Center for Brains, Minds & Machines

CBMM Memo No. 12   **September 2014**

Seeing is *Worse* than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions

by

Andrei Barbu, Daniel P. Barrett, Wei Chen, N. Siddharth, Caiming Xiong, Jason J. Corso, Christiane D. Fellbaum, Catherine Hanson, Stephen José Hanson, Sébastien Hélie, Evguenia Malaia, Barak A. Pearlmutter, Jeffrey Mark Siskind, Thomas Michael Talavage, and Ronnie B. Wilbur

Abstract

We had human subjects perform a one-out-of-six class action recognition task from video stimuli while undergoing functional magnetic resonance imaging (fMRI). Support-vector machines (SVMs) were trained on the recovered brain scans to classify actions observed during imaging, yielding average classification accuracy of 69.73% when tested on scans from the same subject and of 34.80% when tested on scans from different subjects. An apples-to-apples comparison was performed with all publicly available software that implements state-of-the-art action recognition on the same video corpus with the same cross-validation regimen and same partitioning into training and test sets, yielding classification accuracies between 31.25% and 52.34%. This indicates that one can read people's minds better than state-of-the-art computer-vision methods can perform action recognition.

Seeing is *Worse* than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions

Andrei Barbu,[1] Daniel P. Barrett,[2] Wei Chen,[3] N. Siddharth,[4] Caiming Xiong,[5] Jason J. Corso,[6] Christiane D. Fellbaum,[7] Catherine Hanson,[8] Stephen José Hanson,[8] Sébastien Hélie,[2] Evguenia Malaia,[9] Barak A. Pearlmutter,[10] Jeffrey Mark Siskind,[2] Thomas Michael Talavage,[2] and Ronnie B. Wilbur[2]

[1]MIT, Cambridge MA USA
andrei@0xab.com
[2]Purdue University, West Lafayette IN USA
{dpbarret,shelie,qobi,tmt,wilbur}@purdue.edu
[3]SUNY Buffalo, Buffalo NY USA
wchen23@buffalo.edu
[4]Stanford University, Stanford CA USA
nsid@stanford.edu
[5]University of California at Los Angeles, Los Angeles CA USA
caimingxiong@ucla.edu
[6]University of Michigan, Ann Arbor MI USA
jjcorso@eecs.umich.edu
[7]Princeton University, Princeton NJ USA
fellbaum@princeton.edu
[8]Rutgers University, Newark NJ USA
{cat,jose}@psychology.rutgers.edu
[9]University of Texas at Arlington, Arlington TX USA
malaia@uta.edu
[10]National University of Ireland Maynooth, Co. Kildare, Ireland
barak@cs.nuim.ie

Abstract. We had human subjects perform a one-out-of-six class action recognition task from video stimuli while undergoing functional magnetic resonance imaging (fMRI). Support-vector machines (SVMs) were trained on the recovered brain scans to classify actions observed during imaging, yielding average classification accuracy of 69.73% when tested on scans from the same subject and of 34.80% when tested on scans from different subjects. An apples-to-apples comparison was performed with all publicly available software that implements state-of-the-art action recognition on the same video corpus with the same cross-validation regimen and same partitioning into training and test sets, yielding classification accuracies between 31.25% and 52.34%. This indicates that one can read people's minds better than state-of-the-art computer-vision methods can perform action recognition.

Keywords: action recognition, fMRI

1   Introduction

There has been considerable recent interest in action recognition in the computer vision community. By our count, there were 236 papers related to action recognition published in major computer-vision conferences over the past three years.[1] A recent survey paper [17] reports the performance of a variety of different systems on a variety of different datasets. On some datasets with small numbers of classes (e.g. KTH [25], 6 classes; Weizmann [1], 9 classes) the best performance is perfect or near perfect. This has prompted many to conclude that action recognition with small numbers of classes is a solved problem, motivating many to work on datasets with larger numbers of classes (e.g. UCF50 [22], HMDB51 [14], and even UCF101 [26]).

Here we show that this conclusion might be premature. We present a new dataset with only six classes: *carry*, *dig*, *hold*, *pick up*, *put down*, and *walk*. Our dataset is innocuous; there is no attempt to subvert the recognition process. The actions are, arguably, easily interpretable by humans and, similar to KTH [25] and Weizmann [1], occur largely unoccluded in an outdoor setting with an uncluttered background. We applied *every* state-of-the-art, recently published action-recognition system for which code is publicly available (as well as several for which code is not publicly available) to this dataset and obtained classification accuracies between 31.25% and 52.34%. (Chance performance is 16.67%.) As a point of comparison, we showed these same videos to human subjects undergoing *functional magnetic resonance imaging* (fMRI). We trained classifiers on the brain-scan data and obtained an average within-subject classification accuracy of 69.73%. Note that as discussed in the next section, our dataset is *difficult*. While we do not have a precise human-performance baseline against which to compare the above classification accuracies, the next section does discuss human annotation that we have gathered and used to measure the level of difficulty of the corpus.

Figure 1 summarizes our experiment. We train and test state-of-the-art computer vision action-recognition software (C2 [12], Action Bank [24], Stacked ISA [16], VHTK [18], Cao's implementation [2] of Ryoo's method [23], Cao's method [2], and our own implementation of the classifier described in [28] on top of the Dense Trajectories [27–29] feature extractor) to classify video clips depicting one of six action classes and achieve accuracy of about 50%. We show the same video clips as stimuli to human subjects undergoing fMRI and train and test state-of-the-art brain-scan classifiers to classify the same six action classes and achieve accuracy of about 70%. This was an apples-to-apples comparison. Both conditions involved the same eight-fold cross-validation procedure with the same splits of data into training and test sets.

---

[1] 49 (5 oral and 44 poster) in CVPR 2011, 24 (4 oral and 20 poster) in ICCV 2011, 20 (2 oral and 18 poster) in CVPR 2012, 7 (3 oral and 4 poster) in BMVC 2012, 51 (5 oral and 46 poster) in ECCV 2012, 23 (3 oral and 20 poster) in ACCV 2012, 20 (2 oral and 18 poster) in CVPR 2013, and 42 (all poster) in ICCV 2013.
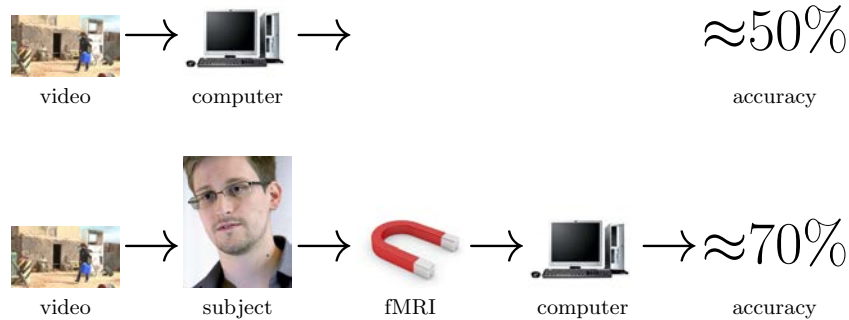
Fig. 1. A summary of our experiment. We train and test state-of-the-art computer-vision action-recognition software to classify video clips depicting one of six action classes and achieve accuracy of about 50%. We show the same video clips as stimuli to human subjects undergoing fMRI, train and test state-of-the-art brain-scan classifiers to classify the same six action classes, and achieve accuracy of about 70%.

2   Dataset

We employed a small portion of the video dataset gathered as part of the Year 2 evaluation for the DARPA Mind's Eye program.[2] (Note that we did not design the corpus or film the video ourselves; it was designed and filmed by DARPA and provided to all teams funded by the Mind's Eye program.) In particular, we used data from two components of that dataset: the portion known as `C-D2b`, which was intended to be used as training data, and the portion known as `y2-evaluation`, what was used as test data for the actual evaluation. Of `C-D2b`, we used solely the `Country_Road` portion (both `Country_Road_1` and `Country_Road_2`), videos filmed on a rural country road depicting the specified action classes. This portion contains 22 video clips ranging in length from about 13.5 minutes to about 41 minutes totaling about 8.5 hours of video. Of `y2-evaluation`, we used all of the videos employed for evaluating the 'Recognition' and 'Description' tasks that were part of the Year 2 evaluation. This portion contains 11 video clips ranging in length from about 6 minutes to about 13 minutes totaling about 2 hours of video. Two of these video clips were filmed in a country-road setting while the remainder were filmed in a 'Safe House' setting, a simulated middle-eastern urban environment. Nominally, this dataset depicts 24 distinct action classes: *approach*, *arrive*, *bury*, *carry*, *chase*, *dig*, *drop*, *enter*, *exchange*, *exit*, *flee*, *follow*, *give*, *hold*, *leave*, *pass*, *pick up*, *put down*, *replace*, *run*, *stop*, *take*, *turn*, and *walk*. However, the video is streaming; action occurrences start and stop at arbitrary points in the time course of the video, and often overlap.

There is no official ground-truth action labeling associated with this dataset. To remedy this, we had five humans annotate the entire `Country_Road` portion of `C-D2b` (both `Country_Road_1` and `Country_Road_2`) and had a different set of five annotators (with one annotator in common) annotate the entire set of videos for the Recognition and Description portions of `y2-evaluation`. Each

---

[2] `http://www.visint.org/datasets#Year_2_Videos`

Fig. 2. Key frames from sample stimuli for each of the six action classes.

annotator annotated the entire corpus portion independently, labeling each occurrence of the 24 specified action classes along with the start and end times for each occurrence. Thus we have five complete redundant annotations of the entire corpus. Having multiple annotators allows us to measure intercoder agreement, which we did for all pairs of annotators. We considered two annotated action occurrences to match when they were labeled with the same action class and temporally overlapped by a minimum specified amount. The temporal overlap was measured using a 1-dimensional variant of the 2-dimension spatial-overlap metric used in PASCAL VOC [6], namely the ratio of the length of the intersection of the two intervals to the length of their union. We then computed the F1 score for each pair of annotators as a function of overlap. The result is shown in Fig. 3. The F1 score naturally decreases monotonically with increasing minimum overlap and goes to zero when the required overlap is 100%, indicating that human annotators never agree on the precise temporal extent of the actions in question. But the F1 score ranged between 0.27 and 0.8 at 50% overlap and between 0.39 and 0.81 at 0% overlap (which still requires temporal adjacency).

This surprisingly low level of human-human intercoder agreement indicates that even in this setting where the actions are easily interpretable by humans and occur largely unoccluded in an outdoor setting with an uncluttered background, the task of delineating temporal extent of action occurrences is ambiguous. Thus we selected a subset of 6 out of the 24 action classes with the highest level of intercoder agreement: *carry, dig, hold, pick up, put down,* and *walk*. For each of these classes, we selected intervals of at least 2.5 seconds where at least two human annotators agreed on the label with at least 50% overlap. From these, we attempted to select 30 random 2.5-second clips for each of the six classes. The 2.5-second clips were chosen to maximally coincide with the intersection of
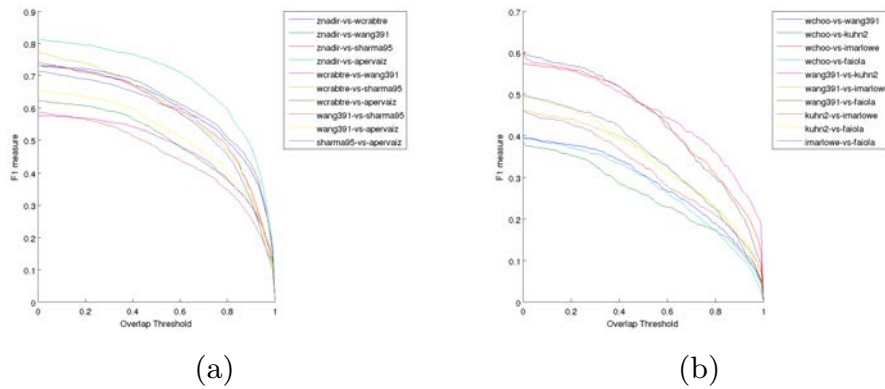
Fig. 3. Intercoder agreement for each annotator pair on (a) the `C-D2b/Country_Road` dataset and (b) the Recognition and Description portions of the `y2-evaluation` dataset that were part of the Year 2 evaluation of the DARPA Mind's Eye program, as a function of requisite temporal overlap.

the human-annotated intervals. However, two classes did not have sufficient clips with the requisite level of intercoder agreement: *dig* with 23 and *hold* with 26. Thus we selected a total of 169 distinct clips across all six action classes with the highest possible level of intercoder agreement.[3] Key frames from sample stimuli are shown in Fig. 2.

We employed a technique to further reduce the potential ambiguity in determining the intended action-class label for each stimulus. This technique was borrowed and adapted from the Natural Language Processing community. Natural language exhibits lexical polysemy: words can have multiple senses, which leads to ambiguity in contexts. WordNet [7, 19] represents word meanings with synsets, unordered sets of words that share a same meaning. A polysemous word with n different meanings occurs in n different synsets, along with its synonyms. For example, the verb *break* is found in the synsets $\{break, interrupt\}$ and $\{break, bust\}$. To further reduce the potential ambiguity in the intended class label depicted by each video, we constructed pairs of video clips with the same label, in the spirit of WordNet's synsets. In other words, we constructed longer stimuli as pairs of different video clips with the same intended action-class label, where each might otherwise be mildly ambiguous as to which action class was intended, but where together, the ambiguity is resolved. Sequences of such video-clip pairs constituted both the stimuli presented to human subjects during fMRI as well as training and test sets for computer-vision action recognition.

3   Action Recognition Software

We sought to try our corpus with as many published action-recognition methods as possible. We searched all papers on action recognition published in all conferences listed under *Computer Vision Paper Indexes*[4] since 2011, namely

---

[3] Code and data at `http://upplysingaoflun.ecn.purdue.edu/~qobi/eccv2014/`.

[4] `http://www.cvpapers.com/index.html`

ICCV 2011 and 2013, CVPR 2011, 2012, and 2013, ECCV 2012, ACCV 2012, BMVC 2012, SIGGRAPH 2011, EUROGRAPHICS 2011, and IJCAI 2011, for indication that their code was publicly available. We sought **end-to-end** implementations that included both feature extraction and classification. (Some authors release only the code for feature extraction, for example binaries for STIP [15][5] and source for Dense Trajectories [27–29][6]. The lack of a compatible released classifier makes it difficult to run and further difficult to compare with the precise published method.) The only papers that we found that indicated such were for C2 [12, 14][7] and Action Bank [24].[8] C2 is particularly relevant to our comparison with fMRI as [14] claims that it

> uses a hierarchical architecture modeled after the ventral and dorsal streams of the primate visual cortex for the task of object and action recognition, respectively.

Additionally, we posted a query for available action-recognition software to CVNet which yielded a single response pointing us to the code for Stacked ISA [16].[9] Furthermore, we contacted Rogerio Feris to see if any code was collected for the study in [17]. He pointed us to a website[10] that yielded only one available system that we hadn't already been aware of, namely Velocity Histories of Tracked Keypoints (VHTK) [18].[11] As far as we can tell, these are the only published action-recognition methods for which there are corresponding publicly available **end-to-end** implementations.

Note that the released code for Stacked ISA is only able to perform binary classification and so must differ from that used to generate the published results which include evaluation of KTH that requires multi-label classification. Also note that for VHTK, the documentation for the released code states that the released code differs from that used to produce the results in the corresponding publication; the actual code used to produce the results in the corresponding publication has not been publicly released. Thus the only publicly available systems that we are aware of that can replicate the associated published results are C2 and Action Bank.

We also have access to two action-recognition software packages that are not publicly available. Cao [2] reports that they reimplemented Ryoo's method [23] as it is not publicly available. We tested against both Cao's implementation [2] of Ryoo's method [23] as well as Cao's method [2]. Further, we implemented our own classifier using the methods described in [28] on top of the publicly available source code for the Dense Trajectories [27–29] feature extraction and tested against this as well (with 4000 GMM components).

---

[5] `http://www.di.ens.fr/~laptev/download.html`

[6] `https://lear.inrialpes.fr/people/wang/download/dense_trajectory_`
   `release_v1.2.tar.gz`

[7] `https://github.com/hueihan/Action_Recognition`

[8] `http://www.cse.buffalo.edu/~jcorso/r/actionbank/`

[9] `http://ai.stanford.edu/~quocle/video_release.tar.gz`

[10] `http://rogerioferis.com/VisualRecognitionAndSearch2014/Resources.html`

[11] `http://www.cs.rochester.edu/~rmessing/uradl/`

4 Overview of FMRI

For a general overview of fMRI see [11]. Modern 3T clinical MRI scanners can perform a high-resolution anatomical scan in about 8 minutes. This yields a spatial resolution of approximately 1mm and produces a 3D image of the brain with about 11 million voxels, with about 13 bits of information per voxel. Functional MRI trades off spatial resolution for scan time, yielding a 3D image containing about 150,000 voxels with a spatial resolution of about 3mm every two seconds. While some state-of-the-art scanners support higher-frequency functional scans about every 250ms, we do not have access to such. Thus, in our experiments, the scan time approximately coincides with the length of the video stimulus.

Most verbs describe state changes that happen over time. For example, *pick up* involves a state change of an object being at rest somewhere to being held by someone. Computer-vision methods can process frame sequences that reflect such changes. Presumably, there are also changes in brain activity to reflect such state changes in the perceived world. But they happen at a time scale that is too short to measure given the temporal resolution of fMRI. A single TR is 2s. The whole video stimulus takes 2.5s. So we get a single brain volume (after the HRF delay) that presumably reflects some smearing of the brain activity during the entire video clip but does not contain explicit information of the time course of processing. This means that while computer-vision action recognition can potentially avail itself of the temporally variant pixel values over the course of a video clip, the fMRI analysis methods we employ cannot, and process a single static brain volume for each video clip.

FMRI does not directly measure neural activity. It measures the *blood oxygenation level dependent* (BOLD) signal. Greater neural activity requires greater energy which in turn requires greater blood flow. Blood flow is shunted to different brain regions according to temporally variant neural activity. However, such shunting is delayed. It takes roughly 8–10 seconds for the BOLD response to peak after a stimulus onset that induces brain activity. Moreover, the deviation in BOLD response can persist for roughly 30 seconds after such. This is called the *hemodynamic response function* (HRF). It induces a smearing in the temporal signature of the brain activity indicated by the BOLD response; adjacent stimuli can induce overlapping BOLD response. To compensate for the HRF, we separate presentation of stimuli to the subject with blanking periods where there is no stimulus except for a fixation crosshair. Moreover, we analyze the brain-scan sample from the third TR after each stimulus, which roughly corresponds to the HRF delay.

Since the spatial resolution of a functional scan is only about 3mm, the scanning process can tolerate a small amount of subject head movement. Subject's heads are confined to a headrest in the head coil and subjects are instructed to attempt to minimize head movement. Preprocessing of BOLD involved correcting for drift, standard motion correction, and between session normalization.

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. [13] perform classification on orthographic nouns, 5

exemplars from each of 12 classes, achieving a mean rank accuracy of 72.4% on a one-out-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" [13, p. 5].) Pereira et al. [20] reanalyze the preceding data in the context of a prior from Wikipedia and achieve a mean accuracy of 13.2% on a one-out-of-12 classification task and 1.94% on a one-out-of-60 classification task. Hanson & Halchenko [9] perform classification on still images of two object classes: faces and houses, and achieve an accuracy above 93% on a one-out-of-two classification task. Connolly et al. [3] perform classification on still images of objects, two instances of each of three classes: bugs, birds, and primates, and achieve an accuracy between 60% and 98% on a one-out-of-two within-class classification task and an accuracy between 90% and 98% on a one-out-of-three between-class classification task. Haxby et al. [10] perform cross-subject classification of image and video stimuli achieving between 60% and 70% between-subject accuracy on image data with 6 to 7 classes and video data with all 18-second clips from *Raiders of the Lost Ark*. To our knowledge, this is the first study that classifies brain scans of subjects observing actions in video, and moreover compares the performance of such to computer-vision action-recognition methods.

5   FMRI Experiment

Video clips were shown to subjects who were asked to think about the action class depicted in the video during imaging. No behavioral or motor response of any kind was elicited. Specifically, subjects were not asked to push buttons or produce words, either oral or visual (written, signed). Subjects were shown sample video prior to imaging and informed of the intended set of action classes.

Because fMRI acquisition times are slow, roughly coinciding with the stimulus length, a single brain volume that corresponds to the brain activation induced by each stimulus was classified to recover the actions that the subjects were asked to think about. Multiple runs were performed for each subject, separated by several minutes, during which no stimuli were presented, no data was gathered, and subjects engaged in unrelated conversation with the experimenters. This separation between runs allowed runs to constitute folds for cross validation without introducing spurious correlation in brain activity between runs.

Imaging used a 3T GE Signa HDx scanner (Waukesha, Wisconsin) with a Nova Medical (Wilmington, Massachusetts) 16 channel brain array to collect whole-brain volumes via a gradient-echo EPI sequence with 2000ms TR, 22ms TE, 200mm×200mm FOV, and 77° flip angle. We acquired 35 axial slices with a 3.000mm slice thickness using a 64×64 acquisition matrix resulting in 3.125mm×3.125mm×3.000mm voxels.

Eight runs were acquired per subject, using a rapid event-related design [13], with stimuli counterbalanced across all six action classes within each run. We presented pairs of 2.5s video clips at 10fps, depicting the same action class. Each such presentation consisted of a 2.5s video clip, 0.5s blanking without a fixation

crosshair, a 2.5s video clip, and 0.5s of fixation, totaling 6s that was aligned to three consecutive TR boundaries. Each such was followed by a minimum of one fixation TR. Each run started with a minimum of four fixation TRs and ended with a minimum of 10 fixation TRs. An additional 48 fixation TRs were randomly placed before, between, and after video-clip-pair presentations. All such fixation TRs were aligned to TR boundaries. Each run comprised 48 presentations spanning 254 captured brain volumes. The 48 stimulus presentations constituted eight instances of each of the six action classes. The eight instances for each action class were selected randomly from a uniform distribution over the set of 23 to 30 video clips for each class. A given clip could appear more than once both within and across runs, but never within a pair. The same stimulus order, both within and across runs, was used for all subjects (and also for the computer-vision action-recognition experiments).

Scan data was gathered for eight subjects and was processed using AFNI [5] to skull-strip each volume, motion correct and detrend each run, and align each subject's runs to each other. Voxels within a run were z-scored, subtracting the mean value of that voxel for the run and dividing by its variance. Since each brain volume has very high dimension, 143,360 voxels, voxels were eliminated by computing a per-voxel Fisher score on the training set and keeping the 4,000 highest-scoring voxels (12,000 for the cross-subject analyses). The Fisher score of a voxel v for a classification task with C classes where each class c has n_c examples was computed as $\frac{\sum_{c=1}^{C} n_c (\mu_{c,v} - \mu)^2}{\sum_{c=1}^{C} n_c \sigma_{c,v}^2}$ where $\mu_{c,v}$ and $\sigma_{c,v}$ are the per-class per-voxel means and variances and μ was the mean for the entire brain volume. The resulting voxels were then analyzed with Linear Discriminant Dimensionality Reduction [8] to select a smaller number of potentially-relevant voxels, selecting on average 1,084 voxels per-subject per-fold (12,000 for the cross-subject analyses). Both stages of voxel selection were performed independently for the training set for each fold of the analysis. The set of voxels to consider was determined solely from the training set. That same subset of voxels was used in the test set for classification.

A linear support vector machine (SVM) [4] was employed to classify the selected voxels. One run was taken as the test set and the remaining runs were taken as the training set. To account for the HRF, the third brain volume after the onset of each stimulus was used for training and classification.

Two kinds of analyses were performed: within subject and cross subject. The within-subject analyses trained and tested each classifier on the same subject. In other words, classifiers were trained on the data for subject s and also tested on the data for subject s. This was repeated for all eight subjects. While we trained and tested on data from the same subject, this does *not* constitute training on the test data since *different* brain scans for *different* video clips were used for training and test. For these, leave-one-out cross validation was performed by run: when testing on run r, the classifiers were trained on all runs *except* run r. Such cross validation precludes training on the test data. Partitioning by run ensures that information could not flow from the training set to the test set through the
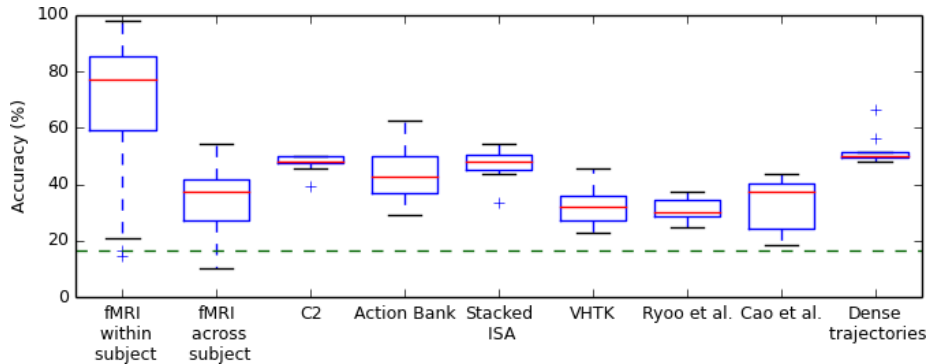
Fig. 4. Box plot corresponding to the results in Table 1, aggregated across subject and run for fMRI and aggregated across run for the computer-vision methods. Red lines indicate medians, box extents indicate upper and lower quartiles, error bars indicate maximal extents, and crosses indicate outliers. The dashed green lines indicates chance performance.

hemodynamic response function (HRF). This was repeated for all eight runs, thus performing eight-fold cross validation.

The cross-subject analyses trained and tested each classifier on different subjects. In particular, a classifier was trained on the data for all subjects except subject s and then tested on the data for subject s. This was repeated for all eight subjects. For these, leave-one-out cross validation was performed by both subject and run: when testing on run r for subject s, the classifiers were trained on all runs *except* run r for all subjects *except* subject s. While there is no potential for training on the test data, even without cross validation by run, there is potential for a different HRF-based confound. Due to the HRF, each scan potentially contains information from prior stimuli in the same run. Since the presentation order did not vary by subject, it is conceivable that classifier performance is due, in part, to the current stimulus in the context of previous stimuli in the same run, not just the current stimulus. One could control for this confound by randomizing presentation order across subject, but this was not part of the experiment design. Cross validation by run is an alternative control for this confound.

The results are presented in Table 1 and Figs. 4 and 5. All results are statistically significant with $p \leq .005$, when aggregated across subject, across run, or both, taking a binomial distribution (repeated independent Bernoulli trials with a uniform distribution over possible outcomes) as the null hypothesis. Assuming independence between trials, with each trial uniformly distributed, is warranted because all runs were counterbalanced. This demonstrates the ability to recover the action class that the subjects were thinking about when watching the video. Note that the confusion matrices are mostly diagonal, with the highest numbers of errors being made distinguishing *carry* and *hold*, *carry* and *walk* (which are both pairs of mutually ambiguous stimuli), and *pick up* and *put down*.[12]

---

[12] The instructions given to subjects delineated *carry*, which required horizontal agent motion, from *hold*, which required the agent to be stationary.

Table 1. Accuracy of within-subject and cross-subject classification of fMRI brain scans of subjects watching video clips on a 1-out-of-6 action-recognition task (chance performance is 0.1666), by subject and run, aggregated across subject, aggregated across run, and aggregated across subject and run. Comparison with seven computer-vision action-recognition methods, by run and aggregated across run.

analysis	subject	mean	stddev	run							
				1	2	3	4	5	6	7	8
fMRI within subject	1	0.7943	0.0783	0.8333	0.8125	0.8958	0.8542	0.7292	0.8125	0.7708	0.6458
	2	0.8880	0.0589	0.8750	0.9375	0.9792	0.9167	0.8958	0.7917	0.8333	0.8750
	3	0.7500	0.0568	0.7917	0.7083	0.7292	0.7500	0.7500	0.6458	0.8125	0.8125
	4	0.3828	0.0945	0.4583	0.5417	0.3750	0.3542	0.3750	0.2083	0.3750	0.3750
	5	0.9063	0.0686	0.8750	0.8542	0.9583	0.9583	0.9583	0.9583	0.9167	0.7708
	6	0.8385	0.0348	0.8750	0.8750	0.8542	0.8333	0.8125	0.8542	0.7708	0.8333
	7	0.5104	0.2260	0.1667	0.1458	0.6875	0.5417	0.6875	0.6875	0.6042	0.5625
	8	0.5078	0.1531	0.2083	0.6458	0.5208	0.6458	0.3958	0.4375	0.6042	0.6042
	mean	0.6973		0.6354	0.6901	0.7500	0.7318	0.7005	0.6745	0.7109	0.6849
	stddev		0.2171	0.3092	0.2557	0.2156	0.2061	0.2136	0.2450	0.1734	0.1694
fMRI across subject	1	0.2917	0.1045	0.2708	0.1458	0.2917	0.3750	0.3542	0.2708	0.1667	0.4583
	2	0.4141	0.0901	0.5417	0.5208	0.3750	0.3958	0.2500	0.3958	0.4167	0.4167
	3	0.3698	0.0761	0.4167	0.4375	0.2917	0.3750	0.3333	0.3125	0.2917	0.5000
	4	0.2917	0.1210	0.4167	0.2292	0.4792	0.2500	0.3958	0.1667	0.2292	0.1667
	5	0.3568	0.0550	0.3958	0.4167	0.3125	0.3333	0.3958	0.3750	0.3750	0.2500
	6	0.4036	0.0695	0.4375	0.3750	0.3333	0.3542	0.3333	0.5208	0.4792	0.3958
	7	0.3698	0.1677	0.1042	0.1042	0.4375	0.4792	0.3958	0.4375	0.5000	0.5000
	8	0.2865	0.0770	0.1458	0.2917	0.2917	0.3958	0.2708	0.2500	0.3750	0.2708
	mean	0.3480		0.3411	0.3151	0.3516	0.3698	0.3411	0.3411	0.3542	0.3698
	stddev		0.1068	0.1527	0.1475	0.0725	0.0647	0.0567	0.1135	0.1173	0.1254
C2 [12]		0.4740	0.0348	0.5000	0.4792	0.3958	0.4792	0.4583	0.5000	0.5000	0.4792
Action Bank [24]		0.4427	0.1112	0.5625	0.4583	0.2917	0.6250	0.3958	0.4792	0.3542	0.3750
Stacked ISA [16]		0.4688	0.0649	0.5208	0.5000	0.5417	0.4583	0.3333	0.5000	0.4375	0.4583
VHTK [18]		0.3255	0.0721	0.3750	0.2708	0.2708	0.3333	0.2292	0.3542	0.4583	0.3125
Ryoo's method*[23]		0.3125	0.0459	0.2500	0.2708	0.2917	0.3750	0.3333	0.2917	0.3750	0.3125
Cao's method [2]		0.3333	0.0964	0.3958	0.2292	0.2500	0.4375	0.1875	0.4167	0.3958	0.3542
Dense Trajectories [27–29]		0.5234	0.0634	0.6667	0.5625	0.5000	0.5000	0.4792	0.4792	0.5000	0.5000

*as implemented in Cao et al. [2]

As expected, the cross-subject average classification accuracy is lower than the within-subject average classification accuracy. This is because there is significant cross-subject anatomical variation. This is ameliorated to an extent, but not completely by warping the scan data to align the subjects to each other. But this process is imperfect. Few fMRI researchers perform cross-subject classification, testing classifiers trained on different subjects [10, 13, 21]. None that we are aware of do so for short video stimuli intended to be classified into object or event classes that correspond to nouns or verbs. Nonetheless, the average cross-subject classification accuracy is *far* above chance and is statistically significant.

6   Computer-Vision Action-Recognition Experiments

We applied C2 [12], Action Bank [24], Stacked ISA [16], VHTK [18], Cao's implementation [2] of Ryoo's method [23], Cao's method [2], and our own implementation of the classifier described in [28] on top of the Dense Trajectories [27–29] feature extractor to the same dataset.[13] When running Action Bank, we used the

---

[13] These experiments were analogous to the within-subject fMRI experiment. It would be meaningless to perform a computational analog of the cross-subject fMRI experiments because there would be no variation between different runs of the same program.
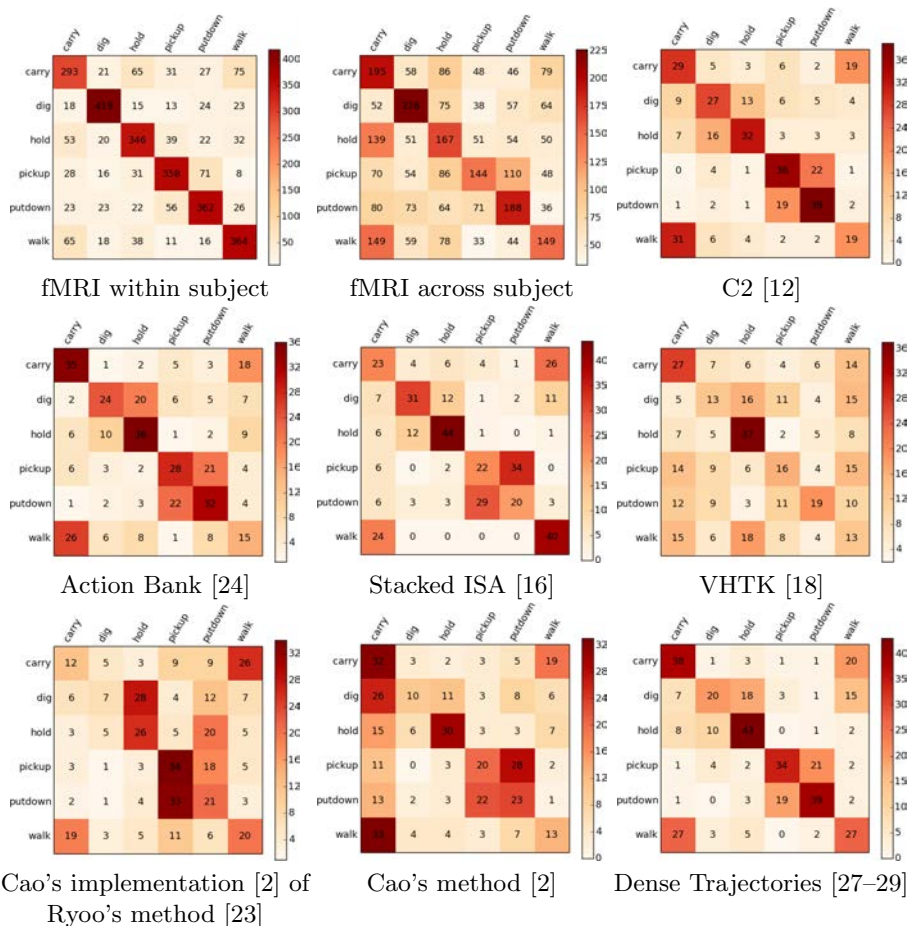
fMRI within subject

fMRI across subject

C2 [12]

Action Bank [24]

Stacked ISA [16]

VHTK [18]

Cao's implementation [2] of
Ryoo's method [23]

Cao's method [2]

Dense Trajectories [27–29]

Fig. 5. Confusion matrices corresponding to the results in Table 1, aggregated across subject and run for fMRI and aggregated across run for the computer-vision methods.

precomputed 205-template bank that was provided with the release. These experiments employed the same eight-fold leave-one-run-out cross validation. One complication arises, however. Since the stimuli were selected randomly from a uniform distribution over the set of available video clips, the same video clip could appear both within a given run and across runs. In the case of computer-vision systems, which directly process the stimuli, this would constitute training on the test data. In particular, several of the computer-vision systems that we evaluated are memory-based and would gain an unfair advantage by recalling from memory the class labels of test videos that occur in the training set. This is not a problem for the fMRI experiments because we did not directly process the stimuli; we process the brain-scan data that was evoked by the stimuli and there is significant natural variation in such.

To ameliorate this problem when performing the computer-vision experiments, we removed from each training set any pair that contained a video clip shared with a pair in the test set. This kept each test set unmodified but resulted in slightly smaller training sets. After removing such pairs, the two video clips from each pair were temporally concatenated in the same order as presented to human subjects to yield the training and test samples for the computer-vision action-recognition experiments. The results are presented in Table 1 and

Figs. 4 and 5. Note that all the computer-vision action-recognition 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.

7   Discussion

Figure 4 illustrates some interesting issues. It shows that Action Bank [24] has lower median accuracy and a higher variance profile that extends to much lower accuracy than C2 [12] and Stacked ISA [16] which predate it. It shows that Cao's implementation [2] of Ryoo's method [23] and Cao's method [2] have lower median accuracy and a much lower span of accuracies than C2 [12], Action Bank [24], and Stacked ISA [16] which predate them. It shows that Cao's method [2] has higher variance than Cao's implementation [2] of Ryoo's method [23] which predates it. Thus generally, the newer methods perform worse than the older ones; it shows that the field is basically not progressing.

Figure 5 gives some indication as to why. It shows that all the computer-vision methods tested confuse *carry* and *walk* much more than fMRI, which could be explained if these methods detected these action classes solely by detecting horizontal motion. It shows that all the computer-vision methods tested confuse *dig* and *hold*, which could be explained if these methods detected these action classes solely by detecting the lack of horizontal motion. It shows that all the computer-vision methods tested confuse *pick up* and *put down*, which could be explained if these methods detected these action classes solely by detecting vertical motion, without detecting the object being picked up or put down and without accounting for the temporal ordering of the motion. It also suggests that the semantics of human perception may play a role in action recognition, which the statistical classifiers cannot pick up. This is all to be expected when one considers that, generally, most current computer-vision methods employ techniques that look solely at local image features at very short spatial and/or temporal scales. Even Action Bank ultimately relies on local image gradients to define its templates. And none of the methods, even Dense Trajectories which can incorporate a person detector, detect the objects being interacted with as part of the action class. In other words, they don't detect the object being *carried*, the shovel used to *dig*, the hole in the ground that is *dug*, or the objects being *held*, *picked up*, or *put down*. Moreover, they don't model the time course of the changing human pose and relative position and orientation of the person and the object interacted with. These are the semantic characteristics of the action class. Thus it shows that none of these methods are, in fact, doing action recognition.

While cross-subject fMRI yields lower accuracy than within-subject fMRI, accuracy that is on par with the computer-vision methods, the confusion matrices indicate that the source of the error in the cross-subject fMRI is different than that in the computer-vision methods. There is less *pick up*:*put down* confusion, far less *dig*:*hold* confusion, and somewhat more *carry*:*walk* confusion. This indicates that even cross subject, the fMRI results appear to be using a degree of semantic inference that is absent in the computer-vision methods and the re-

duced accuracy of cross-subject fMRI is due more to issues of registration than to anything fundamental about the classification process.

8   Conclusion

Despite the explosive growth of interest in action recognition over the past three years and the perfect or near-perfect classification accuracies reported on datasets with small numbers of action classes, we show that the problem remains difficult. Uniformly, the newer methods we tried performed *no better* than or even *worse* than the older methods on this new dataset. One potential explanation is that the field as a whole is collectively overfitting to the datasets, i.e. having individual researchers repeatedly hone their methods to a small number of datasets and having the community collectively perform hill climbing on these datasets is tantamount to training on the test data. We advocate ameliorating this problem by testing methods on *read-once data*, data that has never been processed by the method. We practice what we preach by demonstrating our methods with data gathered *live* on stage during our presentations. Our ability to perform action recognition by reading minds is sufficiently robust to allow us to do something that computer-vision researchers rarely, if ever, do and neuroscientists never do, namely live demos as part of conference presentations. In the past, we have filmed live video during a talk, sent it via wireless internet to a remote imaging center, presented such video as stimuli to a subject waiting in a scanner, scanned them while watching said video, classified the brain scans, and sent the classification results back via wireless internet for live presentation. Moreover, all of the computer-vision methods we tested performed *far* worse than basic machine-learning methods applied to brain-scan data, which is surprising. We classify brain-scan data using SVMs; most computer-vision methods for action recognition do so as well. In essence, what we have done is replace the feature-extraction component with a brain-fMRI combination. This suggests that the computer-vision community may benefit by looking at neuroscience to motivate the development of better feature extraction.

References

1. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: International Conference on Computer Vision. vol. 2, pp. 1395–1402 (2005)
2. Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J.M., Wang, S.: Recognizing human activities from partially observed videos. In: Computer Vision and Pattern Recognition. pp. 2658–2665 (2013)
3. Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.C., Abdi, H., Haxby, J.V.: The representation of biological classes in the human brain. The Journal of Neuroscience 32(8), 2608–2618 (2012)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. Cox, R.W.: AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 29(3), 162–173 (1996)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
7. Fellbaum, C.: WordNet: an electronic lexical database. MIT Press, Cambridge, MA (1998)
8. Gu, Q., Li, Z., Han, J.: Linear discriminant dimensionality reduction. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 549–564 (2011)
9. Hanson, S.J., Halchenko, Y.O.: Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. Neural Computation 20(2), 486–503 (2009)
10. Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J.: A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72(2), 404–416 (2011)
11. Huettel, S.A., Song, A.W., McCarthy, G.: Functional magnetic resonance imaging. Sinauer Associates, Sunderland, MA (2004)
12. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: International Conference on Computer Vision. pp. 1–8 (2007)
13. Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M.: A neurosemantic theory of concrete noun representation based on the underlying brain codes. PloS One 5(1), e8622 (2010)
14. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: International Conference on Computer Vision. pp. 2556–2563 (2011)
15. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2-3), 107–123 (2005)
16. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Computer Vision and Pattern Recognition. pp. 3361–3368 (2011)
17. Liu, H., Feris, R., Sun, M.T.: Benchmarking datasets for human activity recognition, chap. 20, pp. 411–427. Springer (2011)
18. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: International Conference on Computer Vision. pp. 104–111 (2009)

19. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
20. Pereira, F., Botvinick, M., Detre, G.: Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artificial Intelligence 194, 240–252 (2012)
21. Poldrack, R.A., Halchenko, Y.O., Hanson, S.J.: Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychological Science 20(11), 1364–1372 (2009)
22. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Machine Vision and Applications 24(5), 971–981 (2013)
23. Ryoo, M.S.: Human activity prediction: early recognition of ongoing activities from streaming videos. In: International Conference on Computer Vision. pp. 1036–1043 (2011)
24. Sadanand, S., Corso, J.J.: Action Bank: A high-level representation of activity in video. In: Computer Vision and Pattern Recognition. pp. 1234–1241 (2012)
25. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: International Conference on Pattern Recognition. vol. 3, pp. 32–36 (2004)
26. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. Computing Research Repository abs/1212.0402 (2012)
27. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition. pp. 3169–3176 (2011)
28. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision 103(1), 60–79 (2013)
29. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision. pp. 3551–3558 (2013)