Conducting Neuroscience to Guide the Development of AI

Jeffrey Mark Siskind
Purdue University
School of Electrical and Computer Engineering
465 Northwestern Ave.
West Lafayette, IN 47907-2035 USA
qobi@purdue.edu

Abstract

Study of the human brain through fMRI can potentially benefit the pursuit of artificial intelligence. Four examples are presented. First, fMRI decoding of the brain activity of subjects watching video clips yields higher accuracy than state-of-the-art computer-vision approaches to activity recognition. Second, novel methods are presented that decode aggregate representations of complex visual stimuli by decoding their independent constituents. Third, cross-modal studies demonstrate the ability to decode the brain activity induced in subjects watching video stimuli when trained on the brain activity induced in subjects seeing text or hearing speech stimuli and vice versa. Fourth, the time course of brain processing while watching video stimuli is probed with scanning that trades off the amount of the brain scanned for the frequency at which it is scanned. Techniques like these can be used to study how the human brain grounds language in visual perception and may motivate development of novel approaches in AI.

Introduction

A dichotomy exists between two traditional fields that study intelligence for disjoint purposes. AI attempts to engineer synthetic intelligent systems, usually without concern for how natural intelligence works. Psychology and related disciplines attempt to scientifically understand natural intelligence, usually without concern for how to replicate it in engineered systems. Here I advocate a third enterprise: studying neuroscience not for the primary goal of understanding brain function for its own sake but rather as a means for reverse-engineering brain function to serve the goal of helping AI—and the related disciplines of computer vision (CV) and NLP—design better synthetic methods. I present four examples of this. First, I present a novel comparison of state-of-the-art action-recognition methods from CV with state-of-the-art methods for decoding brain scans obtained with fMRI on the same dataset. This suggests methods for improving CV action recognition. Second, I present novel methods for recovering aggregate representations, like *Dan fold shirt on the left*, in brain-scan data, from constituent representations like *Dan*, *fold*, *shirt*, and *on the left*. Novel methods show that the brain regions involved in classifying the aggregate constitute a disjoint union of those involved in classifying the constituents. This suggests methods for decomposing brain activity into primitives to investigate the methods employed by the brain when solving a plethora of standard AI problems. Third, I present novel methods for performing cross-modal decoding of brain-scan data: training models on subjects reading text or hearing speech and then using those models to decode video, and vice versa. This suggests methods for understanding the semantic KR methods employed by the brain. Fourth, I present novel methods for probing the time-course of brain processing: how information flows through different brain regions while processing stimuli. This suggests methods for reverse-engineering the algorithms employed by the brain.

Related Work

State-of-the-art brain-activity classification involves a small number of concept classes, where the stimuli are still images of objects or orthographic presentation of nouns. Just et al. (2010) classify orthographic nouns, 5 exemplars from each of 12 classes, achieving a mean rank accuracy of 72.4% on a 1-of-60 classification task, both within and across subjects. (Note that rank accuracy differs from classification accuracy and denotes "the normalized rank of the correct label in the classifier's posterior-probability-ordered list of classes," Just et al. 2010, p. 5.) Pereira, Botvinick, and Detre (2012) reanalyze the preceding data in the context of a prior from Wikipedia and achieve a mean accuracy of 13.2% on a 1-of-12 classification task and 1.94% on a 1-of-60 classification task. Hanson and Halchenko (2009) classify still images of 2 object classes: faces and houses, and achieve an accuracy above 93% on a 1-of-2 classification task. Connolly et al. (2012) classify still images of objects, 2 instances of each of 3 classes: bugs, birds, and primates, and achieve an accuracy between 60% and 98% on a 1-of-2 within-class classification task and an accuracy between 90% and 98% on a 1-of-3 between-class classification task. Haxby et al. (2011) classify image and video stimuli cross-subject achieving between 60% and 70% between-subject accuracy on image data with 6 to 7 classes and video data with all 18s clips from *Raiders of the Lost Ark*.

Comparison of fMRI Decoding with CV

There has been significant research on action recognition within CV for two decades (see References for a sample).

This work attempts to automatically label short video clips with one of a small set of classes, typically verbs. The predominant approach is *bag of spatio-temporal visual words* (BOW; Schuldt, Laptev, and Caputo 2004). In this approach, features are extracted from the video at a subset of space-time points then pooled and vector quantized, video clips are summarized as histograms of occurrence frequency of codebook entries, and models are trained on such histograms extracted from a training set and then used to classify those extracted from unseen test video. Features used include spatio-temporal interest points (STIP; Schuldt, Laptev, and Caputo 2004) and, more recently, Dense Trajectories (Wang et al. 2011; 2013; Wang and Schmid 2013). Classification often is performed with a support-vector machine (SVM; Cortes and Vapnik 1995).

The fact that BOW methods summarize an entire video clip as a single histogram bears similarity to the standard fMRI methods that classify stimuli from a single brain volume (*i.e.*, a single 3D image of the brain). Moreover, fMRI researchers typically employ SVMs for multivariate pattern analysis (MVPA). Thus BOW methods for action recognition and MVPA methods for fMRI bear a structural similarity. We asked whether they yield similar accuracy, performing an apples-to-apples comparison between CV methods applied to short video clips and fMRI analysis methods applied to scans of subjects watching the same videos (Barbu et al. 2014). Our corpus consisted of 169 2.5s video clips, covering 6 classes *carry*, *dig*, *hold*, *pick up*, *put down*, and *walk*.

We adopted a rapid event-related experiment design (Just et al. 2010). Each of 8 runs for 8 subjects contained 48 stimulus presentations. A single brain volume was captured for each presentation. Stimuli were counterbalanced across all 6 classes within each run with 8 randomly selected videos for each of the 6 classes in each run. Each brain volume consisted of 64×64×35 voxels of dimension 3.125mm×3.125mm×3.000mm. Standard techniques (AFNI; Cox 1996) were employed to process the fMRI data, ultimately reducing the 143,360 voxels in each scan to a 4,000 element vector for within-subject analyses and 12,000 for cross-subject. Such vectors constituted samples for training and testing a linear SVM classifier that employed Linear Discriminant Dimensionality Reduction (Gu, Li, and Han 2011). We performed both within-subject and cross-subject train and test, employing leave-1-run-out and leave-1-run-and-1-subject-out cross validation.

We applied C2 (Jhuang et al. 2007), Action Bank (Sadanand and Corso 2012), Stacked ISA (Le et al. 2011), VHTK (Messing, Pal, and Kautz 2009), Cao *et al.*'s (2013) implementation of Ryoo *et al.*'s (2011) method, Cao *et al.*'s (2013) method, and an implementation of the classifier described in Wang et al. (2013) on top of the Dense Trajectories (Wang et al. 2011; 2013; Wang and Schmid 2013) feature extractor to the same dataset. These experiments employed the same leave-1-run-out cross validation. Results are shown in Fig 1. Note that all the CV systems that we tested on yield similar accuracy to the cross-subject fMRI experiments and *much* lower accuracy than the corresponding within-subject fMRI experiments.

fMRI within subject	69.7%
fMRI cross subject	34.8%
C2	47.4%
Action Bank	44.2%
Stacked ISA	46.8%
VHTK	32.5%
Ryoo *et al.*	31.2%
Cao *et al.*	33.3%
Dense Trajectories	52.3%

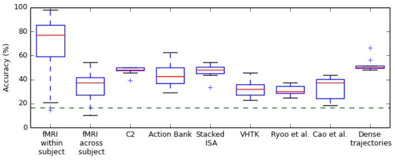


Figure 1: Results from Barbu et al. (2014). Classification accuracy of fMRI data and the 7 CV methods. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green lines indicates chance performance.

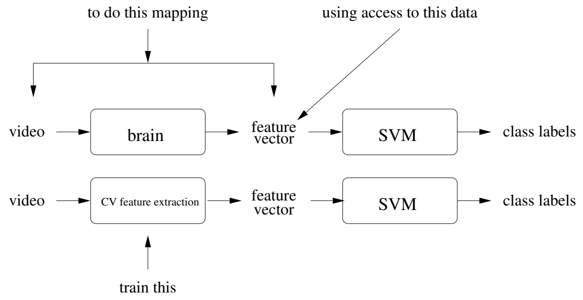


Figure 2: Suggested method for using fMRI to help design better CV algorithms.

This suggests an approach for using fMRI to help design better CV algorithms (Fig. 2). Note that both CV methods and standard fMRI analysis employ similar classifiers, usually SVMs. They differ in the feature vector. CV methods train the entire pipeline with labeled videos, taking the feature vectors to be unsupervised latent variables. FMRI can provide access to that hidden state.

Decomposing Sentences into Words

We conducted another study to determine the feasibility of decomposing brain activity elicited from complex video stimuli into constituents denoting orthogonal aspects of these stimuli. The stimuli were designed as cross products: 1 of 4 actors performing 1 of 3 verbs (*carry*, *fold*, *leave*) on 1 of 3 objects (*chair*, *shirt*, *tortilla*) in either 1 of 2 directions (*leftward*, *rightward*) or 1 of 2 locations (*on the left*, *on the right*), yielding $4 \times 3 \times 3 \times 2 = 72$ combinations.

We employed the same experiment design, capturing a single brain volume for each of 72 stimulus presentations, one for each element in the above cross product, in each run, for each of 8 runs for each of 7 subjects. We performed the same within-subject train-and-test analysis and attempted to decode both the individual constituents (words) as well as aggregates (word pairs and triples as well as entire sentences denoting the entire cross product). Aggregates were decoded using two classifiers: one trained jointly on samples depicting the particular aggregate, and the conjunction of classifiers trained independently on samples depicting each individual constituent component. Results are shown in Fig. 3.

Several things are of note. First, individual constituents, as well as aggregates (pairs, triples, and sentences) can all be

	joint	independent
actor		33.3%
verb		78.9%
object		59.8%
direction		84.6%
location		71.2%
actor-verb	25.7%	26.8%
actor-object	17.5%	20.6%
actor-direction	25.0%	28.4%
actor-location	20.3%	24.0%
verb-object	49.5%	47.9%
verb-direction	71.4%	67.1%
object-direction	39.0%	46.2%
object-location	55.1%	50.0%
actor-verb-object	14.3%	16.8%
actor-verb-direction	21.3%	23.3%
actor-object-direction	8.6%	16.3%
verb-object-direction	32.5%	36.7%
sentence		13.8%

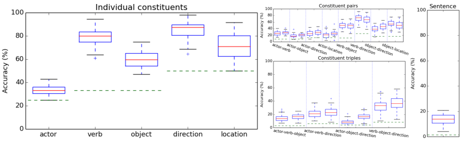


Figure 3: Results of the decomposability study. For pairs and triples, the accuracy of joint (left) *vs.* independent (right) classifiers.

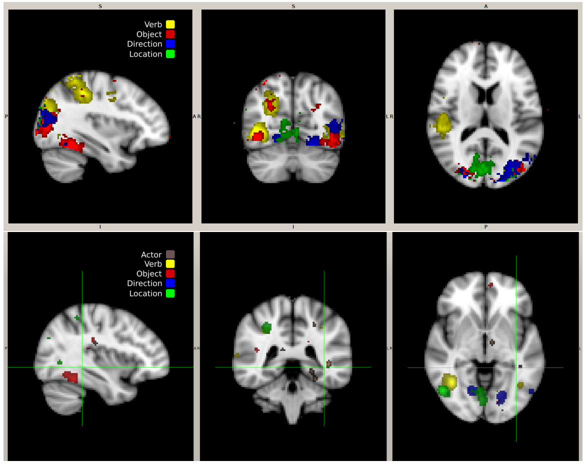


Figure 4: Brain regions for subject 1 from (top) searchlight analysis and (bottom) thresholded SVM weights.

decoded with accuracy *far* above chance. Second, the accuracy of joint classifiers is largely the same as that of the independent classifiers. This indicates both a level of commonality between different aggregate concepts that share a common constituent concept as well as a degree of independence of the processing of the constituents that combine to form the aggregate. To strengthen this analysis, we measured the correlation between the stream of individual judgments produced by both the independent and joint classifiers by computing both the accuracy and the Matthews correlation coefficient (MCC) for multi-class classification (Gorodkin 2004) on samples where the joint classifier was correct, obtaining a surprisingly high degree of correlation (Table 1). To further assess the degree of independent processing, we estimated the brain regions employed for each of the constituents using two methods: searchlight (Kriegeskorte, Goebel, and Bandettini 2006) and backprojection of thresholded SVM weights (Fig. 4 left). This was quantified with two novel analyses: computing the percentage of voxels in the union of the constituents for the independent classifier that are also in the intersection together with the percentage of voxels in the joint classifier that are shared with the independent classifier. The former measures the degree to which the constituent classifiers employ disjoint brain regions. The latter measures the degree to which the joint classifiers constitute the union of the brain regions employed by the independent classifiers. Table 1 indicates that the joint aggregate classifiers are, to a large extent, a disjoint union of the constituent classifiers.

This suggests an approach to deciphering the constituents of algorithms employed by the brain for a variety of intelligent behavior. One can scan subjects performing reasoning, planning, game-playing, language-understanding, and even motor-control tasks (subject to scanner constraints) and use the disjoint-union analysis to determine the primitives employed, how they are shared among the different tasks, and how they combine in ways specific to a particular task. Such can motivate development of new AI algorithms.

Brain Activity from Video and Text Stimuli

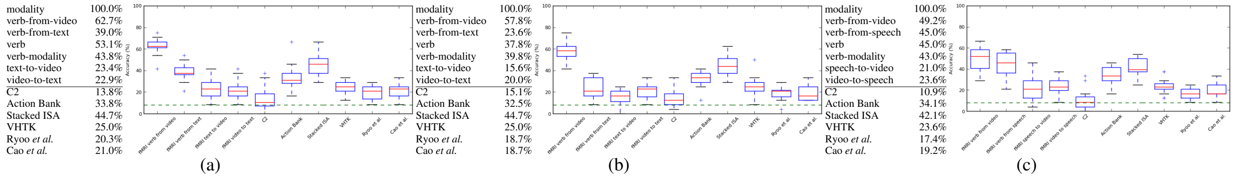
We conducted a new study, video-and-text, to assess the ability to decode verbs cross modally. We asked two questions: can we decode a larger number of classes and can we do so cross modally? For this study, we used a subset of Hollywood 2 (Marszałek, Laptev, and Schmid 2009), a dataset of movie clips with 12 classes (*AnswerPhone*, *DriveCar*, *Eat*, *FightPerson*, *GetOutCar*, *HandShake*, *HugPerson*, *Kiss*, *Run*, *SitDown*, *SitUp*, and *StandUp*) that is used within the CV community to evaluate performance of action-recognition methods. We normalized the resolution, frame rate, duration, and aspect ratio of this subset, selecting 384 2s clips, 32 for each of the 12 classes.

We employed the same experiment design, except that each of 2 subjects underwent 16 runs, each with 48 stimuli. Half were video depictions of the event classes and half were text strings naming the event classes. Each class was presented 4 times per run, twice as video and twice as text. We employed the same within-subject fMRI analysis paradigm as before, using 2,000 voxels. The same cross validation was performed by CV methods as for fMRI. We performed 7 analyses on the fMRI data: **modality** (1 of 2) determine whether the subject is looking at video or text, **verb-from-video** (1 of 12) decode the class from scans of the video stimuli, **verb-from-text** (1 of 12) decode the class from scans of the text stimuli, **verb** (1 of 12) decode the class from scans of all stimuli, **verb-modality** (1 of 24) decode the class and modality, **text-to-video** (1 of 12) decode the class from scans of the video stimuli, having been trained on scans of the text stimuli, and **video-to-text** (1 of 12) decode the class from scans of the text stimuli, having been trained on scans of the video stimuli. Results are shown in Fig. 5(a,b).

Several things are of note. First, it is possible to determine whether the subject is looking at video or text with perfect accuracy. Second, it is possible to decode 1-of-12 verbs from video stimuli with about 57.8%–62.7% accuracy. While this is lower than the accuracy obtained in the earlier study (69.7%; Fig. 1), it is still *far* above chance, on a task that has twice as many classes. Third, performance when decoding verbs from video stimuli is again higher than all of the CV methods. This replicates the earlier result with twice as many classes and with a standard dataset from the CV community. Fourth, it is possible to decode 1-of-12 verbs from text with about 23.6%–39.0% accuracy, again a level *far* above chance. (Subject 1 was a native English speaker while subject 2 was not; this might explain the lower performance on text despite similar performance on video.) Fifth, when the video and text data are pooled for both training

		actor verb	actor object	actor direction	actor location	verb object	verb direction	object direction	object location	actor verb object	actor verb direction	actor object direction	verb object direction
accuracy		0.6607	0.7059	0.6336	0.6763	0.6709	0.7422	0.6544	0.6235	0.8093	0.7403	0.8344	0.7154
MCC		0.3724	0.3430	0.3603	0.2807	0.5959	0.7048	0.5560	0.5067	0.2521	0.3004	0.2352	0.4594
$\dfrac{\left\lvert\bigcap_i \text{single}_i\right\rvert}{\left\lvert\bigcup_i \text{single}_i\right\rvert}$	searchlight	3.30%	6.74%	1.74%	1.32%	14.98%	1.20%	8.43%	4.48%	1.26%	0.06%	0.65%	0.49%
	SVM weights	2.84%	2.54%	1.16%	2.06%	6.05%	3.61%	3.70%	2.36%	0.42%	0.01%	0.00%	0.20%
$\dfrac{\left\lvert\bigcup_i \text{single}_i \cap \text{joint}\right\rvert}{\lvert\text{joint}\rvert}$	searchlight	67.42%	48.64%	68.53%	69.37%	79.53%	74.53%	65.97%	79.15%	24.48%	59.43%	37.78%	55.13%
	SVM weights	58.85%	51.22%	42.42%	27.71%	66.05%	62.38%	52.70%	38.81%	60.68%	56.51%	38.35%	58.25%

Table 1: Correlation between between independent and joint classification for constituent pairs and triples. Quantitative comparison of the brain regions employed by the independent classifiers to those employed by the joint classifiers: the percentage of voxels in the union of the constituents for the independent classifier that are also in the intersection; the percentage of voxels in the joint classifier that are shared with the independent classifier.



(a)	
modality	100.0%
verb-from-video	62.7%
verb-from-text	39.0%
verb	53.1%
verb-modality	43.8%
text-to-video	23.4%
video-to-text	22.9%
C2	13.8%
Action Bank	33.8%
Stacked ISA	44.7%
VHTK	25.0%
Ryoo *et al.*	20.3%
Cao *et al.*	21.0%

(b)	
modality	100.0%
verb-from-video	57.8%
verb-from-text	23.6%
verb	37.8%
verb-modality	39.8%
text-to-video	15.6%
video-to-text	20.0%
C2	15.1%
Action Bank	32.5%
Stacked ISA	44.7%
VHTK	25.0%
Ryoo *et al.*	18.7%
Cao *et al.*	18.7%

(c)	
modality	100.0%
verb-from-video	49.2%
verb-from-speech	45.0%
verb	45.0%
verb-modality	43.0%
speech-to-video	21.0%
video-to-speech	23.6%
C2	10.9%
Action Bank	34.1%
Stacked ISA	42.1%
VHTK	23.6%
Ryoo *et al.*	17.4%
Cao *et al.*	19.2%

Figure 5: Results for (a) subject-1 and (b) subject-2 of video-and-text and (c) subject-1 of video-and-speech.

and test, accuracy of decoding 1-of-12 verbs is about the mean of decoding such from video alone and text alone. The results are similar irrespective of whether or not the task involved simultaneous determination of modality. This is to be expected since modality alone can be decoded perfectly. Sixth, it is possible to decode 1-of-12 verbs from video stimuli using a classifier trained on text stimuli—and vice versa—with about 15.6%–23.4% accuracy. Again, this is *far* above chance and demonstrates a level of modality-independent decoding.

Brain Activity from Video and Speech Stimuli

An additional study, video-and-speech, used the same design as video-and-text except that text stimuli were replaced with speech. Only subject 1 from video-and-text was scanned. The experiment design was otherwise identical to that of video-and-text. Results are shown in Fig. 5(c).

Several things are of note. First, it is again possible to decode modality with perfect accuracy. Second, the accuracy of decoding verbs from video is again higher than all CV methods. Third, it is possible to decode verbs from speech with about 45.0% accuracy, again a level *far* above chance. Fourth, pooling video and speech data again yields the same results. Fifth, it is again possible to decode verbs from video stimuli using a classifier trained on speech stimuli—and vice versa—with about 21.0%–23.6% accuracy. Again, this is *far* above chance and demonstrates a level of modality-independent decoding.

This, together with the above method for analyzing unions and intersections of brain regions, suggests an approach for using fMRI to help decode the KR mechanisms employed by the brain. One can determine those brain regions employed while processing video, text, or speech depictions of the same concept and and compute the intersection and set differences to determine which processing is modality neutral and which is modality specify (Fig. 6).

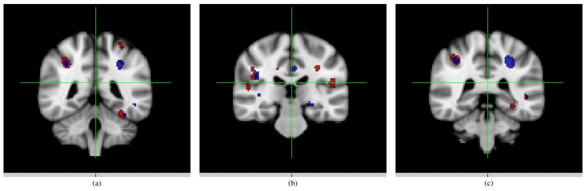


Figure 6: (a) Subset of brain regions for (a) subject 1 and (b) subject 2 in video-and-text and (c) subject 1 in video-and-speech. Red indicates video, blue indicates text/speech. Note that all 3 have roughly the same intersection of video and text/speech suggesting a potential common semantic region for verbs in video, text, and speech across subjects.

Probing the Time-Course of Processing

We conducted another study to determine the feasibility of measuring the time-course of processing by acquiring fewer slices at a higher acquisition rate. We captured 6 axial slices, instead of 35, placed in the parietal lobe a few mm below the top of the brain. This covers the supplementary motor area (SMA) and the premotor cortex: regions of high activity for video stimuli determined by backprojecting the SVM weights from earlier studies. The voxel size was the same, so overall, about 1/6 as much brain volume was covered. But the acquisition rate was more than 6 times as fast. Only video stimuli were presented to a single subject.

The same analysis pipeline was used. A classifier was trained and tested on a sequence of adjacent partial-brain

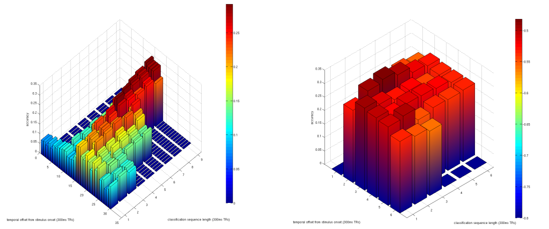


Figure 7: Accuracy as a function of temporal offset from stimulus onset and classification sequence length, both in units of 300ms, for (left) 1000 and (right) 2000 voxels.

volumes, instead of a single full-brain volume. This was done by concatenating the partial volumes into a single long vector. The sequence length was varied, as well as the offset from stimulus onset. Voxel selection and dimensionality reduction were performed on the entire concatenated (partial) volume sequence by the same method as was previously performed on a single (full) volume. Accuracy was computed for 1,000 and 2,000 voxels in the entire (partial) volume sequence (Fig. 7). In each plot, the left axes are the offset from stimulus onset to the start of the partial-volume sequence used for classification. The right axes are the length of the partial-volume sequence used for classification.

Several things are of note. First, for a fixed sequence length the function is almost unimodal. It largely increases monotonically to a single maximum and then decreases monotonically. Second, the asymptotes at either end are at about chance (1/12=8.3%). Third, the maximum for each classification sequence length is around 13–17 TRs, which is 4.2s to 5.4s after stimulus onset. Fourth, the maximum increases with increase in sequence length and peaks at about 9 TRs, which is 2.7s. With increase in sequence length, it is better to start the sequence a little earlier but it is suboptimal to keep the center of the sequence constant. It is best to increase the tail of the sequence more than the head (by about a factor of 2). Fifth, there is a *huge* difference in accuracy with change in offset as well as length. Sixth, the maximal accuracy is 31.7%, which is lower than the best obtained with a single-volume full-brain scan and analysis (45%). We can take the best analysis of sequence length 1 (19.5%) to approximate the best one can obtain with a single-volume partial-brain scan and analysis. Thus volume sequence classification can improve accuracy by 62.5% over single-volume classification.

Newer scanners can perform full-brain scans in 250ms and partial-brain scans even faster. Reprogramming the scanner with specialized pulse sequences, one could focus the scanner on specific brain regions at specific time points. Since scanners are controlled by computers, one could tightly integrate scanner control with real-time analysis of scan data to focus the scanner on brain regions exhibiting that activity most correlated with the stimuli as they are being presented. One can even adapt the stimuli in real time to elicit desired brain activity. This suggests using real-time machine-learning methods to help reverse-engineer brain function and ultimately improve AI algorithms.

Conclusion

The field of AI has experienced many debates over its history: symbolicism *vs.* connectionism, deliberation *vs.* planning, lifted *vs.* ground planning and reasoning, forward *vs.* backward chaining planning, state-space *vs.* plan-space planning, empiricism *vs.* rationalism, and determinism *vs.* stocasticism, just to name a few. While one can imagine attempting to ask which side natural intelligence is on each of these debates through traditional human-subject experiments as performed by psychologists and cognitive scientists, examining the input-output behavior of an organism might fail to tease apart the internal structure of that organism. This is the promise of methods like fMRI (and other sensor mechanisms like EEG, MEG, and PET). One can imagine testing the plausibility of a plethora of representational and algorithm choices made by AI systems, for NLP: the set of labels used by part-of-speech taggers, the tree structures used by parsers, whether parsing is top down or bottom up, the set of thematic roles, and the validity of verb classes; for CV: features such as SIFT, STIP, HOG, and HOF, delineation of object detections as axis-aligned rectangles, sliding window detectors, 2D *vs.* $2\frac{1}{2}$D and 3D representations, and segmentation and group strategies; for robotics: strategies for localization and mapping (SLAM), configuration-space path planning, and strategies for bipedal walking, just to name a few. One can systematically conduct carefully controlled experiments to search for evidence, or lack thereof, for each side of the above fundamental AI questions. The studies reported here are a first step in addressing two such questions: the neural plausibility of BOW approaches to activity recognition *vs.* time-series classifiers, and propositional joint probabilistic modeling of the grounding of language in vision *vs.* relational compositional approaches.

Acknowledgments

This research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. Imaging was performed at the Purdue MRI Facility, administered by the Weldon School of Biomedical Engineering, a core facility of the NIH-funded Indiana Clinical & Translational Science Institute. Imaging used a 3T GE Signa HDx scanner (Waukesha, Wisconsin) with a GE 8-channel head coil array for the experiments that employed audio stimuli and a Nova Medical (Wilmington, Massachusetts) 16-channel head coil array for all other experiments. A. Barbu, D. P. Barrett, Z. Burchill, W. Chen, J. J. Corso, C. D. Fellbaum, C. Hanson, S. J. Hanson, S. Hélie, E. Malaia, B. A. Pearlmutter, N. Siddharth, T. M. Talavage, G. Tamer, R. B. Wilbur, C. Xiong, and H. Yu contributed in various ways to various studies.

References

Barbu, A.; Barrett, D. P.; Chen, W.; Siddharth, N.; Xiong, C.; Corso, J. J.; Fellbaum, C. D.; Hanson, C.; Hanson, S. J.; Hélie, S.; Malaia, E.; Pearlmutter, B. A.; Siskind, J. M.; Talavage, T. M.; and Wilbur, R. B. 2014. Seeing is worse than believing: Reading people's minds better than computer-vision methods recognize actions. In *ECCV*, 612–627.

Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; and Basri, R. 2005. Actions as space-time shapes. In *ICCV*, volume 2, 1395–1402.

Cao, Y.; Barrett, D.; Barbu, A.; Narayanaswamy, S.; Yu, H.; Michaux, A.; Lin, Y.; Dickinson, S.; Siskind, J. M.; and Wang, S. 2013. Recognizing human activities from partially observed videos. In *CVPR*, 2658–2665.

Connolly, A. C.; Guntupalli, J. S.; Gors, J.; Hanke, M.; Halchenko, Y. O.; Wu, Y.-C.; Abdi, H.; and Haxby, J. V. 2012. The representation of biological classes in the human brain. *The Journal of Neuroscience* 32(8):2608–2618.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Cox, R. W. 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29(3):162–173.

Gaidon, A.; Harchaoui, Z.; and Schmid, C. 2014. Activity representation with motion hierarchies. *International Journal of Computer Vision* 107(3):219–238.

Gorodkin, J. 2004. Comparing two *K*-category assignments by a *K*-category correlation coefficient. *Computational Biology and Chemistry* 28(5):367–374.

Gu, Q.; Li, Z.; and Han, J. 2011. Linear discriminant dimensionality reduction. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 549–564.

Hanson, S. J., and Halchenko, Y. O. 2009. Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Computation* 20(2):486–503.

Haxby, J. V.; Guntupalli, J. S.; Connolly, A. C.; Halchenko, Y. O.; Conroy, B. R.; Gobbini, M. I.; Hanke, M.; and Ramadge, P. J. 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72(2):404–416.

Jhuang, H.; Serre, T.; Wolf, L.; and Poggio, T. 2007. A biologically inspired system for action recognition. In *ICCV*, 1–8.

Just, M. A.; Cherkassky, V. L.; Aryal, S.; and Mitchell, T. M. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS One* 5(1):e8622.

Kriegeskorte, N.; Goebel, R.; and Bandettini, P. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America* 103(10):3863–3868.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *ICCV*, 2556–2563.

Laptev, I. 2005. On space-time interest points. *International Journal of Computer Vision* 64(2-3):107–123.

Le, Q. V.; Zou, W. Y.; Yeung, S. Y.; and Ng, A. Y. 2011. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 3361–3368.

Li, L.-J., and Fei-Fei, L. 2007. What, where and who? Classifying events by scene and object recognition. In *ICCV*, 1–8.

Liu, J.; Kuipers, B.; and Savarese, S. 2011. Recognizing human actions by attributes. In *CVPR*, 3337–3344.

Liu, J.; Luo, J.; and Shah, M. 2009. Recognizing realistic actions from videos "in the wild". In *CVPR*, 1996–2003.

Marszałek, M.; Laptev, I.; and Schmid, C. 2009. Actions in context. In *CVPR*.

Messing, R.; Pal, C.; and Kautz, H. 2009. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 104–111.

Niebles, J. C.; Chen, C.-W.; and Fei-Fei, L. 2010. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 392–405.

Pereira, F.; Botvinick, M.; and Detre, G. 2012. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence* 194:240–252.

Rodriguez, M. D.; Ahmed, J.; and Shah, M. 2008. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 1–8.

Ryoo, M. S. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 1036–1043.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *CVPR*, 1234–1241.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local SVM approach. In *ICPR*, 32–36.

Tang, K.; Fei-Fei, L.; and Koller, D. 2012. Learning latent temporal structure for complex event detection. In *CVPR*, 1250–1257.

Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *ICCV*, 3551–3558.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR*, 3169–3176.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision* 103(1):60–79.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An approach to pose-based action recognition. In *CVPR*, 915–922.

Yamoto, J.; Ohya, J.; and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. In *CVPR*, 379–385.