This document explains the procedure used by the Gallant lab team (Alex Huth, Mark Lescroart and Anwar Nunez-Elizalde) to label the various HCP movies. Two types of labels were created: motion-energy labels that describe the low-level structural features of the movies, and semantic-category labels that describe the high-level semantic features of the movies. Motion-energy labels were assigned using the method presented in Nishimoto et al., 2011, and semantic category (WordNet) labels were assigned using the method presented in Huth et al., 2012. The document is broken into three parts. Part 1 discusses the various label files and their contents. Part 2 explains the procedures used to produce the motion energy labels. Part 3 explains the procedures used to produce the semantic category labels.

1. HCP movie label files

Labels for the motion-energy model (Nishimoto, et al., 2011) and the semantic category model (Huth, et al., 2012) are provided in separate hdf5 files. Each file contains one feature matrix for each HCP movie. All feature matrices have a temporal resolution of 1Hz to match the 7T fMRI sampling rate (TR = 1 sec). hdf5 files can be read in python (using pytables or h5py), matlab (using h5read), or other programming languages. Note that rows and columns (time / feature dimensions) of the matrices may be transposed depending on the software used to load hdf5 files.

Each movie contains separate sections intended for estimation and validation of regression models. Thus we also provide indices to distinguish these sections within each movie. The estimation set markers (*{movie}_est)* and validation set markers (*{movie}_val)* are binary variables that indicate for each second of the movie whether that second is part of the estimation set or the validation set. These estimation and validation indicator variables exclude the first six seconds from the beginning of each clip within each movie, to account for stimulus onset responses plus a hemodynamic lag.

The exclusion of stimulus onset responses is an important consideration, for these two feature spaces as well as any others developed for these movies. Each of the four movies is itself made up of 4 or 5 movie clips. The movie clips are separated by rest periods, which are silent and nearly all black. The difference between movies and rest periods is likely to exert a much larger influence on BOLD responses than changes within the movies. The stimulus onsets, in particular, can create spurious correlations between feature spaces and may complicate model interpretation. For example, if the rest periods are not excluded, an auditory model may predict activity in visual regions, simply because visual regions also respond strongly to the non-rest periods. This problem can be mitigated by excluding the rest periods from model estimation and validation. Thus we strongly recommend that any model estimation should use the estimation and validation indices that we provide here.

MotionEnergyFeatures.hdf5

This file contains the motion energy feature matrices (* denotes a field within the hdf5 file). Each feature matrix is <n> TRs x 4025 motion energy channels, where <n> varies with the movie.
* MOVIE1_CC1: Motion energy feature matrix for 7T_MOVIE1_CC1_v2.mp4
* MOVIE2_HO1: Motion energy feature matrix for 7T_MOVIE2_HO1_v2.mp4
* MOVIE3_CC2: Motion energy feature matrix for 7T_MOVIE3_CC2_v2.mp4
* MOVIE4_HO2: Motion energy feature matrix for 7T_MOVIE4_HO2_v2.mp4
* {movie}_est: Estimation set markers for {movie}
* {movie}_val: Validation set markers for {movie}
* gabor_parameters: Parameters used for generating the motion-energy filters (See Part 2 for details)

WordNetFeatures.hdf5

This file contains binary matrices indicating when each WordNet label was present (* denotes a field within the hdf5 file). Each feature matrix is <n> TRs x 859 WordNet labels.
* MOVIE1_CC1: WordNet feature matrix for 7T_MOVIE1_CC1_v2.mp4
* MOVIE2_HO1: WordNet feature matrix for 7T_MOVIE2_HO1_v2.mp4
* MOVIE3_CC2: WordNet feature matrix for 7T_MOVIE3_CC2_v2.mp4
* MOVIE4_HO2: WordNet feature matrix for 7T_MOVIE4_HO2_v2.mp4
* synsets: A list containing the 859 WordNet synset names that correspond to each column
* {movie}_est: Estimation set markers for {movie}
* {movie}_val: Validation set markers for {movie}

For completeness, we also provide the text files containing the manually assigned labels before they were converted to the WordNet synsets. (See Part 3 for labeling details). These labels can potentially be used to develop other semantic feature spaces for the movies.

WordNet_MOVIE1_CC1.txt: Manually labeled synsets for 7T_MOVIE1_CC1_v2.mp4
WordNet_MOVIE2_HO1.txt: Manually labeled synsets for 7T_MOVIE2_HO1_v2.mp4
WordNet_MOVIE3_CC2.txt: Manually labeled synsets for 7T_MOVIE3_CC2_v2.mp4
WordNet_MOVIE4_HO2.txt: Manually labeled synsets for 7T_MOVIE4_HO2_v2.mp4

2. Motion Energy model (following Nishimoto et al 2011)

The HCP movies are about 50 minutes in total length. To extract motion-energy features from these movies they were were processed through a software pipeline like that described in Nishimoto et al (2011). Each frame was first transformed into the Commission Internationale de l'Éclairage (CIE) L*A* B* color space, and the color channels were discarded. The luminance channel was then passed through a bank of several thousand spatiotemporal Gabor filters. Each filter was defined by unique combination of position, size, orientation, motion direction, spatial and temporal frequency, and phase. Motion energy was calculated by squaring and summing pairs of Gabor filters in quadrature phase. Motion energy signals were then passed through a compressive (log) nonlinearity and temporally down-sampled to the fMRI sampling rate (1 Hz). See Nishimoto et al (2011) for further details.

Each motion energy channel was also normalized to have zero mean and unit variance (z scored). To normalize the data, preprocessed stimulus matrices for all movies were concatenated. Rest periods and periods intended to be used for model validation were removed, and the remaining (estimation) data were z scored. The means and standard deviations of the estimation movies were used to normalize the validation movies.

Note that the HCP movies were both slightly smaller and less symmetric than those used by Nishimoto et al (2011) (20ºx15º vs 20ºx20º). To ensure that the motion-energy features were as similar as possible to those used in Nishimoto et al. (2011), the maximum spatial frequency of the Gabor filters was set to be the same in both experiments (1.6 cycles/degree). This resulted in fewer motion energy channels overall for the HCP data (4,025 for HCP data vs 6,555 in Nishimoto et al).

A unique combination of the following parameters defines each motion energy feature channel:

1 - x position
2 - y position
3 - direction of motion (or orientation)
4 - spatial frequency
5 - temporal frequency
6 - Gaussian envelope size

The hdf5 field in *MotionEnergyFeatures.hdf5* labeled <gabor_parameters> contains the definition of each channel, according to the six parameters enumerated above (x and y position, direction, spatial frequency, temporal frequency, and Gaussian envelope size). All these parameters, for each channel, are provided in a 6 x 4025 matrix. The order of columns (channels) in <gabor_parameters> is consistent with the order of columns in each preprocessed stimulus matrix. The order of the rows follows the order enumerated above (1-6).

Both x and y position dimensions (first and second rows of gabor_parameters) are scaled from 0-1. To calculate the location of each Gabor filter in image pixels (from the top left of the screen), the x and y values should be multiplied by the pixel dimensions of the stimulus (x=1024, y=720). To calculate the

location of each filter in visual angle degrees (from the top left of the screen), the x and y values should be multiplied by the visual angle dimensions of the stimulus (x=20, y=15).

Direction of motion (orientation) is expressed in degrees; 0 is a vertically-oriented Gabor wavelet drifting to the right; 45 is a wavelet drifting 45º upwards and to the right. Other orientations continue similarly in counter-clockwise rotation.

Spatial frequency is expressed in cycles per degree. Zero spatial frequency means a circular Gaussian with no grating (sinusoidal) component.

Temporal frequency is expressed in Hertz. Zero temporal frequency means no temporal variation.

Gaussian envelope size is expressed as standard deviation (sigma) in the same 0-1 units as x and y position. Note that the x and y dimensions are not equal. The Gaussian envelopes are constrained to be approximately circular with the larger of the two dimensions (x) governing the size. Thus to get the 1-sigma radius of the Gaussian envelope for a given motion energy Gabor channel in degrees of visual angle, you would multiply the gabor_parameters value by the horizontal (x) size in degrees (20º).

3. Semantic category model (following Huth et al 2012)

To extract semantic categories features from the HCP movies they were labeled by hand using WordNet synsets, as described in Huth et al. (2012). Our intent was to produce a set of labels that capture the semantic object and action categories in each scene. To that end, we labeled salient object and action categories at a resolution of 1 second (that is, each 1-second clip was labeled separately).

Labels were assigned using the WordNet semantic taxonomy. This has two advantages over simple text labels. First, WordNet labels (called "synsets") are numbered to disambiguate between homographs. For example, "plant" is ambiguous, but "plant.n.01" refers to a factory building, "plant.n.02" refers to a living thing in the plant kingdom, and "plant.v.01" refers to the act of planting something in the ground. Second, WordNet includes information about *is-a* relationships between categories, which are used to supplement the manually tagged labels. For example, "wolf" is an instance of "canine", which is an instance of "carnivore", "placental mammal", "mammal", and so on. Adding these superordinate categories tends to improve the performance of encoding models by allowing poorly sampled categories to share information with their WordNet neighbors.

To ensure coverage, consistency, and accuracy of the labels, we used a multi-stage labeling procedure. First, one observer labeled all the scenes. Second, a team of 8 observers each checked and corrected the labels for 1/8 of the movies. Third, one observer (different from the first stage) did a pass through all the movies to ensure label consistency.

The following is the list of instructions to the labelers.

1. Label salient things, but also try to get complete coverage of the scene.

    Imagine that you are going to paint over every label — we want the whole scene to get covered in paint. But that doesn't mean that you need to label every little thing. If there is a shot of one person in front of a cluttered room, you don't need to label every individual object in the room. Just "person" and "room". Here's a hint for saliency: if an actor in the scene interacts with an object, it's probably salient.

2. Do not label parts of wholes (except where they are highly salient).

    If a scene contains an outside view of a house, don't label "door", "window", "roof", "eaves" and whatever else. The exception is if a part is particularly salient. For example if a scene is an outside view of a house and there is a person opening and emerging from a door, label the door. Also, if only a part of the whole is visible (e.g. only a person's hands) then only the part should be labeled.

3. Use labels consistently.

Try to look through the labels and see how they are used before you change anything. For example, video of people talking should be labeled “talk.v.02”.

4. For each labeled verb, there should be a corresponding noun (or pair of nouns, for a transitive verb).

For example if a person is cutting a shrub with shears, you’ll want to label “person”, “shear (verb)”, “shears (noun)”, and “shrub”.

5. Make labels as specific as possible.

Because we use the WordNet taxonomy to infer higher-level categories, information is gained by being as specific as possible. For example, use the label “giraffe”, not “animal”, and “male_child”, not “person”.

6. Any English word will likely map to multiple labels (synsets) in WordNet. Carefully choose the one that is most appropriate.

For instance, “eat.v.01” refers to consuming food, but “eat.v.02” refers to the act of eating a meal. Choose the most appropriate for the scene.

Additionally, when choosing a label, consider both the definition given in WordNet and the WordNet hierarchy above the label. Sometimes the definition is an incomplete description but the hierarchy gives more clues. For instance, ‘tomato’ has a separate synset for the word as part of a plant hierarchy, as well as part of a food hierarchy, so checking the hypernym paths clarifies which synset could be more appropriate for the scene.

Some WordNet synsets are specified by multi-word phrases, separated by underscores (e.g. “telephone_wire”).

7. If the 1-second clip contains a cut, only label categories that appear for 20%+ of the clip (except if those categories are not present in earlier or later clips).

For example, if a cut occurs in the final 4 frames of a 30-frame clip, do not label categories following the cut.

8. For scenes that occur indoors, include the label “room” or, if possible, a more specific label.

For example, a scene occurring in a living room should include “living room”, and the same goes for “kitchen”, and so on.

9. For scenes that include multiple humans, use the following rules:

a) For every salient person in a scene, use “person” or, if possible, a more specific label (i.e., “old_woman”, “mailman”, etc.).

b) If there is a set of two or more distinguishable but non-salient people in a scene, then use

"people".
c) If there is a set of numerous non-distinguishable and non-salient people, then use "crowd".

10. For scenes that include a motor vehicle being driven, use the following rules:
a) If you can see the driver, use the transitive verb "drive.v.01".
b) If you can see only salient passenger(s) or cargo in a moving vehicle, use "drive.v.02".
c) If you can see only the vehicle moving but not the driver, use the intransitive verb "drive.v.16".

11. Passive verbs referring to a person/people should be labeled in addition to active verbs.

For instance, if a person is in a room, synsets like "sit.v.01", "stand.v.01", "lie.v.02", etc. should be labeled as often as "sit_down.v.01", "arise.v.01", "lie_down.v.01", etc. when appropriate.

This also applies to what a salient person is focusing on with their eyes. If they are staring at something, use "gaze", but if their eyes or head are shifting focus, use "look".

Similarly, if a salient person's blinking is salient in the scene, use the verb "blink".

12. Underwater scenes should be labeled with "water.n.01".

13. If a plant type (i.e., "tree", "bush", "grass", etc.) is not clearly definable, default to "vegetation".

14. If there is text on the screen (either overlaid in post-production or in the scene) it should be marked as "text.n.01". Subtitles are marked using the additional label "subtitle.n.01". When the title of the film is shown on screen it should be marked "title.n.08". End credits of a film are marked with "credits.n.01".

References

Huth, A.G., Nishimoto, S., Vu, A., & **Gallant, J.L.** (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76, 1210-1224. [DOI: 10.1016/j.neuron.2012.10.014, PMCID: PMC3556488, NIHMSID: NIHMS418681]

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641–6. doi:10.1016/j.cub.2011.08.031