Convergent and Invariant Object Representations for Sight, Sound, and Touch

Kingson Man,^{1,*} Antonio Damasio,¹ Kaspar Meyer,^{1,2} and Jonas T. Kaplan¹

¹Brain and Creativity Institute, University of Southern California, Los Angeles, California, 90089

²Institute of Anesthesiology, University Hospital, University of Zurich, Zurich, Switzerland

Abstract: We continuously perceive objects in the world through multiple sensory channels. In this study, we investigated the convergence of information from different sensory streams within the cerebral cortex. We presented volunteers with three common objects via three different modalities—sight, sound, and touch—and used multivariate pattern analysis of functional magnetic resonance imaging data to map the cortical regions containing information about the identity of the objects. We could reliably predict which of the three stimuli a subject had seen, heard, or touched from the pattern of neural activity in the corresponding early sensory cortices. Intramodal classification was also successful in large portions of the cerebral cortex beyond the primary areas, with multiple regions showing convergence of information from two or all three modalities. Using crossmodal classification, we also searched for brain regions that would represent objects in a similar fashion across different modalities of presentation. We trained a classifier to distinguish objects presented in one modality and then tested it on the same objects presented in a different modality. We detected audiovisual invariance in the right temporo-occipital junction, audiotactile invariance in the left postcentral gyrus and parietal operculum, and visuotactile invariance in the right postcentral and supramarginal gyri. Our maps of multisensory convergence and crossmodal generalization reveal the underlying organization of the association cortices, and may be related to the neural basis for mental concepts. *Hum Brain Mapp* 36:3629–3640, 2015. © 2015 Wiley Periodicals, Inc.

Key words: fMRI; multivariate pattern analysis; multisensory; perception; crossmodal; auditory; visual; tactile; concepts

INTRODUCTION

We perceive objects in our environment through multiple sensory modalities. For example, we may recognize a dog by hearing its bark, seeing its picture, or rubbing its fur. These characteristic features are registered as radically different patterns of energy at our sensory portals—the ears, eyes, and skin—and are then represented in specialized, segregated regions of the brain. As far as the cerebral cortex is concerned, the initial representations take place in the early auditory, visual, and somatosensory cortices, respectively. In spite of this diversity at the perceptual and neural levels, we easily arrive at the same concept of an object regardless of which sensory modality was used to

Contract grant sponsor: The Mathers Foundation; Contract grant sponsor: the National Institutes of Health; Contract grant number: NIH/NINDS 5P50NS019632.

*Correspondence to: Kingson Man; Brain and Creativity Institute, University of Southern California, 3620A McClintock Ave, Los Angeles, CA, 90089-2921. E-mail: kman@usc.edu

Received for publication 16 December 2014; Revised 21 May 2015; Accepted 21 May 2015.

DOI: 10.1002/hbm.22867

Published online 5 June 2015 in Wiley Online Library (wileyonlinelibrary.com).

perceive it. This suggests that information streams from the different sensory channels converge somewhere in the brain and form representations that are invariant to the input modality. This study aimed to map the integration of sensory information from multiple modalities in the human brain.

Nearly half a century ago, the lesion tracing studies of Jones and Powell [1970] in macaque described a stepwise convergence of sensory pathways among the auditory, visual, and somatosensory modalities. The three modalities gradually converged in a region homologous to human angular and supramarginal gyrus. More recently, diffusion tensor imaging of the human brain showed that regions activated by sight-, sound-, or touch-implicating words shared converging connections to the angular gyrus [Bonner et al., 2013]. Several parietal regions were implicated in a functional magnetic resonance imaging (fMRI) study of activations evoked by motion stimuli in the auditory, visual, and tactile modalities [Bremmer et al., 2001]. An fMRI coactivation study found that the fusiform gyrus was sensitive to the auditory, visual, and tactile presentation of manipulable objects, as compared to textures [Kassuba et al., 2011]. The same study found that fusiform gyrus and angular gyrus were sensitive to crossmodal matching among the three modalities (e.g., simultaneous presentation of the sight, sound, and touch of the same object, as compared to mismatched stimuli). Several additional regions of the brain are known to be activated by more than one sensory modality [reviewed by Amedi et al., 2005].

In this study, we assessed the information content of these multisensory coactivations. We performed fMRI of the human brain as subjects heard, viewed, or touched three common objects with distinct features in each modality. We used multivariate pattern analysis (MVPA) to answer two specific questions. First, where in the brain do we find bi- or trimodal information convergence? In other words, which brain regions contain object-specific representations from multiple sensory modalities? Second, where in the brain do we find sensory invariance? In other words, which brain regions contain object representations that are similar across different modalities of object presentation?

MATERIALS AND METHODS

Subjects

Nineteen right-handed subjects were originally enrolled in the study. One subject was excluded from the analysis due to poor coverage of the brain by the scan's field of view. The data presented come from the remaining 18 participants, 8 female and 10 male. The experiment was undertaken with the informed written consent of each subject in a protocol approved by the University of Southern California's University Park Institutional Review Board and in accordance with the Declaration of Helsinki.

Stimuli

Auditory, visual, and tactile stimuli were generated for three common objects: a whistle, a wine glass, and a strip of Velcro. The objects were chosen based on their mutually distinct auditory, visual, and tactile qualities. Video and audio recordings were made of the objects being actively manipulated by one of the authors (K. Man); the wine glass was tapped with a fingernail, the whistle was blown, and the pieces of Velcro were attached and separated. Auditory stimuli were generated from the audio tracks of the videos and played over a black screen. Visual stimuli were generated from the video tracks and played in silence. The audio and video extracts were all 5-s long and did not overlap in time, to prevent a correspondence between the sight and sound of the same object at the level of simple temporal dynamics. Tactile stimuli were delivered by placing the objects themselves into the hands of subjects for bimanual free exploration for 5 s. An additional set of audiovisual clips, containing both audio and video tracks, was generated for the trimodal stimulus condition, used only in the functional localizers.

Stimulus Presentation

Prior to scanning, subjects were briefly introduced to the three objects and practiced exploring each one with both hands. The participants were instructed to pay close attention to the sensory qualities of each stimulus as it was presented, and to keep their eyes open during all runs. Inside the fMRI scanner, subjects were presented with stimuli in separate auditory (A), visual (V), and tactile (T) runs, 12 in total, in the order AVTVATATVTVTA. Each run contained 15 stimulus presentations: each of the three objects was presented five times in pseudorandomized order with no consecutive repeats. One stimulus was presented every 11 s. A sparse-sampling scanning paradigm was used to ensure that all stimuli were presented in the absence of scanner noise: a single whole-brain volume was acquired starting 2 s after the end of each 5-s stimulation period. The 2-s image acquisition was followed by a 2-s pause, after which the next trial began. Timing and presentation of the auditory and visual stimuli was controlled with MATLAB 7.9.0 (The Mathworks), using the Psychophysics Toolbox 3 software [Brainard, 1997]. Sound intensity of the audio clips was adjusted to the loudest comfortable level for each subject. Video clips were displayed on a rear-projection screen at the end of the scanner bore, which subjects viewed through a mirror mounted on the head coil. The same sound clip or video clip was used to depict each object over the course of the experiment. Each clip was repeated 20 times in total. For tactile stimulation, the experimenter placed the objects into the waiting hands of the subject and removed them after 5 s. Subjects were instructed, during T runs, to place their hands together in a shallow-cupped position at waist level with palms facing up to receive the object. As soon as the object was

received, subjects actively explored it with both hands. Subjects were instructed to avoid looking at, or producing sounds with, the objects while they manipulated them. The experimenter, who was present for all T runs, verified good compliance of active bimanual exploration of the objects.

Prior to the 12 runs for MVPA, we performed two runs of a functional localizer. During the localizer runs, stimuli were presented either unimodally (sound only, sight only, and touch only) or trimodally (simultaneous sound, sight, and touch). Each of these 12 unique stimuli was presented three times, for a total of 36 stimulus presentations per run. The stimuli were presented in pseudorandomized order with no consecutive repeats. The duration of each trial was randomly jittered up to 1 s about the mean duration of 15 s, and images were acquired continuously according to a slow event-related design. These functional localizers served as an independent dataset on which we performed a conventional univariate analysis to select voxels and generate masks for MVPA.

Image Acquisition

Images were acquired with a 3-Tesla Siemens MAGNETOM Trio System. Echo-planar volumes for MVPA runs were acquired with the following parameters: repetition time (TR) = 11,000 ms, acquisition time (TA) = 2,000 ms, echo time (TE) = 25 ms, flip angle = 90°, 64×64 matrix, in-plane resolution 3.0×3.0 mm², 41 transverse slices, each 2.5-mm thick, covering the whole brains of most subjects. Volumes for functional localizer runs were acquired with the same parameters except in continuous acquisition with TR = 2,000 ms. We also acquired a structural T1-weighted MPRAGE for each subject (TR = 2,530 ms, TE = 3.09 ms, flip angle = 10°, 256×256 matrix, 208 coronal slices, 1 mm isotropic resolution). The structural scan was collected after the sixth MVPA run, serving as a passive rest period for the subject in the middle of the 12-run sequence.

Univariate Analysis of Functional Localizer Runs

We performed a univariate analysis of the functional localizer runs using the FMRIB Software Library (FSL) 5.0 [Smith et al., 2004]. Data preprocessing involved motion correction [Jenkinson et al., 2002], brain extraction [Smith, 2002], slice-timing correction, spatial smoothing with a 5-mm full-width at half-maximum Gaussian kernel, high-pass temporal filtering using Gaussian-weighted least-squares straight line fitting with sigma (standard deviation of the Gaussian distribution) equal to 60 s, and prewhitening [Woolrich et al., 2001].

The four stimulus types, auditory (A), visual (V), tactile (T), and trimodal (AVT), were modeled separately with four regressors derived from a convolution of the task design and a gamma function to represent the hemodynamic response function. Motion correction parameters were included in the design as additional regressors. The

two functional localizer runs for each participant were combined into a second-level fixed-effects analysis, and a third-level intersubject analysis was performed using a mixed-effects design.

Registration of the functional data to the high-resolution anatomical image of each subject and to the standard Montreal Neurological Institute (MNI) brain was performed using the FSL FLIRT tool [Jenkinson and Smith, 2001]. Functional images were aligned to the high-resolution anatomical image using a six-degree-of-freedom linear transformation. Anatomical images were registered to the MNI-152 brain using a 12-degree-of-freedom affine transformation.

Five activation maps were defined using the functional localizer data: three “unimodal” activation maps of the areas activated during the presentation of each modality compared to rest ($A > R$, $V > R$, $T > R$), a “multisensory” map of regions activated during simultaneous AVT stimulation compared to rest ($AVT > R$), and a “superadditive” map of areas in which the trimodal stimuli evoked activity greater than the sum of the activations from the three unimodal stimuli ($AVT > A + V + T$). The activation maps were thresholded with FSL’s cluster thresholding algorithm using a minimum Z-score of 2.3 ($P < 0.01$) and a cluster size probability of $P < 0.05$.

Localizer-Based Voxel Selection for MVPA

We used the localizer results to select voxels for seven different masks within which MVPA was performed. First, three unimodal masks were constructed by selecting the 250 voxels with the highest *t*-values from each subject’s corresponding A, V, and T unimodal functional localizer maps. Additionally, three bimodal masks, AV, AT, and VT, and a trimodal mask, AVT, were created from the overlaps between the unimodal maps. For instance, to create the AV mask, we identified voxels that were significantly activated in both the $A > R$ and the $V > R$ maps. Voxel values were then rescaled to be the percentage of the maximum value of each respective map. These values were summed across the two maps, and the top 250 voxels in this combined map were taken to form the AV mask. The analogous procedure was used to create the AT, VT, and AVT masks.

Mask-Based MVPA

MVPA was performed using the PyMVPA software package [Hanke et al., 2009] in combination with LibSVM’s implementation of the linear support vector machine [Chang and Lin, 2011]. Data from the 12 MVPA runs of each subject were concatenated and motion-corrected to the middle volume of the entire time series, then linearly detrended and converted to Z-scores by run. We performed intramodal and crossmodal classification within the relevant localizer-based masks. There were

three possible two-way discriminations among the three stimuli: for three stimuli (W)histle, (G)lass, and (V)elcro, they are [W,G], [W,V], and [G,V]. For intramodal classification, we performed a leave-one-trial-out cross-validation procedure. A classifier was trained on data from all but one of the stimulus presentations and then tested to decode the left-out trial. This procedure was repeated with data from each trial serving as the testing set once. All of the test accuracies from all of the two-way discriminations were averaged to arrive at the overall intramodal classifier accuracy. For crossmodal classification, two-way discriminations were performed with a classifier trained on all stimulus presentations in one modality and tested on all stimulus presentations in a different modality. Test accuracies from the three possible two-way discriminations were averaged. For the three modalities A, V, and T, there were six orderings of training and testing: [A,V], [V,A], [A,T], [T,A], [V,T], and [T,V]. For each pair of modalities, two orderings of training and testing were averaged to arrive at the overall crossmodal accuracy. For example, [A,V] and [V,A] accuracies were averaged to yield the overall A–V crossmodal classifier accuracy.

Searchlight Analyses

To conduct a mask-independent search for brain regions containing information relevant to intramodal or crossmodal classification, we performed a searchlight procedure [Kriegeskorte et al., 2006]. In each subject, a classifier was repetitively applied to small spheres ($r = 3$ voxels) centered on every voxel of the brain. The classification accuracy for each sphere was mapped to its center voxel to obtain a whole-brain accuracy map for each of three intramodal and six crossmodal classifications. Our searchlight maps included only brain voxels that were imaged in all 18 subjects. The extent of our brain maps in MNI space was bounded by $Z = 70$ dorsally and $Z = 32$ ventrally, which omitted some coverage of the ventral surface of the anterior temporal lobe.

The intramodal classifier was trained on data from all but one of the stimulus presentations, and then tested on the remaining presentation (leave-one-trial-out cross-validation). Crossmodal classifiers were trained on all trials of one modality and tested on all trials of the other modality. Searchlights were performed for each of the three different two-way discriminations of the stimuli and then averaged together. Maps from each subject were then warped into the standard space and visualized with FSL and MRIcroGL [Chris Rorden, <http://www.cabiatl.com/mricro/mricrogl>]. The three unimodal searchlight maps were overlaid and regions of overlap were found. Overlap maps were generated by rescaling each unimodal searchlight map to be the percentage of the maximum value of that map, then selecting only voxels that were significant in both (or all three) of the component maps and summing them. The crossmodal searchlight maps, A–V, A–T, and

V–T, were generated by averaging across both orderings of training and testing for each pair of modalities.

Statistical Analyses

Given that our hypothesis was directional—classifier performance on two-way discriminations should be higher than the chance result of 0.5—we used one-tailed t -tests across 18 subjects to assess the statistical significance of our classifier accuracies in the mask-based analyses. For the searchlight maps, we performed a nonparametric one-sample t -test using the FSL randomize program [Winkler et al., 2014] with default variance smoothing of 5 mm. Under the null hypothesis, the signs of the difference measures may be randomly permuted, or flipped, without affecting the overall distribution of values. Here, our difference measure was classifier performance in excess of chance. We subtracted chance level from the observed accuracy values, resulting in a set of positive- and negative-signed values denoting above- and below-chance performance. We quantified the likelihood of observing a set of accuracies as high as in our original data, from a distribution composed of sets of accuracies with sign-flipped data. The signs of the data were randomly negated for 5,000 iterations for each of the six classification types (intramodal A, V, T and crossmodal A–V, A–T, V–T). We thresholded our searchlight maps at the 99th percentile of the empirical null distributions (P level of 0.01; $t > 2.567$; degrees of freedom (DOF) = 17). We next corrected for multiple comparisons with a cluster mass threshold of $P < 0.05$. We selected only voxels that were part of clusters with a mass greater than the 95th percentile of cluster mass maxima across the 5,000 sign-flipped, t -thresholded maps. Crossmodal maps underwent the same procedure with the addition of a premasking step: we only considered regions in which intramodal decoding of both relevant modalities was successful. This step was added because a necessary precondition of decoding stimuli across two modalities is that the stimuli must be discriminable within each of the modalities.

RESULTS

Functional Localizer Activation Mapping

At the group level, auditory, visual, or tactile stimuli, as compared to rest, produced activity clusters in the auditory, visual, and somatosensory/motor cortices, respectively (Fig. 1). The presentation of simultaneous AVT stimuli compared to rest activated broad regions of the cortex resembling a concatenation of the unimodal maps, with peaks in auditory, visual, and somatosensory/motor cortices (Fig. 1, AVT map). However, the superadditive map ($AVT > A + V + T$) revealed that activity in the auditory, visual, and tactile cortices was greater during simultaneous trimodal stimulation than during isolated unimodal stimulation. Thus, the superadditive map

Auditory > Rest **Visual > Rest** **Tactile > Rest** **AVT > Rest** **AVT > Auditory + Visual + Tactile**

revealed a multimodal potentiation effect. All brain maps presented in the text may be downloaded as volumetric statistical images at any of the following mirrored links:

<http://www-bcf.usc.edu/~jtkaplan/data.php>
<http://goo.gl/30a30c>
<http://goo.gl/0kGndK>
<http://dx.doi.org/10.6084/m9.figshare.1394725>

Classifier Performance in Localizer-Based Masks

For each subject, MVPA was performed within the masks based on that subject's own functional localizer data (see Materials and Methods section). Group-level accuracies were averaged across 18 subjects and three pairwise discriminations of the three objects.

Intramodal classification

Each type of stimulus (sounds, videos, and touches) could be classified significantly better than chance from neural activity in the corresponding unimodal mask (auditory, visual, tactile; Table I). Several types of stimuli could also be predicted at higher-than-chance level in masks of a different modality. Thus, sounds could be distinguished at levels significantly better than chance in the visual mask, visual stimuli were classified significantly better than chance within the auditory mask and the tactile mask, and tactile stimuli were classified significantly better than chance within the auditory mask and the visual mask.

Crossmodal classification

Classifiers trained in one modality and tested in another were, in general, not significantly better than chance

Figure 1.

Univariate analysis of brain regions activated by auditory, visual, and tactile stimuli. Hearing sounds most strongly activated the superior temporal lobe, peaking in Heschl's gyrus and the planum temporale. Watching (silent) videos broadly activated occipital and parietal cortices, peaking in the occipital pole. Touching objects resulted in an activation pattern that peaked in pre- and postcentral gyri and included additional, broad aspects of the parietal lobe. The simultaneous presentation of objects in all three modalities (AVT), as compared to rest, resulted in a broad cortical activation pattern which peaked in the auditory, visual, and somatosensory cortices, and resembled a concatenation of the unimodal maps. A "superadditive" map of regions in which simultaneous AVT presentation evoked greater activity than the sum of the separate unimodal activations revealed a multimodal potentiation effect. All slices are in radiological convention (right side of image corresponds to left side of brain). MNI coordinates specify the slices displayed, and not the locations of peak voxels. All volumetric statistical images in this article may be downloaded at the links provided in the text. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE I. Classifier performance in localizer-based masks at the group level—intramodal classification accuracies

Stimulus modality \ mask	Auditory mask	Visual mask	Tactile mask
Auditory stimuli	0.725****	0.547*	0.539 (ns)
Visual stimuli	0.591****	0.769****	0.658****
Tactile stimuli	0.564***	0.556**	0.606****

Sensory stimuli were well decoded in their respective sensory cortical masks (shaded cells). Stimuli were also decoded at levels better than chance in the sensory cortical masks of a different modality, for all combinations of stimulus modality and sensory mask, with the exception of sounds in the tactile mask. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$; ns, not significant.

within the corresponding localizer-based masks (Table II). As the only exception, crossmodal decoding of videos and touches was successful in the V–T mask.

Searchlight Information Mapping

Intramodal searchlights

Group-level searchlight analyses evidenced the highest levels of classification accuracy for each type of stimulus within the corresponding sensory cortices (Fig. 2, Left). Auditory stimuli were decoded most accurately along Heschl's gyrus and the planum temporale, bilaterally. Visual stimuli were decoded most accurately in the medial and lateral occipital lobe and fusiform gyrus, bilaterally. Tactile stimuli were decoded most accurately in the ventral aspect of the postcentral gyrus, bilaterally. However, Figure 2 (left panel) illustrates impressively that intramodal classification of sounds, videos, and touches was also successful in vast regions of the cortex outside the corresponding early sensory areas. Thus, classification accuracies were significantly above chance level in several of the regions broadly defined as “association” cortices, as well as in the early sensory cortices of modalities other than that of the stimulus.

There were areas of overlap among the A, V, and T unimodal searchlight maps (Fig. 2, Right). Regions that discriminated both sounds and videos included Heschl's gyrus and planum temporale, superior temporal gyrus and sulcus, the ventral part of postcentral gyrus and adjacent postcentral and central sulcus, and the posterior part of fusiform gyrus, all bilaterally. Regions that discriminated both sounds and touches included planum temporale, posterior parietal operculum, and supramarginal gyrus, bilaterally. Regions that discriminated both videos and touches included superior parietal lobule, supramarginal gyrus, parietal operculum, the superior and inferior divisions of lateral occipital cortex, and inferior temporal-occipital cortex, bilaterally. Regions that discriminated all three types of stimuli included posterior planum temporale, parietal operculum, supramarginal gyrus, and secondary somatosensory cortex, bilaterally.

Crossmodal searchlights

Crossmodal invariance was detected in lateralized and highly focal regions of the association cortices (Fig. 3). A cluster of audiovisual invariance was found in the right hemisphere, at the junction of lateral occipital cortex and posterior temporal cortex. Audiotactile invariance was found in the left secondary somatosensory cortex, along the postcentral gyrus and parietal operculum. Visuotactile invariance was found in two distinct clusters in the right hemisphere, one spanning postcentral gyrus, superior parietal lobule, supramarginal gyrus and parietal operculum, and the other in medial premotor cortex.

DISCUSSION

In this study, we mapped brain regions containing information about the identity of common objects presented in three sensory modalities. The intramodal classification results indicated that auditory, visual, and tactile stimuli are discriminated not only within their corresponding early sensory cortices, but also in association cortices and even in early sensory areas of different modalities. As is evident from Figures 1 and 2 (Left), the intramodal

TABLE II. Classifier performance in localizer-based masks at the group level—crossmodal classification accuracies

Modalities \ mask	Auditory	Visual	Tactile	AV	AT	VT	AVT
A–V	0.514 (ns)	0.503 (ns)		0.515 (ns)			0.519 (ns)
A–T	0.506 (ns)		0.488 (ns)		0.498 (ns)		0.503 (ns)
V–T		0.491 (ns)	0.528 (ns)			0.556*	0.506 (ns)

Crossmodal classifications were performed in the relevant unimodal and bimodal masks (e.g., audiovisual crossmodal classification was performed in the auditory, visual, and audiovisual masks), as well as in a trimodal mask. In general, crossmodal classification in the localizer-based masks was indistinguishable from chance level. There was one exception: crossmodal classification of videos and touches was successful in the VT bimodal mask. Chance performance for all classifications is 0.5; the statistical significance of better-than-chance accuracy at the group level was assessed with one-tailed t -tests. * $P < 0.05$; ns, not significant.

Figure 2.

Intramodal searchlight analysis. Left, Group-average classification accuracy for sounds, videos, and touches. Auditory stimuli were most accurately decoded in the auditory cortices, peaking along Heschl's gyrus. Visual stimuli were most accurately decoded in the calcarine sulcus and occipital pole. Tactile stimuli were most accurately decoded along the pre- and postcentral gyri. Successful classification, for each modality, also extended beyond the respective sensory cortices, into the association cortices as well as the early sensory cortices of the other modalities. Maps were thresholded at the P level of 0.01 ($t > 2.567$; $DOF = 17$) and then a cluster mass P level of 0.05, as determined by a nonparametric one-sample t -test (5,000 permutations of random sign-flipping of the data). Right, Overlaps among the intramodal classifications. A crossmodal overlap score was calculated by sum-

ming the normalized accuracy values of the input maps and selecting only the shared voxels. Auditory and visual stimuli were jointly best decoded in parietal operculum, supramarginal gyrus, posterior temporal cortex, and temporal and occipital fusiform gyri. Auditory and tactile stimuli were jointly best decoded in planum temporale, parietal operculum, and supramarginal gyrus. Videos and touches were jointly best decoded in superior parietal lobule, supramarginal gyrus, parietal operculum, and lateral and inferior occipital cortex. Stimuli from all three modalities were jointly best decoded in the parietal operculum, supramarginal gyrus, and secondary somatosensory cortex, as well as lateral and ventral occipital cortices. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

classification maps closely resembled the corresponding localizer maps. This indicates that nearly all regions that increase their activity during sensory stimulation do so in a manner specific to the stimuli, that is, they carry information about the identity of the presented object. The observation that stimuli of a specific sensory modality are encoded in regions of the cortex that reach far beyond the early sensory areas of that modality, and even into the early cortices of other modalities, contributes to a growing appreciation that “modality specific” cortices may be less specific than commonly assumed [Schroeder and Foxe, 2005], and that neocortex may be “essentially multisensory” [Ghazanfar and Schroeder, 2006]. In this regard, our current results complement earlier work showing that

primary sensory cortices of the auditory, visual, and somatosensory modalities contain information that can determine in which other modality a stimulus was presented, as well as the location at which it was presented [Liang et al., 2013]. We also replicate the findings of various other studies showing that information about a stimulus presented in a given sensory modality can be found in the cortices of a different modality: auditory information can be detected in visual [Vetter et al., 2014] and in somatosensory [Etzel et al., 2008] cortices; visual information can be detected in auditory [Meyer et al., 2010; Hsieh et al., 2012; Man et al., 2012] and in somatosensory [Meyer et al., 2011; Smith and Goodale, 2015] cortices; tactile information can be detected in auditory [Kayser et al., 2005, in macaque] and in visual [Oosterhof et al., 2012] cortices.

The overlap maps between the intramodal searchlights (Fig. 2, Right) indicate that several regions discriminated stimuli of two or even all three modalities (the latter being the case in parietal operculum, supramarginal gyrus, and lateral occipitotemporal cortices). However, comparing these overlap maps with the maps of successful crossmodal classification (Fig. 3), it becomes obvious that only a subset of the regions that could decode stimuli in two modalities also represented these stimuli in similar fashion across the modalities. For example, multiple regions could distinguish both sounds and videos, but only a focal region of right occipitotemporal cortex represented corresponding sounds and videos in a similar manner. This discrepancy between overlapping intramodal and true crossmodal classification reveals that the colocalization of information from different sensory modalities does not necessarily imply invariant representation among those modalities. We do expect, however, that larger and additional regions of invariance will be found with improved methodology.

A-V invariance

A-T invariance

V-T invariance

t-stat
2.567 [max]

Figure 3.

Crossmodal searchlight analysis. Classifiers trained to distinguish objects presented in one sensory modality were used to distinguish the same objects presented in a different sensory modality. Thus, above-chance performance indicates the presence of modality-invariant information about the identity of the object. Top, Audiovisual invariance was detected in a right hemisphere cluster located at the junction of lateral occipital cortex and posterior temporal cortex. Middle, Audiotactile invariance was detected in left secondary somatosensory cortex, along the inferior postcentral gyrus and parietal operculum. Bottom, Visuotactile invariance was detected in two right hemisphere clusters: one that extended from postcentral gyrus and superior parietal lobule to supramarginal gyrus and into parietal operculum, and the other in medial premotor cortex. Maps were thresholded at the P level of 0.01 ($t > 2.567$; $DOF = 17$) and then a cluster mass P level of 0.05, as determined by a nonparametric one-sample t -test (5,000 permutations of random sign-flipping of the data). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We intentionally chose objects that enjoyed rich sensory associations to maximize our ability to detect such associations. Motion and physical interaction, being highly diagnostic features of these objects, likely contributed to our successful detection of invariant representations. Our real-world objects, as opposed to the majority of synthetic laboratory stimuli, also possess well-established associations acquired through multiple other channels, including emotion or olfaction, and these undoubtedly contributed to our results as well. The strong and perhaps overlearned nature of these objects, as well as the response-free design of our study, may contribute to the null result for invariant representation in the prefrontal cortices. We expect that prefrontal cortices would be recruited while learning novel crossmodal associations, or under a more demanding task.

Due to the small set of objects used in the study ($n = 3$), one could argue that successful crossmodal classification may not truly reflect modality-invariant representations of stimulus identity but, instead, some other confounding or nuisance variable. For example, audiotactile classification may succeed due to a systematic relation of attentional capture or vividness of imagery among the objects, rather than crossmodal invariance regarding object identity. A nuisance variable would have to produce an identical ordering of the objects in both modalities of presentation. The reduction of possible orderings as the number of objects to be distinguished decreases makes this problem especially severe for two-class decoding, but still worthy of consideration in our three-class problem. From a different perspective, however, it may be argued that the identifying information of an object is precisely the set of all such “confounding” variables that systematically distinguish objects from each other. Our operational definition of invariant representations—content-specific patterns of activity that are preserved across different stimulation modalities—was fulfilled.

Crossmodal Invariances in Context

Many recent studies have used multivariate crossmodal classification, in which an algorithm is trained to classify data from one sensory modality and tested to classify data from a different sensory modality, to detect modality invariant representations [reviewed by Kaplan et al., 2015]. In our previous study of audiovisual invariance [Man et al., 2012], we also found evidence for common coding at the junction of the temporal, parietal, and occipital lobes. Although that region is not far from the cluster of audiovisual invariance identified in this study, its location was more anterior and superior. Intriguingly, in keeping with the current results, audiovisual invariance was found almost exclusively in the right hemisphere. A right-lateralization for audiovisual integration was also found in a meta-analysis of studies of audiovisual speech signals: the activation likelihood estimate cluster for “validating,” or congruent, crossmodal speech signals was more prominent in the right superior temporal sulcus than in the left

[Erickson et al., 2014]. A handful of other studies have used MVPA to identify audiovisual invariant representations. Peelen et al. [2010] decoded emotional categories across auditory and visual expressions of emotion in left pSTS and medial prefrontal cortex, which were attributed to their roles in mentalizing and theory of mind. Akama et al. [2012] decoded the categories of animals or tools across the spoken or printed names of objects belonging to each category, using voxels from throughout the brain. Ricciardi et al. [2013] discriminated actions from nonactions across their auditory and visual presentations using voxels from the right temporo-parietal-occipital junction, but not from the left. Ventral premotor and superior parietal voxels were also recruited bilaterally for this classification. Simanova et al. [2014] discriminated the object categories of tools or animals across formats that included auditory and visual presentations, primarily in the right ventrolateral prefrontal cortices and the left TPJ and temporo-occipital cortices. Summing up, the temporo-occipital boundary regions located between auditory and visual cortices are often involved in representing auditory and visual stimuli in an invariant manner. The presence of lateralization and the additional cortical regions identified may vary with the nature of the stimuli and the tasks involved.

We are aware of one prior MVPA study of invariant representation involving the auditory and motor/tactile modalities: Etzel et al. [2008] found that bilateral premotor cortices discriminated hand and mouth actions that were either heard or performed. Their classification was not significantly above chance in either left or right secondary somatosensory cortices, in contrast to our detection of A–T invariance exclusively in the left secondary somatosensory cortex and parietal operculum. This discrepancy may be due to the differences between experimental tasks: the performance of specific actions versus sensorimotor exploration. A univariate fMRI study found that bilateral parietal operculum was selectively activated by the haptic perception of texture as opposed to shape [Stilla and Sathian, 2008]. Although that study focused on visuohaptic convergence, we suggest that texture exploration may also have a significant auditory component. In materials of appropriate hardness, the spatial frequency of texture features may be directly perceived by auditory feedback: by running one’s hand across the surface and hearing the different sounds produced. Thus, the Stilla and Sathian [2008] finding of haptic texture-selectivity in parietal operculum may be reconciled with our identification of the region, at least in the left hemisphere, to contain audiotactile invariant representations.

Pietrini et al. [2004] reported visuotactile invariant representations for the categories of shoes and bottles in bilateral inferior temporal cortex, in which activity patterns correlated across seeing or touching those object types. Their study also highlighted the impact of stimulus selection on a cross-classification study: visuotactile invariance was detected for small manmade objects but not for faces. In a study of visuotactile invariance for object-directed actions, Oosterhof et al. [2010] located clusters of

Figure 4.

Overview of results. **(A)** Intramodal classifications and their overlaps (multimodal convergence). Red, green, and blue voxels indicate successful decoding of auditory, visual, and tactile stimuli, respectively. Arrows indicate the overlay of different maps. Overlapping regions are represented in additive hues. **(B)** Crossmodal classifications (multimodal invariance). The A–V (yellow), A–T (magenta), and V–T (cyan) invariant regions are a subset of the regions identified by the corresponding intramodal overlaps

from A (presented using the same colors). The partially transparent central brain shows the overlay of the three types of invariance, with an inset brain showing the view angle. **(C)** A schematic view of modal and supramodal information processing. A stepwise convergence of sensory information occurs at border regions of the relevant sensory cortices. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

crossmodal decoding in occipitotemporal cortex and medial premotor cortex, both bilaterally, and intraparietal sulcus and postcentral gyrus in the left hemisphere (although a cluster of nonsignificant, above-chance accuracy was also seen in the right postcentral gyrus). Our major cluster of visuotactile invariance, extending along the length of the right postcentral gyrus, seems to be a mirror image of the left postcentral cluster of Oosterhof et al. [2010].

The findings of prior studies of specific crossmodal pairs—A–V, A–T, or V–T—show partial convergence with our results. We would attribute some of the differences in lateralization and the identification of additional cortical regions to differences in study design and stimulus selection. By virtue of presenting the same set of objects in three different modalities, and performing all three types of crossmodal classification, our study provides context for the prior piecemeal findings.

Comparison to Other Approaches

A different way to detect modality-invariant information is with the semantic congruency effect. In delayed match

to sample, priming, or adaptation paradigms, brain activity is modulated by the correspondence of object identity across presentations in different modalities. Audiovisual congruency effects were reviewed by Doehrmann and Naumer [2008] and reported by van Atteveldt et al. [2010] and Doehrmann et al. [2010]; audiotactile and visuotactile congruency effects were found in studies by Kassuba et al. [2013a,2013b]. Depending on the study design and stimulus presentation characteristics, crossmodal congruency may result in either suppression or enhancement of activity. With consecutively repeated presentations of identical stimuli, the study of van Atteveldt et al. [2010] found that adaptation of congruent audiovisual letters was stronger in certain areas than adaptation of incongruent audiovisual letters. However, as alluded to in our prior study [Man et al., 2012], crossmodal decoding allows more specific conclusions than crossmodal congruency effects: decoding determines *which* of several objects was presented across modalities, whereas congruency only determines *that the same* object was presented. This latter match or mismatch signal may not necessarily contain representational information about the identity of the stimulus.

Our analytical approach mapped representations as multivariate patterns composed of both activations and deactivations, in contrast to the univariate criteria of sub-additivity, superadditivity, mean activity, or max activity. When these methods were directly compared in the same dataset, the stringent criterion of superadditivity was found to be problematic for fMRI data [Beauchamp, 2005]. Our “superadditive” map ($AVT > A + V + T$) was presented only as a useful comparison to the “multisensory” map ($AVT > R$), and in fact revealed a multimodal potentiation effect. This was an ancillary finding arising from the functional localizer scans, the main purpose of which were to generate masks for MVPA in an independent dataset.

We may also compare classification performance in our masks with performance in our searchlights. Masks were defined by univariate activation levels in the functional localizers, whereas searchlight maps were defined by information mapping. This comparison tests the hypothesis that greater activity correlates with greater representational information. This hypothesis was supported for intramodal decoding but not for crossmodal decoding. The intramodal masks consisted of the most strongly activated voxels in each unimodal contrast, and in practice, selected the early sensory cortices of each modality. Intramodal classification was successful in these sensory masks. Our crossmodal masks, identifying voxels with elevated coactivity, were for the most part unsuccessful. Therefore, modality invariance may represent a stage of information integration that is less-well correlated with evoked activity.

Multistage Sensory Convergence

Generally speaking, the finding of multimodal invariance supports a model of multiple stages of information processing in the sensory and association cortices. As sensorimotor information traverses through deeper stages of processing, modality-specific representations are bound together by systematically more abstract cross-linkages between two and possibly three sensory modalities [Fig. 4; cf. Quinn et al.’s, 2014 “mixed model”]. Our results support the suggestion by Wallace et al. [2004], who proposed that bimodal and trimodal convergence would follow the topographic principle of occurring at the borders separating their respective sensory cortices, following known patterns of anatomical connectivity [Jones and Powell, 1970].

We have previously proposed that a neuroarchitectural framework of convergent–divergent (CD) signaling among hierarchically placed cortical regions registers associations among lower-level features from various sensory modalities [Damasio, 1989a, 1989b]. Our maps of convergence and invariance among sight, sound, and touch provide a neuroanatomical grounding for the CD framework. The bottom-up convergence to form associations is reciprocated, at all levels, by top-down retroactivation of the component fea-

tures. This activity of time-locked retroactivation may provide a mechanism for CD signaling to participate in conceptual representation [Damasio, 1989c; Meyer and Damasio, 2009; Man et al., 2013]. The “empirical” philosophical position on concepts defines them as “copies or combinations of copies of perceptual representations” [Prinz, 2002]. A particular concept, reaching full bloom, will invoke the representations of its sensory features in addition to the CD region coordinating their activity [Coutanche and Thompson-Schill, 2014]. A mental concept may thus be neuroanatomically characterized as dependent on multisite coordinated activity in the multisensory brain.

In conclusion, we mapped the cortical representations of real-world objects presented in audition, vision, and touch. We found sensory representations in large sectors of the sensory and association cortices, and even in the early sensory cortices of modalities other than that of the presented stimulus. We mapped the bi- and trimodal convergences of this information among the three sensory modalities. Finally, we provided evidence for modality invariant representation of objects, identifying focal regions at the junctions of the temporal, parietal, and occipital cortices for further study.

REFERENCES

- Akama H, Murphy B, Na L, Shimizu Y, Poesio M (2012): Decoding semantics across fMRI sessions with different stimulus modalities: A practical MVPA study. *Front Neuroinformatics* 6:24
- Amedi A, Von Kriegstein K, Van Atteveldt NM, Beauchamp MS, Naumer MJ (2005): Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res* 166:559–571.
- Beauchamp M (2005): Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* 3:93–113.
- Bonner MF, Peelle JE, Cook PA, Grossman M (2013): Heteromodal conceptual processing in the angular gyrus. *NeuroImage* 71:175–186.
- Brainard DH (1997): The psychophysics toolbox. *Spat Vis* 10:433–436.
- Bremmer F, Schlack A, Shah NJ, Zafiris O, Kubischik M, Hoffmann K, Zilles K, Fink GR (2001): Polymodal motion processing in posterior parietal and premotor cortex. *Neuron* 29:287–296.
- Chang CC, Lin CJ (2011): LIBSVM A Library for Support Vector Machines. *ACM TIST* 2:27.
- Coutanche MN, Thompson-Schill SL (2014): Creating Concepts from Converging Features in Human Cortex. *Cereb Cortex*.
- Damasio A (1989a): The brain binds entities and events by multi-regional activation from convergence zones. *Neural Comput* 1:123–132.
- Damasio A (1989b): Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33:25–62.
- Damasio A (1989c): Concepts in the brain. *Mind Lang* 4:24–28.
- Doehrmann O, Naumer MJ (2008): Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Res* 1242:136–50.

- Doehrmann O, Weigelt S, Altmann CF, Kaiser J, Naumer MJ (2010): Audiovisual functional magnetic resonance imaging adaptation reveals multisensory integration effects in object-related sensory cortices. *J Neurosci* 30:3370–3379.
- Erickson LC, Heeg E, Rauschecker JP, Turkeltaub PE (2014): An ALE meta-analysis on the audiovisual integration of speech signals. *Hum Brain Mapp* 35:5587–5605.
- Etzel JA, Gazzola V, Keysers C (2008): Testing simulation theory with cross-modal multivariate classification of fMRI data. *PLoS One* 3:e3690.
- Ghazanfar AA, Schroeder CE (2006): Is neocortex essentially multisensory? *Trends Cogn Sci* 10:278–285.
- Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S (2009): PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7: 37–53.
- Hsieh P-J, Colas JT, Kanwisher N (2012): Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. *J Neurophysiol* 107:3428–3432.
- Jenkinson M, Smith S (2001): A global optimisation method for robust affine registration of brain images. *Med Image Anal* 5: 143–156.
- Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17:825–841.
- Jones EG, Powell TP (1970): An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain* 93:793–820.
- Kaplan JT, Man K, Greening S (2015): Multivariate Cross-Classification: Applying machine learning techniques to characterize abstraction in neural representations. *Front Hum Neurosci* 9: 151.
- Kassuba T, Klinge C, Hölig C, Menz MM, Ptito M, Röder B, Siebner HR (2011): The left fusiform gyrus hosts trisensory representations of manipulable objects. *NeuroImage* 56:1566–77.
- Kassuba T, Klinge C, Hölig C, Röder B, Siebner HR (2013a): Vision holds a greater share in visuo-haptic object recognition than touch. *NeuroImage* 65:59–68.
- Kassuba T, Menz MM, Röder B, Siebner HR (2013b): Multisensory interactions between auditory and haptic object recognition. *Cereb Cortex* 5:1097–1107.
- Kayser C, Petkov CI, Augath M, Logothetis NK (2005): Integration of touch and sound in auditory cortex. *Neuron* 48:373–384.
- Kriegeskorte N, Goebel R, Bandettini P (2006): Information-based functional brain mapping. *Proc Natl Acad Sci* 103:3863–3868.
- Liang M, Mouraux A, Hu L, Iannetti GD (2013): Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature Commun* 4:1979.
- Man K, Kaplan JT, Damasio A, Meyer K (2012): Sight and sound converge to form modality-invariant representations in temporoparietal cortex. *J Neurosci* 32:16629–16636.
- Man K, Kaplan JT, Damasio H, Damasio A (2013): Neural convergence and divergence in the mammalian cerebral cortex: From experimental neuroanatomy to functional neuroimaging. *J Comp Neurol* 521:4097–4111.
- Meyer K, Damasio A (2009): Convergence and divergence in a neural architecture for recognition and memory. *Trends Neurosci* 32:376–382.
- Meyer K, Kaplan JT, Essex R, Webber C, Damasio H, Damasio A (2010): Predicting visual stimuli on the basis of activity in auditory cortices. *Nat Neurosci* 13:667–668.
- Meyer K, Kaplan JT, Essex R, Damasio H, Damasio A (2011): Seeing touch is correlated with content-specific activity in primary somatosensory cortex. *Cereb Cortex* 21:2113–2121.
- Oosterhof NN, Wiggett AJ, Diedrichsen J, Tipper SP, Downing PE (2010): Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. *J Neurophysiol* 104:1077–1089.
- Oosterhof NN, Tipper SP, Downing PE (2012): Visuo-motor imagery of specific manual actions: A multi-variate pattern analysis fMRI study. *NeuroImage* 63:262–271.
- Peelen MV, Atkinson AP, Vuilleumier P (2010): Supramodal representations of perceived emotions in the human brain. *J Neurosci* 30:10127–10134.
- Pietrini P, Furey ML, Ricciardi E, Gobbini MI, Wu W-HC, Cohen L, Guazzelli M, Haxby JV (2004): Beyond sensory images: Object-based representation in the human ventral pathway. *Proc Natl Acad Sci* 101:5658–5663.
- Prinz J (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA, US: MIT Press.
- Quinn BT, Carlson C, Doyle W, Cash SS, Devinsky O, Spence C, Halgren E, Thesen T (2014): Intracranial cortical responses during visual-tactile integration in humans. *J Neurosci* 34:171–181.
- Ricciardi E, Handjaras G, Bonino D, Vecchi T, Fadiga L, Pietrini P (2013): Beyond motor scheme: A supramodal distributed representation in the action-observation network. *PLoS One* 8:e58632.
- Schroeder CE, Foxe J (2005): Multisensory contributions to low-level, “unisensory” processing. *Curr Opin Neurobiol* 15:454–458.
- Simanova I, Hagoort P, Oostenveld R, van Gerven Maj (2014): Modality-independent decoding of semantic information from the human brain. *Cereb Cortex* 24:426–34.
- Smith SM (2002): Fast robust automated brain extraction. *Hum Brain Mapp* 17:143–155.
- Smith FW, Goodale MA (2015): Decoding visual object categories in early somatosensory cortex. *Cereb Cortex* 25:1020–1031.
- Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23 Suppl 1:S208–19.
- Stilla R, Sathian K (2008): Selective visuo-haptic processing of shape and texture. *Hum Brain Mapp* 29:1123–1138.
- van Atteveldt NM, Blau VC, Blomert L, Goebel R (2010): fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci* 11:11.
- Vetter P, Smith FW, Muckli L (2014): Decoding sound and imagery content in early visual cortex. *Curr Biol* 24:1256–1262.
- Wallace MT, Ramachandran R, Stein BE (2004): A revised view of sensory cortical parcellation. *Proc Natl Acad Sci* 101:2167–2172.
- Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014): Permutation inference for the general linear model. *NeuroImage* 92:381–397.
- Woolrich MW, Ripley BD, Brady M, Smith SM (2001): Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage* 14:1370–1386.