perception task and the yoked oddball task in terms of "temporal" detection points. Clearly, in the event perception task these detection points are driven substantially by expectancies rather than attentional vigilance. We propose that these expectancies comprise the "top down" influence that can be isolated to contrast these two kinds of tasks.

This study is consistent with past research that implicates top-down and bottom-up control in visual spatial attention, language processing, and cognitive control functions as well as breaking new ground by using fMRI bold data to do exploratory graphical modeling of the recruitment and interaction of context sensitive brain functions during top-down and bottom-up processing. We will describe first the event perception task and relevant behavioral data collected from subjects viewing one of three different videos of familiar action sequences. After describing the basic methods and behavioral results, we will turn to the fMRI data and computational modeling.

Methods for behavioral data collection

Event perception

We used scripted videos, similar to other studies using naturalistic movie viewing (e.g., Hausson et al. 2004), where actors were given very simple and general instructions, such as, "make a pot of coffee" or "open this box and assemble the chair inside". The videos were continuously filmed from a single camera angle throughout the action sequence (this is in contrast to commercial movies which often involve multiple camera angles, action and

Fig. 1 Frames from two different action sequences used as stimuli in event perception (*left*) the "chair" video, showing a person putting a chair together, (*right*) the "coffee" video showing a person making and drinking a cup of coffee (each frame shows a >0.9 event change point)

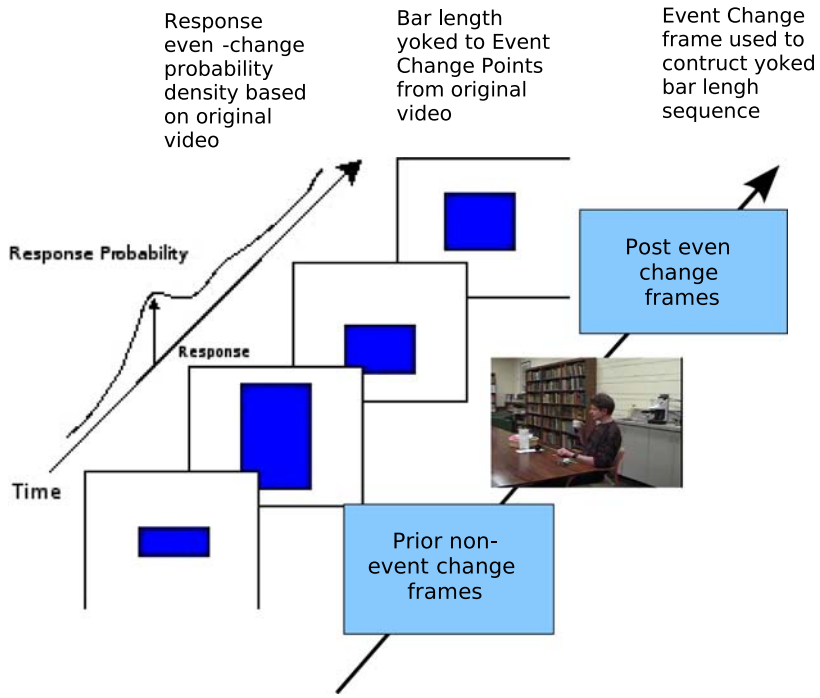


Fig. 2 Construction of the Yoked control sequence based on the TRD (temporal response density or probability), points in the action sequence where subjects had high common agreement of an event change point (single frame for reference-not actually shown with *rectangles*) was associated with a increase in bar length that was greater than 2 standard deviations within a distribution of bar lengths randomly sampled from a gaussian distribution with mean 10 mm and standard deviation 2 mm

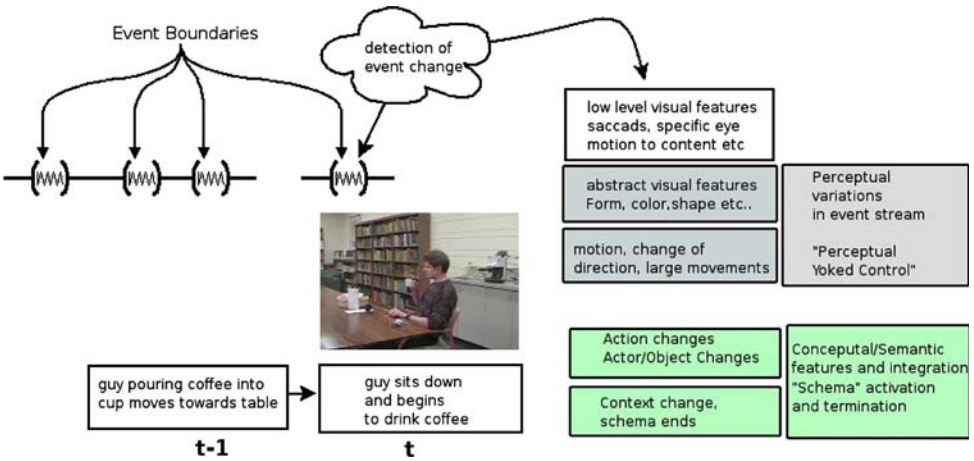


Fig. 3 Hypothetical processing of event change detection shown in time from one frame prior to an event change point and the event change point, that invokes various perceptual and cognitive processes which we roughly segregate into "bottom-up" (*gray boxes*) and "top-down" (*green boxes*), the yoked control is designed to only control for more abstract visual feature, motion and detection of change of state as in an "oddball task". We are explicitly not controlling for specific eye motion or low level featural processing that may be related to detailed content with the yoked control (*white box*). Moreover, in order to focus on brain tissue that is associated with more abstract visual features that elicit detection of event change, we concatenated all three video sequences (study, chair, coffee) in our GLM analysis—see Fig. 8

conceptual cuts). All videos were less than 5 min long and consisted of a single actor interacting with relevant objects. Example frames from two of the scripted videos are shown above in Fig. 1, A third video, of a student entering a room and working on laptop, which we called the "study" video, was also used.

Subjects were asked to indicate "significant event changes" by pressing a button while they were watching the action sequences. Subjects typically find this task natural and immediately begin parsing (viewing and button pressing) once the video begins, usually without any further clarification of the instructions.

Results: basic event perception experiments

A convenient way to visualize the subject agreement over time is based on a group-wise analysis of temporal

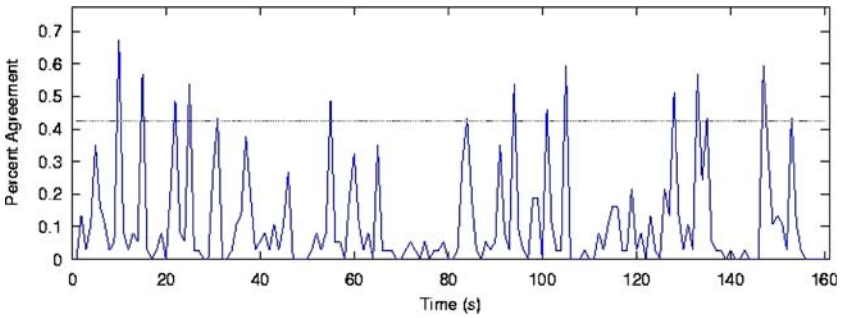


Fig. 4 Button press responses from 25 subjects while observing the "study" action video accumulated in one second bins resulting in a temporal response density (*TRD*) indexing an instantaneous estimate of change point in the action sequence. The *horizontal dashed line* indicates the 95% confidence interval for the change point estimate which when thresholded produces 17 change points. For individuals each TRD is a binary sequence of 1 s time points, which is convolved with the HRF and used as an event change point regressor for the GLM (see Fig. 8)

response points from all subjects ($N = 26$), where their responses are binned (one second resolution). This procedure results in a response probability density over time, what we will refer to hereafter as a temporal response density (TRD) and is shown in Fig. 4. In Fig. 4, note that the horizontal line shows the 99% confidence interval for subject agreement event change points over the whole action sequence. Event change points tend to involve rapid body movements ("standing up", "large arm movements" etc.) as well as conceptually dependent initiation (e.g., "opening a box'", "filling coffee pot with water' etc.), which are not necessarily correlated with change in body state.

Surprisingly, viewings of different scripted videos results in very similar parsing rates for each subject as evidenced by high correlation of parsing rates between two different action sequences (see Fig. 5; experiments 1 and 2 repeat viewings of the same video). Thus, although subjects can adjust their parsing rates, they seem to have a *preferred* rate of parsing visual action sequences. Shown in Fig. 7 is the parsing rate of for one scripted video and one cartoon video with a correlation of 0.86 over 26 subjects (each point in the graph is a subject parsing rate), although the autocorrelation of their temporal point responses (see Fig. 6) is not significant (horizontal line is 95% confidence limit). These data indicate that the visual features of the action sequence exert a strong primary control over parsing behavior as opposed to some sort periodic (internal) response based control. Behaviorally, therefore, subjects have reliable parsing rates that mirror the event structure of the action sequence and will, when queried, accurately summarize the action sequence (e.g., ".. the guy was making coffee").

Results: yoked control for bottom up

As subjects watched each video, their parsing responses in *real-time* were used to construct a schema-free video which in the visual field jittered up and down according to Gaussian sampled noise with fixed standard deviation. At each of the subject's response time points in the scripted video, the bar jittered 2 s.d above that of the background jittering (see Fig. 2). After viewing the scripted video

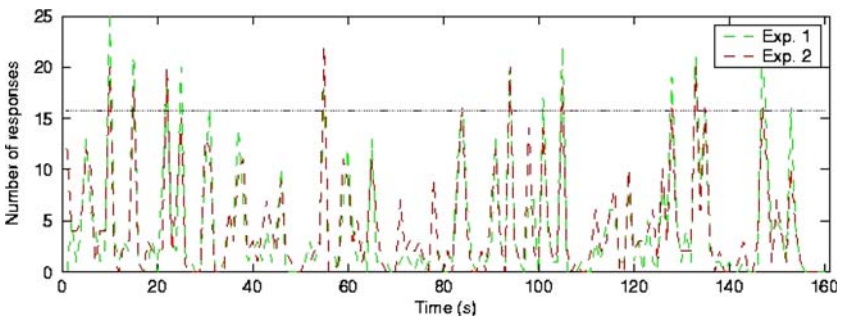


Fig. 5 Button presses collected in one second bins while observes watched the "House" video. Similar to Fig. 4, these are response time densities indicating the instantaneous estimate of event change point. There are two TRDs in the figure, each one composed of the same 25 subjects indicated event change points in two different sessions. The *horizontal dashed line* indicated the 95% significance level from which can be inferred 14 change points. Note the strong overlap which represents a high correlation ($r = 0.93$) implying that the estimate is highly stable

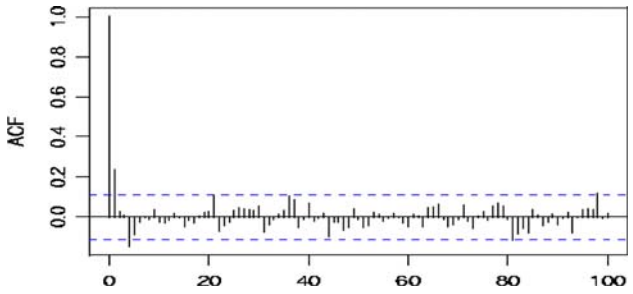


Fig. 6 Autocorrelation of time series in Fig. 4, the blue dash lines are the 95% confidence interval for correlations, this type of profile is typical of a random sequence indicating that the time series alone and consequently independent of the action sequence cues has no simple internal structure

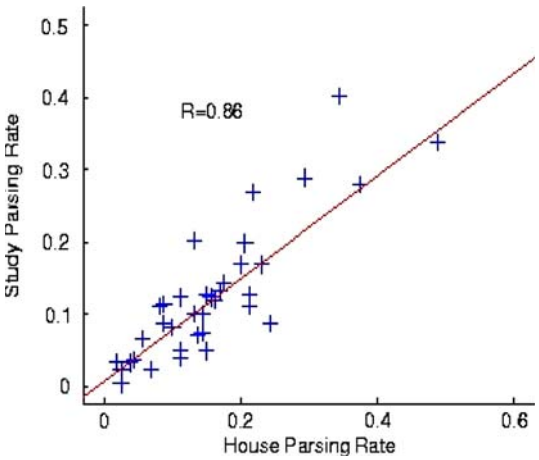


Fig. 7 Correlation of parsing rates for 26 subjects in both the study action sequence and the house action sequence. Note the tendency for subjects to have a invariant parsing rate over time, note strong linear relationship over an order of magnitude over parsing rates

subjects waited 2 min and then were asked to view the schema-free video. Subjects, would be told to press a button when they detected a significant change in the video sequence and were provided no information about the relationship between the first video and the schema-free video. None of the subjects noticed the relationship between the jittering of the bar in the schema-free video and their own responses during the scripted video. This "yoked" condition therefore controlled for the button pressing, and "bottom-up" attentional and cognitive control processing associated with event change detection in the scripted video. As was found for the scripted videos, TRDs showed significant change points but with lower subject agreement. In fact, the correlation between the scripted video and its yoked perceptual counterpart dropped to near zero ($r = 0.07$; independent of the change point button presses, which with were detected with high probability, .85). It is not surprising, therefore, that the pattern of button presses around change points in the schema-rich and schema-free videos were significantly different, despite the high probability of detection of the *specific* change points. Consequently, removing top-down cues changes the pattern of change point detection and TRDs, and indicates that action sequence content is critical for invoking the natural individual parsing rates and patterns.

Neuroimaging methods

We replicated these experiments while collecting fMRI from 32 subjects across the three video action sequences and their yoked perceptual sequences as previously described. Imaging was performed with a 3T Siemens Allegra head only MRI scanner (Erlangen, Germany). We used a 3D magnetization prepared rapid acquisition gradient recalled echo (MP-RAGE) T1-weighted scanning sequence with 2 mm isotropic resolution to acquire structural images for each participant. A T2*-weighted asymmetric spin-echo, echo planar sequence with flip angle of 90 and a 30 ms time to echo (TE) was used to acquire functional data. There were 32, 4 mm slices with each slice consisting of 3.75×3.75 mm cells in 64×64 grid were acquired in whole volumes. The time to repetition (TR) for each volume was 2 s. All data analysis was done using FSL (Smith et al. 2001). MP-RAGE for all subjects were registered to standard atlas using FSL's FLIRT subprogram to the Montreal Neurological Institute (MNI) atlas template. Data were realigned and detrended using the standard FSL FLIRT tool. All localization was done in the Talairach and Torneax (1988) atlas using appropriate affine transformations from MNI registered T1 and BOLD images.

GLM analysis and node clustering

In order to increase the detection of tissue that is more independent of detailed action sequence content we also aggregated over different video content. Consequently, we had three conditions consisting of three separate schema-rich videos with their corresponding yoked control schema-free videos, and two schema-poor videos (these highly stylized and consisted of similar cartoon sequences in which various geometric shapes moved around the display). Initial preprocessing steps included performing an event-related GLM, clustering in the subsequent brain maps to detect potential candidate node clusters which were further filtered by subject agreement (>50% of subject agreement in order to retain a cluster point), and temporal coherence within the cluster assessed by eigenvector analysis (see below and Fig. 9). Specifically, GLM analysis was performed on the fMRI using individual subjects responses (binary time series) weighted with the

group-wise TRD (continuous probability density), thus producing conservative estimate of the event change judgments during action sequence viewing. Each weighted TRD was further convolved with the HRF (hemodynamic response function) and then used as a continuous regressor (these would look similar to the TRDs in Figs. 4, and 5 except they would be phase-shifted and smoothed) in a type of event-related GLM design. These preprocessing steps produced Z-maps (see Fig. 8) for each subject that were then submitted to a mode density clustering method that jointly clustered over both spatial maps and subjects. Crucial to the detection of graphical structure is the initial detection of contiguous dense clusters of voxels that possess similar covariance with other dense ROIs.

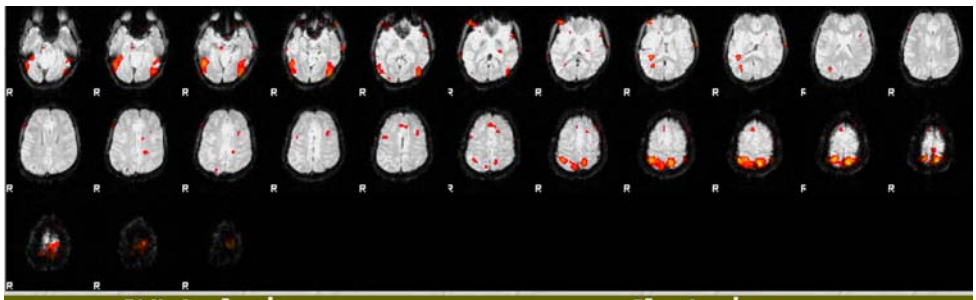
At the same time, we constrained the mode seeking algorithm to detect those voxels across subjects that were in similar spatial coordinates (all subjects are registered in MNI space). This mode density clustering ensures that we extract time series from each area that contains highly similar time series, and while the standard ROI clustering might produce similar results, have both less temporal coherence and less density spatially and therefore less structure to model. More details and benchmarks about the Dense Mode Clustering algorithm can be found in Hanson et al. (2007a).

Results GLM event related analysis

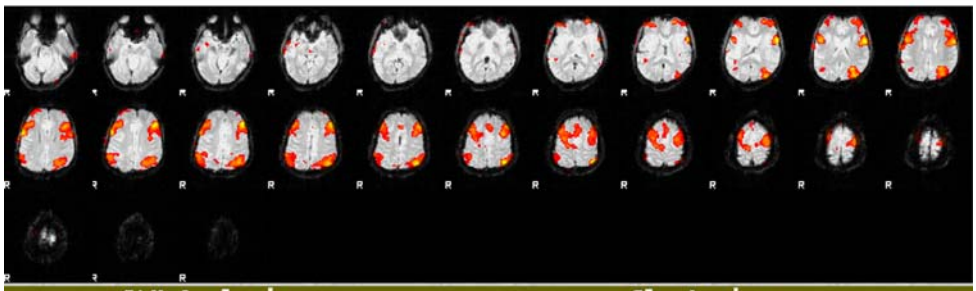
The results for all subjects across all scripted conditions are shown in Fig. 8a, b. Significant areas (shown in red by condition) were only included if both the z score was significant (<0.01) and subject agreement exceeded at least half the subjects (>50%, see Fig. 8). These consistent areas are shown in the tables below for each of the group-wise brain maps. Note in this analysis that there are both common and distinct areas between the schema-rich condition (Fig. 8a) and the schema-poor condition could have focused on only common areas as we constructed the graphical analysis, which might have made it easier to compare across conditions, clustering finds distinct areas that should be treated as a variable in graphical analysis, rather than being ignored. We tended to bias towards the most common graphs nodes/constituents we could and to construct graphs based on all the criterion based voxel clusters. Common areas between the two conditions included inferior parietal lobule (the IPL cluster included some voxels in angular gyrus and supramarginal gyrus), superior temporal sulcus and parts of the gyrus (cluster including STS/STG) and medial frontal gyrus (MedFG). Distinct areas that characterized the schema-rich condition included middle temporal gyrus (MTG or what is usually considered MT+) and middle frontal gyrus (MFG including DLPFC) while the cingulate gyrus (anterior to middle) is unique to the yoked perceptual (schema-free) condition. Thus, the main difference between the scripted tasks that tended to invoke more "top-down" or schema-rich expectations and the yoked perceptual detection (schema-free) task were MTG (MT+) and MFG, areas associated with motion detection and working memory/storage functions. Less surprising perhaps, was the unique area that characterized schema-free condition was the anterior cingulate (ACC), which often is associated with novelty or stimulus change point detection (such as in oddball tasks which the schema-free case closely resembles). Given that common areas are recruited to these two very different tasks and that unique areas appear that are specific to each task, it seemed more productive to ask how these areas may be interacting. In effect, we sought to characterize how networks of these identified brain areas interact in order to modulate top-down influence in this type of visual comprehension task (e.g., McIntosh 1999, 2001).

Brain interactivity: graph modeling

Different strategies have been used recently to fit graphical models to data. Each of these approaches have focused on different aspects of the time series data and priors on the model. More critically, they have varied in the assumptions underlying the kinds of parameters that can be estimated about connectivity of brain regions. Friston et al. (2003) and Penny et al. (2004) have focused on dynamic causal modeling (DCM), a means of estimating the specific coupled parameters in a bilinear dynamical system that are a function of hemodynamic (based on an extended balloon model, Buxton et al. 1998) and neurodynamic state variables. The "effective connectivity" in DCM is based on three types of connections: (1) the "intrinsic connections" which specify regional connections and their direction, (2) a set of connections from exogenous inputs with connectivity to specified regions, and (3) connections which define which of the regional connections can be modified by which exogenous inputs. These sets of connectivity parameters and their specification (estimation) comprise the basic assumptions about the model structure. This is an ambitious model in that it attempts to estimate node (region of interests chosen by an initial GLM analysis by the user) "causal influence" based on these three types of connection estimates while maintaining a stable dynamical system. Fundamental to this type of dynamic causal model is the expectation that there is a single best model given the coupled time series that can be selected from $N\hat{}m$ possible cyclic or acyclic (where N equals edges and m are the number of nodes) models underlying the temporal covariations between ROIs. Although this approach is both



GLM Analysis						Clustering		
Brodman Area	Voxel tot.	Z	x	y	z	%Agreement	Mean	SD
PreCun	228	3.1	2	-70	35	21	55	27
SFG	149	2.6	-13	10	56	21	32	25
IPL	89	4.01	60	-22	29	76	4.2	5
Pre/Post	53	4.1	-62	26	38	88	5.2	4.3
STS	30	3.85	53	5	-1	77	5.7	4.9
MT	22	3.88	-59	-68	15	84	62	46
MFG	13	5.10	33	-5	53	69	3.2	2.5
IFG	13	3.7	-61	7	18	23	1.2	0.2
MedFG	11	3.9	-2	5	43	72	11	7.8



GLM Analysis						Clustering		
Brodman Area	Voxel tot.	Z	x	y	z	%Agreement	Mean	SD
Pre/Post	880	4.1	-62	26	38	98	15	20
MFG	627	5.1	33	-5	53	44	24	12
IPL	214	4.1	60	-22	29	100	24	20
SFG	169	2.6	-43	35	31	21	32	12
IFG	157	3.7	-61	7	18	23	12	22
STS	37	3.9	53	5	-1	83	20	15
Insula	20	3.8	44	-2	10	22	8	4
Cing	17	3.9	-2	5	43	92	27	14
MedFG	16	4.7	9	-9	52	74	55	40

Fig. 8 GLM analysis showing all subjects and videos for (*upper*) schema rich action sequences and (*below*) for schema poor videos. Clustering both within brain maps and across subjects identified areas that were common to at least >0.5 of all subjects with same cluster centroids

impressive in its scope and complexity, it requires relevant priors and penalties to reduce the risks of divergence in what is already a huge model search space (see below). Without strong priors on such models it is very likely that estimation algorithms may degenerate as the number of nodes increase beyond 5 or 6. Nonetheless, DCM is one of the most widely used graphical modeling approaches, at least partly due to its access in a popular analysis package (e.g., SPM). At the opposite end of this spectrum are the graph approaches by McIntosh (1999, 2001), which start with a user specified model. In this approach a confirmatory strategy is taken which depends on either an existing theory about how areas may be interacting or constraints arising from plausible anatomical pathways. In these models graphs are fit using partial least squares or factor analysis, and confirmed by "failing to reject" the graph as a likely model. This provides plausible confidence in the fitted graph as not being inconsistent with covariances and variances of the data. The method we propose could be considered to be somewhere in between these two approaches to graph estimation. Specifically, on oriented (directed) labeled graphs and on the similarity of covariances (as in the McIntosh confirmatory approach), rather than the dynamics of time series. However, we will also be interested in exploratory analysis, that may identify *novel* graph influence. Two problems immediately arise. One problem involves the size of graph space as we model more ROIs and the natural increase in the number of Markov equivalent graphs with the increasing graph space. We have shown (Hanson et al. 2004b) that graph space

increases as 3^((*n* − 1) × *n*)/2) which for even 5 ROIs (equal to "*n*" in the equation before) is well over 60 k graphs for 8 ROIs nearly 10 billion! Without strong priors on graph construction this will clearly be an intractable search problem. One prior that does help is that there are few brain area network type theories in cognitive neuroscience that are likely to involve more than seven brain areas at present. Even discussion of the so-called "mirror system" which tends to be very distributed in humans often involves no more than four to six areas (MT+, STS, pre-Motor, Inferior Frontal Gyrus, IPS/IPL) and typically two or three areas at a time. This limit on the number of relevant brain areas can be used strategically to limit the data structure of interest and make the search in these smaller graph spaces computationally possible. Across many such tasks if the same network of regions area engaged, then the difference between them can only be seen in their potential interconnectivity and influence. Our approach therefore is only slightly stronger than calculating the thresholded inter-correlations in that we are extracting a directed graph from the reconstruction of the covariance from the model. The other problem involves the fact that different graphs will fit the same data with nearly the same goodness of fit. This occurs when the different graphs have the same implications for the partial correlations existing in the data. Consequently, in order to get a data-driven measure of directionality, we prefer a voting procedure to estimate the presence of an edge and then the overall directionality of the edge. All GOF (goodness of fit) equivalent graphs (within some Epsilon) will thereby provide a weight of evidence proportional to their presence in the equivalence class. This provides a "best graph" in the sense of all possible graphs that might have fit the covariance data and an unbiased estimate of edges and influences between ROIs. We have used simple Chi-square methods, but prefer Akaike (AIC), which tradeoffs between degrees of freedom in the model and the graph fit.

The graph fitting procedure and methods

Since the validity of the graph identified is critically dependent on the ROIs or nodes identified, some commitment to the node identification must be made in order to assure they are stable, temporally coherent, and roughly similar in size (from sampling error arguments). Some methods (dynamic causal modeling) put the ROI identification in the hands of the users. Unfortunately, without guidance, an experimenter could pick the same ROIs from each condition, even though, ROIs might consist of roughly 50 voxels in one condition and one voxel in the other. (Even though it would be just as easy for the user to pick the same number of voxels in both conditions, it would not guarantee the similarity of their spatial density). This comparison could produce different graphs, but not necessarily on the basis of the condition difference since the temporal coherence, sparseness of the ROI, and volume all could contribute to any observed differences in graphs. One apparent difficulty is that the present method could estimate different nodes across the graphs in each condition making the comparisons noncommensurate. We think this misses the point. Although one could force the same "nodes" for each graph, this would have the effect of reducing the validity of each graph in each condition. This is because any constraint that gives preference to ROIs that have lower probability of appearing in the graph in the first place introduces uncertainty in the graph estimate itself. Moreover, if the conditions are actually minimally different, then the same or "related" areas should appear in the analysis as stable and predictive. Can valid comparisons be made across such graphs? If the same nodes do not appear across conditions, there are instabilities in the presence or absence of a node, which would seem to be diagnostic of those conditions. On the other hand, if the same nodes do appear then estimates of causal influence are still valid in the context of the other nodes and can be scaled (e.g., normalized) in that graph context. Of course, absolute influence estimates would not be valid, unless the graph edge estimation across conditions are made in some more global way. Nonetheless, the relative influence of any edge within a given graph is meaningful within that context and provides some evidence for the modulation of that influence between regions across conditions. So, for example, if in condition 1 area A influences area B twice as much as area C, then this relative comparison is valid with respect to condition 2 where such influence might be absent. Thus ordinal comparisons among conditions for the same areas are clearly interpretable, even if interval scale information about graph influence across conditions is not. Finally, in the present method one can also include a prior on the same areas appearing across graphs (perhaps dropping those that are not common prior to graph estimation) that are also the most spatially dense. This of course, becomes an empirical issue that should not be resolved by model-based assumptions or the convenience of absolute comparisons, but rather the actual data-structure's validity in terms of presence of nodes and their estimated influences.

Consequently, we invest in a new method for clustering that finds dense modes in brain maps that also have high agreement across subjects. This new clustering method (Hanson et al. 2007a), what we call dense mode clustering (DMC), does 3-dimensional density estimation in brain maps and iteratively searches through multi-scale spatial filters, the most dense ROIs that are also most sparse in the overall brain space. Simultaneously we identify over all subjects the supra-thresholded common (>50%) clusters

that according to DMC are also dense regions. Hence these ROIs are both the most dense earlier, and are also found in most subjects. We think this is the minimum criterion for picking candidate nodes, although one could provide more requirements concerning the time series or even their underlying dynamics. The time series from all voxels per region were concatenated (as opposed to averaged) over all subjects (creating a "super subject"). In this way, a time series across each subject for each video could contribute in an individual way to the underlying patterns represented as a principle component. All scripted (schema-rich) action sequences (3 videos) were submitted to an SVD (singular value decomposition) to extract a common time series based on the minimum number of eigenvectors to reconstruct at least 80% of the original time series variance. This was done for each ROI (five in the case of scripted action sequences) over concatenated variances. A similar procedure (see Fig. 9 for general data flow and pipeline analysis) was used for the perceptual (schema-free) tasks over all subjects for each ROI (four in the case of the yoked perceptual detection task).

A covariance measure across all ROIs in the scripted and perceptual conditions were computed over all time points (for the "super subject" 780 time points-schema-rich videos were 305, 260, 225 s with a TR of 2 s). Covariance matrices for each subject over all action sequences and perceptual videos were also computed for subsequent cross-validation.

Graph fitting results-schema-rich versus schema-free

Each covariance matrix was submitted to structural equation modeling (LISREL). All possible graphs ($N = 5$, 58,024, $N = 4$, 664) were fit and sorted by goodness of fit (e.g., Akaike, although "P" value and BIC were compared and typically consistent) and graph averaged over the top 5%. This produced a "best graph" in terms of connectivity and direction of edges. In Fig. 10 we show brain interactivity estimates from this new method (Hanson et al. 2007b), contrasting the scripted (schema-rich) and yoked perceptual (schema-free).

In order to test the generality of the graph structures shown, we cross-validated the "super subject" graph to each individual subject covariance in the same condition producing excellent agreement (>90%) for each individual subject covariance (note that the group covariance, which was not an average, is not required to reproduce the individual graphs due to individual differences etc. Each edge of the graph shows "influence" (in a regression sense) of one area upon another during action sequence parsing. Given the critical contrast between scripted (schema-rich) and perceptual (schema-free) (i.e., "top-down"/"bottom-up" contrast) we propose that the estimated schema-rich graph represents two kinds of schema influences during parsing. First is a schema activation sub-network consisting of areas (IPL/IPS, MFG and MedFG) that we argue primarily focused on retrieving, activating and comparing

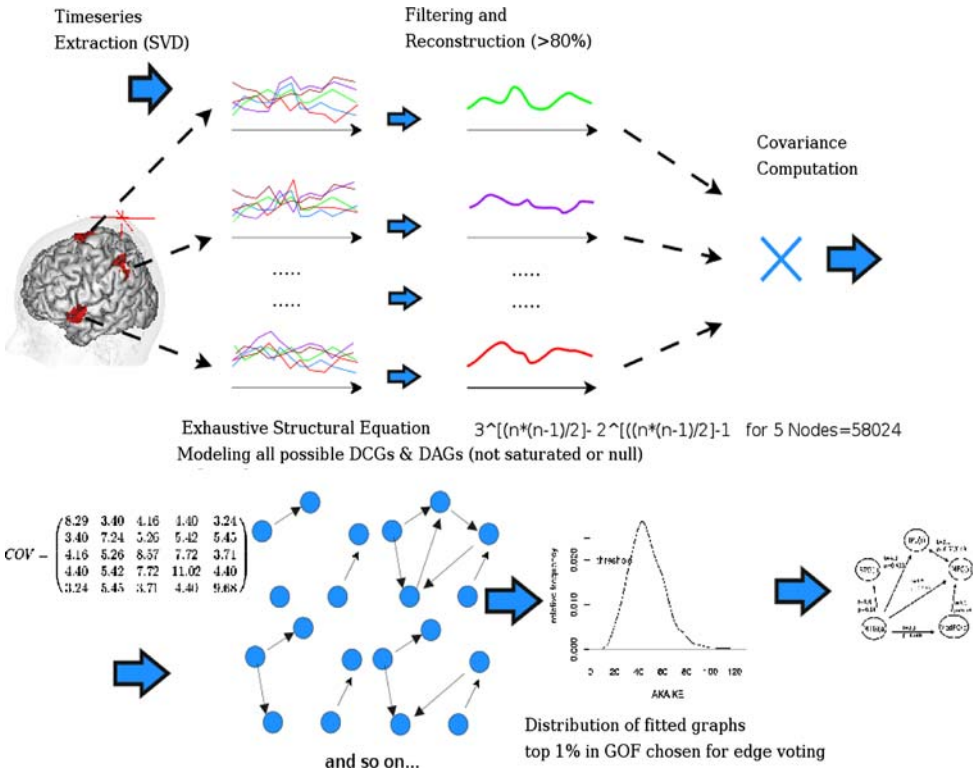


Fig. 9 Structural equation modeling procedure; we begin with dense mode clusters and extract time series based on each voxel that are filtered with the HRF convolved TRDs, the remaining set are submitted to an SVD producing an eigen-time series that can reconstruct at least 80% of the original time series variance. These eigen-time series are used to construct an ROI x ROI covariance matrix which is subsequently fit with all possible directed cyclic and acyclic graphs within the same Markov equivalence class, for 5 nodes there are 58,024 such graphs. These are sorted by goodness of fit (GOF) and the top 1% of these graphs are used to vote for edge presence and direction. Statistics are aggregated from individual edges and reported on the voted graph

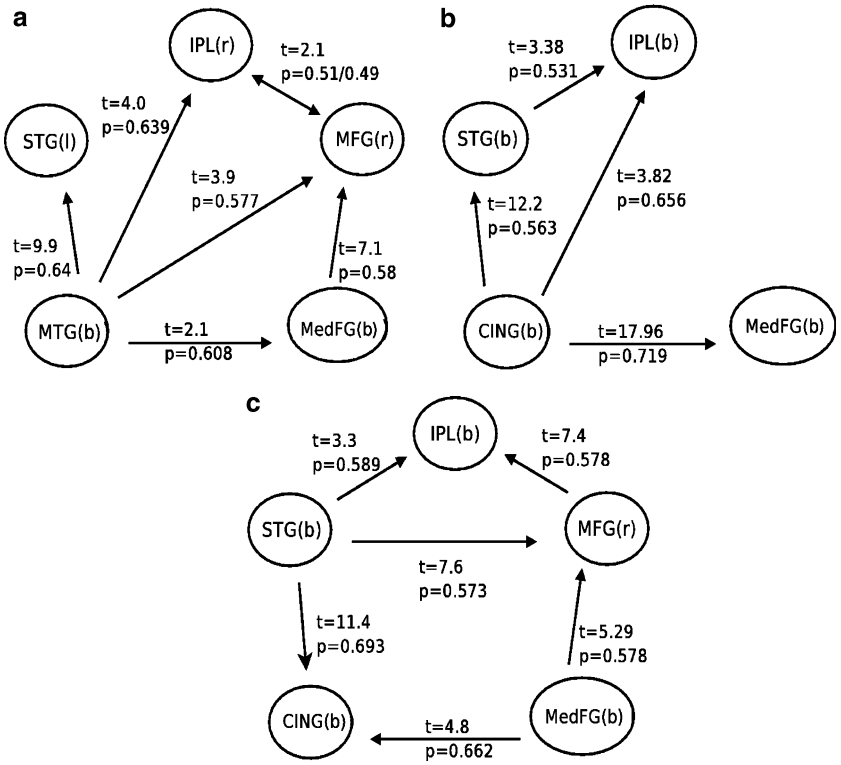


Fig. 10 (**a** *above*) Schema-rich graph "scripted" and its (**b** *right*) yoked schemafree "perceptual" counterpart. Note shared areas include Medial frontal gyrus STS and IPL. Cingulate is specific to yoked perceptual task. The "*t*" values on each edge indicates the significance of the influence and the "*P*" indicates the probability of the edge in the top 5% of the best fitting graphs, schema-poor brain interactivity graph. The schema-poor (**c** *bottom*) graph represents the response of subjects over 2 video sequences of a animation of a circle moving through a geometrically constrained space through a random trajectory. Some subjects would make up stories (e.g., "the circle was searching for something") about the circle-changed directions in the trajectory. In this case as expected there is a blending of the two previous graphs from the schema-rich and schema-free cases in the panel (**a**) and (**b**)

perceptual features in familiar action sequences. Second, is a subnetwork, consisting of areas (MT+, STS and IPL/IPS), that we propose is activated concept schema to clustered feature configurations to expectations based on activated schema.

Schema-poor predictions

A test of the present graph structures is to compare the schema-rich graph and schema-free to an intermediate case. Such a case was first studied by Heider and Simmel (1944) consisting of cartoon geometric objects moving in arbitrary paths through a space of other cartoon objects. In this case, subjects still reliably parsed the video sequence despite its unfamiliarity. Significant change points were seen as occuring when a change of direction, pausing of the object near edges and other coincidental smooth motion interruptions. Some subjects attempted to provide elaborate accounts of the arbitrary motion sequences. Because this task is somewhere between schema-rich sequences and schema-free perceptual tasks, we predicted would involve components of the perceptual detection task as well as components of the rich video sequence task inasmuch subjects have a bias to "tell a story" about any type of animation sequence. Consequently, we might expect to see both the sub-network that was associated with purely bottom-up processing (i.e., cingulate, STS, MT) and the schema activation sub-network (IPL/IPS, MFG, MedFG). Shown in Fig. 10c is a graph estimated as before over 10 subjects, 2 video sequences and in this case, 510 time points per subject. Note that as expected the schema-poor graph is a hybrid of the subnetworks implicating both schema activation and bottom-up visual search such as that seen in both schema-rich and schema-free graphs.

Discussion

In the present research we provide a framework for detecting networks of brain regions that reflect different levels of processing associated with classical "top-down" and "bottom-up" perceptual/cognitive functions. We show

how brain areas that have been associated in many kinds of visual attention and recognition tasks (IPL/IPS, MFG, MedFG) appear as sub-networks that form the basis for hypotheses concerning constituent schema activation/identification functions. These kinds of graphs are focused on influence connectivity as opposed to anatomical (known or not) connectivity and show how brain areas can influence each other in real-time cognitive/perceptual tasks. They do not however constitute a theory about event perception or how each of these specific brain areas may influence perceptual processing. In the next section we discuss a model first proposed by Hanson and Hanson (1996). This model consists of a specific recurrent neural network in which "state" or memory information is maintained over time, conditionally dependent on present stimulus input, and can seamlessly transition from one stored schema to another in order to minimize prediction error. The canonical theory of this type was first posited by Neisser (1976) and Schank (1975) in different forms and different contexts, but with substantially the same goal to explain how we organize and use memory of similar episodes to predict, filter and comprehend present sensory information flow.

A model of event perception

Schema or script processing theories first appeared in the 1970s in order to account for the continuity of perceptual processing and narrative story comprehension (Rumelhart and Ortony 1977). Schema as a concept was first adapted by Bartlett (1932) to refer to a perception/action/goal knowledge structure. This concept was further extended by Rumelhart in the 1970s to include story comprehension. According to Rumelhart a story was encoded as an episodic structure or event that could be traced to an actor-action-goal sequence. These episodic events would be implemented by schemata which would include nested subgoals that would recursively satisfy the larger schemata goal. Many results and details concerning memory organization and story summarization seemed consistent with this account. Computational accounts that attempted to capture the schema concept were introduced by Schank (1975) and provided simple data structures ("scripts") that were filled in as expectations within a story were realized. Earlier Neisser (1967) had proposed a schema-processing model for the what he termed the "perceptual cycle", a perceptual processing account of how the world remains stable and smooth in the face of constant changing stimulus variation and diversity (see Fig. 11). In Neisser's scheme, a stimulus object or event would initiate some specific schema which in turn would generate expectations which in turn would collect more data about the world and the object context. Confirmation of the same schema continues the perceptual cycle, while new data or data inconsistent with the existing schema modifies or initiates new schema. Computationally this was captured in a simple model proposed by Hanson and Hanson (1996), using a recurrent neural network (see Fig. 12). This network was constructed to be similar in form to the perceptual cycle (see Fig. 11) and captured the concept of a controlled search through a schema space. In this case stimulus input provided the "object" initiation or modification of the schema, while the hidden layer stored or coded for schema that might be more general then the input stimulus itself, while smoothly predicting the event change points and event stable points. This model was able to simulate many known experimental results concerning memory organization and to exhibit schema type behavior. In the next section we will map our graphs to the schema RNN model and propose how functional areas and their relation to the different graphs are consistent with known interpretations of brain function in schema activation, maintenance, top-down filtering.

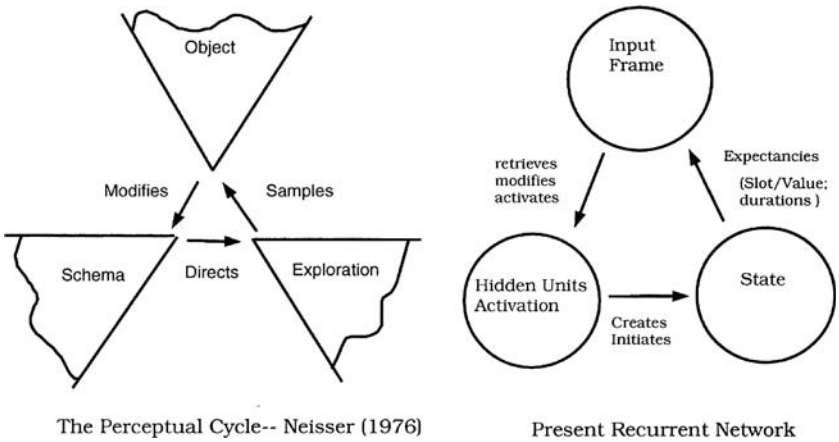


Fig. 11 Model of Event Perception (*left*) Neisser's Perceptual Cylce, (*right*) the conceptual implementation of the perceptual cycle as a recurrent neural network see Fig. 12

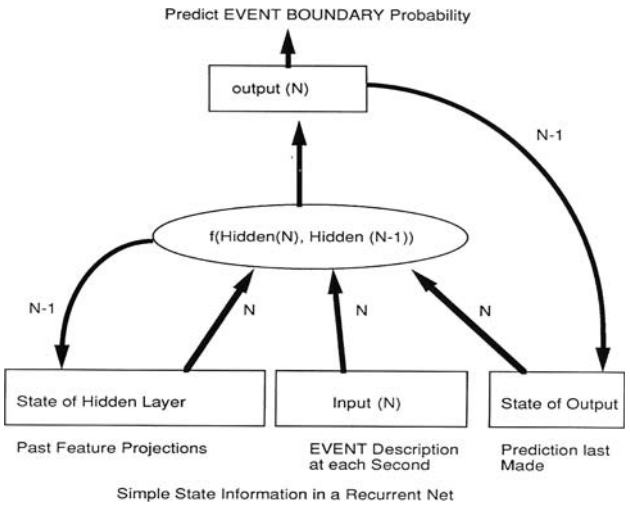


Fig. 12 Recurrent neural network conceptualized in Hanson and Hanson (1996) used to model event perception task. Note Fig. 11 right part of figure showing relationship to Neisser's perceptual cycle

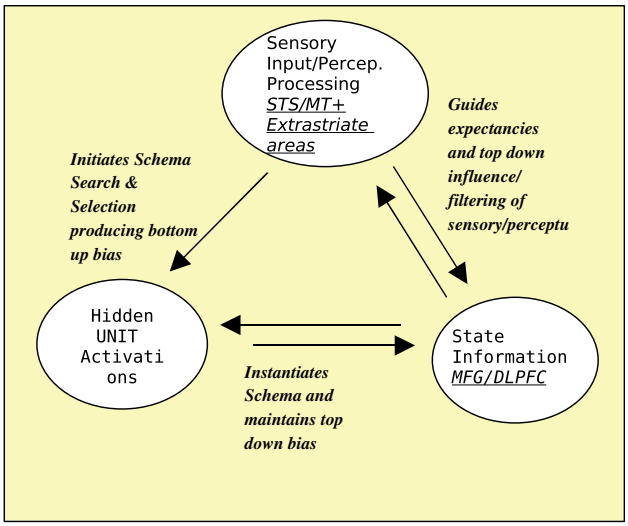


Fig. 13 Theoretical functional interactions between brain areas during event perception tasks. Although this is a feedback cycle, if entered at Sensory processing (bottom-up) we are assigning STS/MT+ and extriate areas for initatiating a search for relevant schema which in turn interacts with prefrontal areas to guide (top-down) expectancies and filtering sensory processing

Neural correlates of event processing

The present model begins with the identification of MT+ and STS for event change detection, which are known to be activated by various motion cues and have been shown previously to be associated with event change (Zacks et al. 2001). Based on the graphical models we propose that these areas provide cues for schema search and selection to medial frontal gyrus. Medial frontal gyrus is further mediated by inferior parietal lobule and nearby angular gyrus, both of which have been previously identified with attention shifting and secondary cuing. Given a schema is triggered by appropriate MT+/STS (as well as extrastriate areas and probably higher order featural processing in IT, although not in the present graphs) the IPL and MedFG help instantiate and select a specific schema which is maintained and implemented by PFC (MFG/DLPC), as well as guiding expectancies and filtering and influencing perceptual/sensory input (cf. Miller and Cohen 2001; Johnson et al. 2005). This type of implementation (see Fig. 13) can produce smooth transitions and maintain stable representations in the face of sensory or perceptual variation; much as Neisser proposed in his "perceptual cycle". We extend this model, however, in the present proposal by providing both neural mechanisms, represented by known prefrontal/parietal and perceptual brain areas, as well as providing a potential computational account that has sequential and temporal constraints. Unlike other approaches to event perception for creating stable representations (Carpenter et al. 1991), we do not expect a change point response or "reset" during event perception (which might very well implicate ACC or superior frontal gyrus), unless there are surprising sensory data not within the prediction scope of the activated schema.

Graphical models and ROIs: component theory

Almost every identifiable area of the brain has been shown to have multiple functions in response to diverse kinds of cognitive tasks. Recent reassessments of various areas whose function had been previously well established are now being reinterpreted in more abstract forms (e.g., Broca's area, IFG, as response selection, Thompson-Schill et al. 2005; fusiform "face" area, Kanwisher et al. 1997; Haxby et al. 2001; Hanson et al. 2004a; Hanson and Halchenko 2007, as complex feature reduction). What we believe is more likely to be unique about specific brain function is the pattern of constituent areas that are recruited for the task and the way they influence one another during that task (cf., McIntosh 2001). For the subnetworks we identified with bottom-up processing, consider a snapshot of the various functions ascribed to the constituents of this subnetwork that have been proposed in the literature:

IPL/IPS: attention shifting, theory of mind, visual awareness, spatial visual attention, spatial action intention, agency inference, etc.

MFG (DLPC) working memory, sustained memory functions, target detection, expectancy based decision making, effective attention, auditory categorization, object naming reading, language processing, tool use… etc.

MedFG: conflict monitoring, control of voluntary action, selection response set, "high level executive functions", decision processing, language supporting functions, processing of reward, verbal working memory, interracial judgments, numerical calculation, theory of mind... etc.

Theses types of observations have lead to a view that brain activity is more likely to reflect a complex sort of modularity where constituent functions are recruited opportunistically in the context of other such functions.

Complex modularity: Donder's Revisited Much of the logic of hypothesis testing in neuroimaging is based on a "factor and localize" framework often referred to as Donder's method (Posner and Raichle 1994). This method (originally defined for reaction time) assumes that brain activity can be factored into additive effects that can be localized by the appropriate baseline condition. "Rest" conditions are often used as baseline, but this could simply introduce more variability given the lack of control on subject's "rest" states. Nonetheless this factoring strategy has been successful in neuroimaging and does lead to functional localization, but perhaps not in any simple additive way. Graphical modeling in individual subjects has the potential for providing a compromise between approaches that focus on isolated modules and those focused on distributed computation. What is left to resolve in such schemes is the nature of the underlying constituents, levels of resolution, and the specific brain function. It should be clear that any given task and its variations (e.g.,) certainly involves cognitive functions (working memory, executive functions... etc.), and there may that might cause a radical restructuring of the graph edges and constiuent nodes (coherent brain areas). Fodor (1983) describes two kinds of modularity that are logically possible for the organization of mental computations. One is what he refers to as an horizontal modularity. That could be thought of as an array with cognitive functions listed horizontally in no particular order and with no particular level of granularity: "spatial memory" "language", "attention", "object recognition", and so on. A second possibility is vertical modularity, often the default hypothesis in cognitive neuroscience, e.g., "language organs", "face areas", "morality centers" etc. One could also specify a (yet to be specified) constituent modularity, one that reuses horizontal elements across many kinds of cognitive function independent of domain (e.g., "sequential order of information" instead of a "syntax area"). Of course, hierarchical organization of functional relations among neuronal groupings is familiar in vision and in motor control, e.g., in saccadic eye movements. The identification and decomposition of cognitive tasks that have well correlated neural circuitry of the kind we have considered here is an ongoing research task.

Although mapping cognitive function to localized brain areas has been the fundamental task driving the human brain mapping and cognitive neuroscience fields for the past decade, it may be obvious from these types of observations that identifying the *unique* function of specific areas of the brain is likely to be a an unsatisfying agenda, if such function is context sensitive relative to other recruited areas. Therefore, despite the apparent success of the present normative program in identifying memory, attention categorization, perception areas, it's typical for many other areas to be engaged during basic cognitive/perceptual tasks that are often considered "background", "secondary" or just irrelevant. As the neuroimaging field matures, theories of cognitive neuroscience may naturally involve hypotheses about interactions among as well as the causal influence that one brain area may have upon another, what has been termed effective connectivity (Friston et al. 2003) such as we have demonstrated here. Whether we consider language processing, working memory, or simple detection tasks, cognitive and perceptual processes are likely to include networks of regions that uniquely define a kind of computation. The key observation in this work is that constituents of brain activity may organize in networks where underlying brain activity forms some stable, but perhaps transient, computational function. Consequently, our ability to model underlying networks depends critically on detecting larger data structures (e.g., graphs) rather than local regions of interest as normative approaches in cognitive neuroscience do now.

References

Bartlett FC (1932) Remembering: an experimental and social study. Cambridge University Press, Cambridge

Buxton RB, Wong EC, Frank LR (1998) Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Mag Reson Med 39(6):855–864

Carpenter GA, Grossberg S, Rosen DB (1991) ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. Neural Netw 4(4):493–504

Desimone R, Duncan J (1995) Annu Rev Neurosci 18:193

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modeling. Neuroimage 19(4):1273–1302

Fodor JA (1983) The modularity of mind: an essay on faculty psychology. MIT, Cambridge

Hanson C, Hirst W (1989) On the representation of events: a study of orientation, recall, and recognition. J Exp Psychol General 118(2):136–147

Hanson C, Hanson SJ (1996) Development of schemata during event parsing: Neisser's perceptual cycle as a recurrent connectionist network. J Cogn Neurosci 8:119–134

Hanson S, Halchenko Y (2007) Support vector machines for object recognition: there is no face identification area. Neural Comput (in press)

Hanson SJ, Negishi M, Hanson C (2001) Connectionist neuroimaging. Emerg Neural Comput Architect Based Neurosci, pp. 560–576

Hanson SJ, Matsuka T, Haxby JV (2004a) Combinatoric codes in ventral medial temporal lobes for objects: Haxby revisited: Is there a "face" area? NeuroImage 23:156–166

Hanson SJ, Matsuka T, Hanson C, Rebbechi D, Halchenko Y, Zaimi A, Pearlmutter B (2004b) Structural equation modeling of neuroimaging data: exhaustive search and Markov Chain Monte Carlo. HBM-2004

Hanson SJ, Rebbechi D, Matsuka T, Halchenko Y, Hanson C (2007a) Methods for graphical modeling of brain interactivity. Neuroimage (submitted)

Hanson SJ, Rebbechi D, Zaimi A, Hanson C, Halchenko Y (2007b) Clustering brain maps with mode seeking algorithms. Magnetic Resonance Imaging (in press)

Hausson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. Science 303(5664):1634–1640

Haxby JV, Gobbini E. MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425

Heider F, Simmel M (1944) An experimental study of apparent behaviour. Am J Psychol 57(2):243–259

Johnson MK, Raye CL, Mitchell KJ, Greene EJ, Cunningham WA, Sanislow CA (2005) Using fMRI to investigate a component process of reflection: prefrontal correlates of refreshing a just-activated representation. Cogn Affective Behav Neurosci 5:339–361

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17(11):4302–4311

Kastner S, Ungerleider LG (2000) Annu Rev Neurosci 23:315

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortext function. Ann Rev Neurosci 24:167–202

McIntosh AR (1999) Mapping cognition to the brain through neural interactions. Memory 7(5/6):523–548

McIntosh AR (2001) Towards a network theory of cognition. Neural Netw 13:861–876

Neisser U (1967) Cognitive psychology. Appleton Century-Crofts, New York

Neisser U (1976) Cognition and reality, San Francisco, Freeman

Newtson D (1973) Attribution and the unit of perception of ongoing behavior. J Pers Soc Psychol 28:28–38

Penny WD, Stephan KE, Mechelli A, Friston K (2004) Compaing dynamic causal models. Neuroimage 22:1157–1172

Posner MI, Raichle ME (1994) Images of mind. W. H. Freeman, New York

Posner MI, Gilbert CD (1999) Attention and primary visual cortex. Proc Natl Acad Sci USA 96(6):2585–2587

Rumelhart DE, Ortony A (1977) The representation of knowledge in memory. Anderson RC, Spiro RJ, Montague WE (eds) Schooling and the acquiation of knowledge. Lawrence Erlbaum Assoc., Mahwah, pp 99–135

Smith S, Bannister P, Beckmann C, Brady M, Clare S, Flitney D, Hansen P, Jenkinson M, Leibovici D, Ripley B, Woolrich M, Zhang Y (2001) FSL: new tools for functional and structural brain image analysis. In: Seventh international conference on functional mapping of the human brain

Schank RC (1975) Conceptual information processing. North-Holland Publishing Co, Amsterdam

Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. Thieme, New york

Thompson-Schill SL, Bedney M, Goldberg RF (2005) The frontal lobes and the regulation of mental activity. Curr Opin Neurobiol 15:219–224

Zacks J, Braver TS, Sheridan MA, Donaldson DI, Snyder AZ, Ollinger JM, Buckner RL, Raichle ME (2001) Human brain activity time-locked to perceptual event boundaries. Nature Neurosci 4(6):651–655

Zacks JM, Tversky B (2001) Event structure in perception and conception. Psychol Bull 127(1):3–21