Cogn Neurodyn (2014) 8:389–397
DOI 10.1007/s11571-014-9291-3

RESEARCH ARTICLE

Encoding brain network response to free viewing of videos

Junwei Han · Shijie Zhao · Xintao Hu · Lei Guo · Tianming Liu

Received: 24 November 2013 / Revised: 2 April 2014 / Accepted: 15 April 2014 / Published online: 20 April 2014

Abstract A challenging goal for cognitive neuroscience researchers is to determine how mental representations are mapped onto the patterns of neural activity. To address this problem, functional magnetic resonance imaging (fMRI) researchers have developed a large number of encoding and decoding methods. However, previous studies typically used rather limited stimuli representation, like semantic labels and Wavelet Gabor filters, and largely focused on voxel-based brain patterns. Here, we present a new fMRI encoding model to predict the human brain's responses to free viewing of video clips which aims to deal with this limitation. In this model, we represent the stimuli using a variety of representative visual features in the computer vision community, which can describe the global color distribution, local shape and spatial information and motion information contained in videos, and apply the functional connectivity to model the brain's activity pattern evoked by these video clips. Our experimental results demonstrate that brain network responses during free viewing of videos can be robustly and accurately predicted across subjects by using visual features. Our study suggests the feasibility of exploring cognitive neuroscience studies by computational image/video analysis and provides a novel concept of using the brain encoding as a test-bed for evaluating visual feature extraction.

J. Han · S. Zhao · X. Hu (✉) · L. Guo
School of Automation, Northwestern Polytechnical University, Xi'an, China
e-mail: xintao.hu@gmail.com

T. Liu
Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, University of Georgia, Athens, GA, USA

Keywords fMRI · Encoding · Computer vision · Brain networks

Introduction

Brain encoding models provide effective means to understand how brain activity varies along with the variation in external stimuli and how well the brain activity can be predicted from the quantitatively measured external stimuli. It has been receiving increasing interest and a number of papers have been published in recent few years. Especially, several surveys by (Haynes and Rees 2006; Naselaris et al. 2011; Kay and Gallant 2009; Hasson et al. 2010; Sugase-Miyamoto et al. 2011), and Chen et al. (2014) have provided a broad overview of approaches for encoding including image analysis methodologies, functional magnetic resonance imaging (fMRI) analysis algorithms, machine learning algorithms and region of interest (ROI) selection methods and so on. An encoding model mainly consists of four components: structural substrates for brain response modeling, brain response modeling, external stimuli modeling, and the mapping from stimuli to brain response (Naselaris et al. 2011; Chen et al. 2014).

Although previous studies have yielded remarkable results, in our opinion, three problems are required to be revisited in current encoding studies. The first one is that the quantified external stimuli used in previous works are limited. In most fMRI studies [e.g. (Shirer et al. 2012; Haxby et al. 2001; Sterzer et al. 2008; Peelen et al. 2009; Mitchell et al. 2008; Nishimoto et al. 2011)], visual features were used to represent external stimuli of image/video, which includes image grid intensity, color (Naselaris et al. 2011; Nishimoto et al. 2011; Miyawaki et al. 2008), semantic category labels [e.g., (Mitchell et al. 2008)], and

participants rated scores about the external stimuli such as face and human body in a naturalistic video stream (Bartels and Zeki 2004). However, those representations are generally qualitative and subjective, and thus substantially limit the power of encoding models. To alleviate this problem, researchers tried to model the external stimuli via computational image/video descriptors. For example, Kay (Kay et al. 2008) and Naselaris (Naselaris et al. 2009) adopted Wavelet Gabor filters to model the texture feature of input image. Bartels (Bartels et al. 2008) used a motion energy model to describe visual stimuli of free viewing of movie segments. The computer vision community has developed a large amount of visual feature descriptors to represent image/video from different perspectives, for example, color, shape and motion. These features are typically objective and can be automatically derived by computer vision algorithms. However, whether those computer vision based features are feasible for fMRI encoding models has not been fully examined yet. Furthermore, in the computer vision field, the visual features are typically evaluated by conducting recognition or classification experiments based on image/video benchmarks with their human-labeled ground truth. However, this evaluation mechanism is from an engineering view without taking human brain cognition into full consideration. It is of great interest to explore the feasibility of applying brain encoding models to evaluate and compare various visual features.

The second problem is in the component of structural substrates for functional brain response modeling. The structural substrates provide the base for extracting meaningful information from fMRI data. In existing encoding models, voxel-based and ROI-based methods (Thirion et al. 2007; Polyn et al. 2005; Naselaris et al. 2011; Dumoulin and Wandell 2008; Mitchell et al. 2008) have been widely adopted. Voxels and ROIs were determined manually based on neuroscience domain knowledge or automatically based on activation detection using task-fMRI. Although voxels and ROIs-based methods are easy to implement and effective in many existing works, their reproducibility, generalizability and reliability have been limited due to the lack of a common and individualized representation of human brain architecture as pointed out in (Liu 2011; Chen et al. 2014). To be specific, voxel-based methods pose difficulties in assessing the consistency of encoding models across subjects due to the intrinsic variability of brain structure and functions and thus the lack of precise voxel-wise correspondence between subjects (Liu 2011). Recently, we developed and validated a novel data-driven strategy, namely DICCCOL (dense individualized and common connectivity-based cortical landmarks) (Zhu et al. 2012, 2013), to discover consistent and corresponding structural landmarks across various brains. In total, 358 consistent and corresponding functional landmarks were identified, each of which was optimized to possess maximal group-wise consistency of DTI-derived fiber shape patterns (Zhu et al. 2012). Moreover, this set of the 358 structural brain landmarks can be accurately and reliably predicted in a subject based only on DTI data (Zhang et al. 2012). The DICCCOL system provides an appropriate representation of human brain network and enables the opportunity of exploring the consistency of encoding and decoding brain network responses across subjects.

The third problem is in the component of brain response modeling. In previous encoding literatures (Naselaris et al. 2011; Kay et al. 2008; Miyawaki et al. 2008), fMRI Blood Oxygen-level Dependent (BOLD) intensities have been widely utilized to measure the brain's functional response. However, many literature reports (Logothetis et al. 2001; Chen et al. 2014; Heeger and Ress 2002) have pointed out that fMRI BOLD signals are often sensitive to physiological motion effect and some non-neuronal noises, which may reduce the reliability of encoding models. Another group of methods adopted the brain activation patterns measured by the GLM (general linear model) (e.g., Haxby et al. 2001; Naselaris et al. 2011; Sterzer et al. 2008; Walther et al. 2009; Mitchell et al. 2008) to construct encoding models. Recently, the results reported in the literatures (Richiardi et al. 2011; Shirer et al. 2012) suggest us that functional connectivity is a new, alternative school of methodologies for quantitatively measuring functional brain response. Essentially, brain function is resulted from large-scale functional connectivities (Haynes and Rees 2006; Lynall et al. 2010; Friston 2009; Hagmann et al. 2010). The brain's comprehension of visual stimuli can be precisely represented by these functional connectivities and interactions among relevant brain networks (Friston 2009; Hagmann et al. 2010; Lynall et al. 2010). Notably, a few recent studies (Hu et al. 2012; Han et al. 2013) have demonstrated that functional connectivity is an effective tool to model brain response to free viewing of videos.

These three above described problems motivates us to develop a novel fMRI encoding model to predict the brain's responses to free viewing of videos. The architecture of proposed encoding model is illustrated in Fig. 1. To represent the visual stimuli, we adopt a number of representative features in computer vision research included RGB histogram, color moments, Histogram of Oriented Optical Flow (HOOF) (Nayak et al. 2011) and RGB-SIFT (Van De Sande et al. 2010). To model the universal brain activity in response to video stimuli across subjects, we use the DICCCOL system (Zhu et al. 2013) to localize large-scale cortical ROIs and measured the functional connectivities among them. Afterwards, the encoding model bridging feature space and response space is trained via least-squares support vector regression (LSSVR) (Suykens

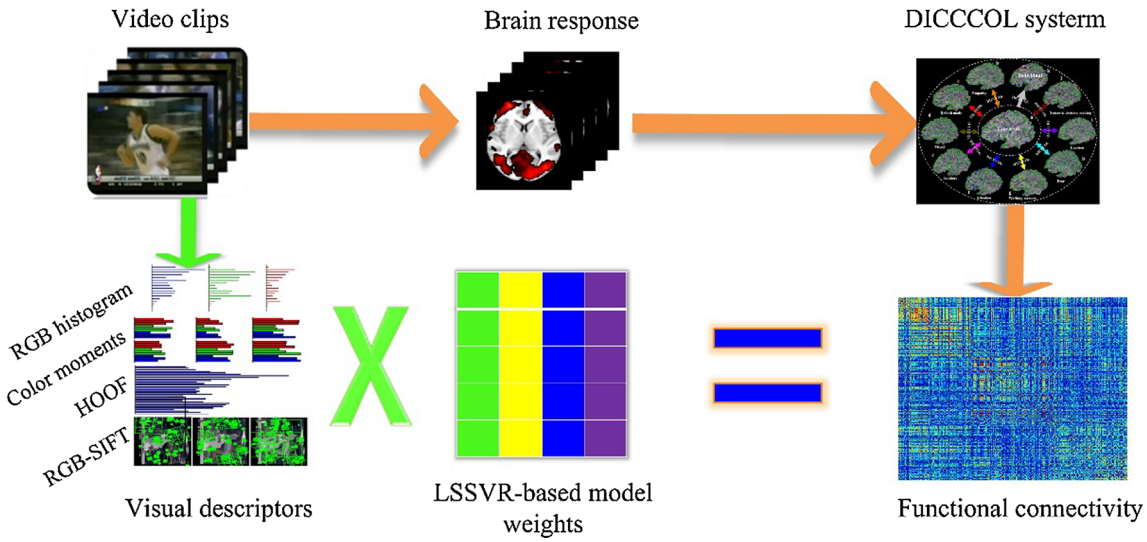

Fig. 1 The framework of the proposed encoding model. A number of representative features in computer vision research (Van De Sande et al. 2010) are adopted to model the input video stimuli and the DICCCOL system (Zhu et al. 2013) is used to localize large scale cortical ROIs, based on which the brain responses are quantified as the functional interactions among the ROIs. Afterwards, the encoding model bridging feature space and response space is trained via least-squares support vector regression (LSSVR) (Suykens and Vandewalle 1999)

and Vandewalle 1999). Experimental results demonstrated that brain network responses during free viewing of videos can be robustly and accurately predicted by those visual features and across different subjects.

The rest of this paper is organized as follows. Section "Materials and methods" describes the materials and methods adopted in this paper, including the brain response feature representation procedure and the visual feature extraction pipeline and the specifics of the proposed encoding model. The experiments and results are reported in "Experimental results" section. Finally, the discussion and conclusions are drawn in "Conclusion" section.

Materials and methods

Data acquisition and pre-processing

Subjects and stimuli

Three subjects participated in the study, which was approved by the University of Georgia IRB. All participants are young male student aged between 20 and 30 years old and they were in good health with no past history of psychiatric or neurological diseases. Participants all had normal or corrected-to-normal vision.

Natural stimulus fMRI (N-fMRI), e.g., during video watching in this paper, provides an uncontrolled environment to study the functional mechanism of the human brain. We randomly selected 51 shots including 12 commercials, 19 weather reports and 20 sports from the TRECVID 2005 data set (Smeaton et al. 2006). Each video clip is lasting 60 s or so.

MRI data acquisition

During fMRI scan, these clips were presented to these subjects via MRI-compatible goggles. The E-prime software (Schneider et al. 2002) was used for the strict synchronization between movie viewing and fMRI scan. Every participating subject took the multimodal DTI and fMRI scans in three separate scan sessions. The acquired DTI data of each participant was used to localize their DICCCOL ROIs.

Functional images were acquired on a GE 3T Signed MRI system using an 8-channel head coil at The University of Georgia Bioimaging Research Center. We set the scan parameters as follows: 30 axial slices, matrix size 64×64, 4 mm slice thickness, 220 mm FOV, TR = 1.5 s, TE = 25 ms, ASSET = 2. Diffusion tensor imaging data was also acquired for DICCCOL landmarks localization. DTI data was acquired using the isotropic spatial resolution 2 mm × 2 mm × 2 mm and the specific parameters were: TR = 15.5 s, TE = min-full, b-value = 1,000 for 30 DWIs and 3 B0 volumes.

Data preprocessing

FMRI data were preprocessed using the FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). The preprocessing of fMRI data includes skull removal, motion correction, spatial smoothing, temporal prewhitening, slice time correction,

and global drift removal. To predict ROIs for each subject based on DICCCOL system, the preprocessing of DTI data includes skull removal, motion correction and eddy current correction. Fiber tracking was performed via MEDINRIA (http://www-sop.inria.fr/asclepios/software/MedINRIA/).

Brain response modeling

Localizing reproducible and accurate cortical ROIs that are consistent and correspondent across individuals is a critical problem for brain network studies. Recently, we developed and validated a novel data-driven discovery approach that identified 358 consistent and corresponding DICCCOL ROIs in over 240 brains (Zhu et al. 2013). The intrinsic neuroscience foundation of the approach is that each brain's cytoarchitectonic area possess a unique set of extrinsic in and out, entitled the "connectional fingerprint" in (Passingham et al. 2002), which principally determines the functions of each brain area. A variety of recent studies (Laird et al. 2009; Passingham et al. 2002; Zhu et al. 2013; Zhang et al. 2012) have confirmed and replicated this close relationship between structural connection pattern and brain function. In addition, this set of 358 structural brain landmarks can be accurately and reliably predicted in an individual subject based only on DTI data (Zhang et al. 2012), demonstrating the remarkable reproducibility and predictability. Therefore, in this paper, we employ the DICCCOL system to localize dense cortical ROIs for each subject.

We first use the brain ROI prediction approach in (Zhang et al. 2012) to localize the 358 DICCCOLs in the scanned subjects with DTI data. Then, after linearly transforming the ROIs to the fMRI image space, each stimulus fMRI signals were extracted for each of these 358 DICCCOLs. Afterwards, we applied the PCA (principal component analysis) on the multiple fMRI time series within each ROI to extract a representative fMRI signal (Zhu et al. 2012). Finally, the eigenvector corresponding to the largest eigenvalue was defined as the representative fMRI signal for this ROI. With the 358 ROIs for each subject, the functional connectivity between any pair of ROIs is measured as the Pearson correlation coefficient between their N-fMRI signals, resulting in a 358×358 matrix for each video sample. Since the functional connectivity between ROIs is symmetric and the correlation between the same ROI is nonsense, we obtained 63903-dimensional functional response vector for each video sample.

Video stimuli representation

A large amount of feature descriptors have been developed and used by the computer vision community. A recent work (Van De Sande et al. 2010) reviewed a number of color descriptors commonly used in computer vision field and quantitatively evaluated them based on the accuracy of performing object and scene recognition tasks on image/video benchmarks. The work of (Nayak et al. 2011) discussed a number of state-of-the-art features used in activity recognition which are adequate to the representation of videos' motion patterns. Motivated by the work (Van De Sande et al. 2010) and (Nayak et al. 2011), in this paper, we selected four representative visual descriptor to characterize video clips, which are RGB histogram, color moments, RGB-SIFT, and HOOF. The former two descriptors measure the video color distribution while the RGB-SIFT characterize the local shape and spatial information and the HOOF describes the global motion information of the video.

RGB histogram

A 48-dimensional color histogram was extracted in RGB color space to describe the global color distribution in the video. The RGB histogram is a combination of three 1-D histograms calculated on R, G, and B channels of the RGB color space.

Color moments

Although color moments (Amir et al. 2003) of an image in the RGB space are simple to calculate, they are very effective for image/video analysis. In this paper, an image is firstly partitioned into 2 * 3 sub-blocks, and then the color moments of each block in each channel are calculated and concatenated. Similar to (Amir et al. 2003), we use three central moments which are mean, standard deviation and skewness to represent an image's color distribution. Thus, we obtained a 54 dimensional color moments descriptor for each key frame.

RGB-SIFT

The SIFT descriptor proposed in (Lowe 2004) is one of state-of-the-art techniques to characterize the local shape of a region based on edge orientation histograms derived from the gradient information. The RGB-SIFT (Van De Sande et al. 2010) calculated SIFT descriptors in the RGB color space.

HOOF

Histogram of Oriented Optical Flow (Nayak et al. 2011) is a popular and effective scale-invariant global feature to represent the motion in an entire frame using optical flow (Baker et al. 2011) in computer vision community. HOOF

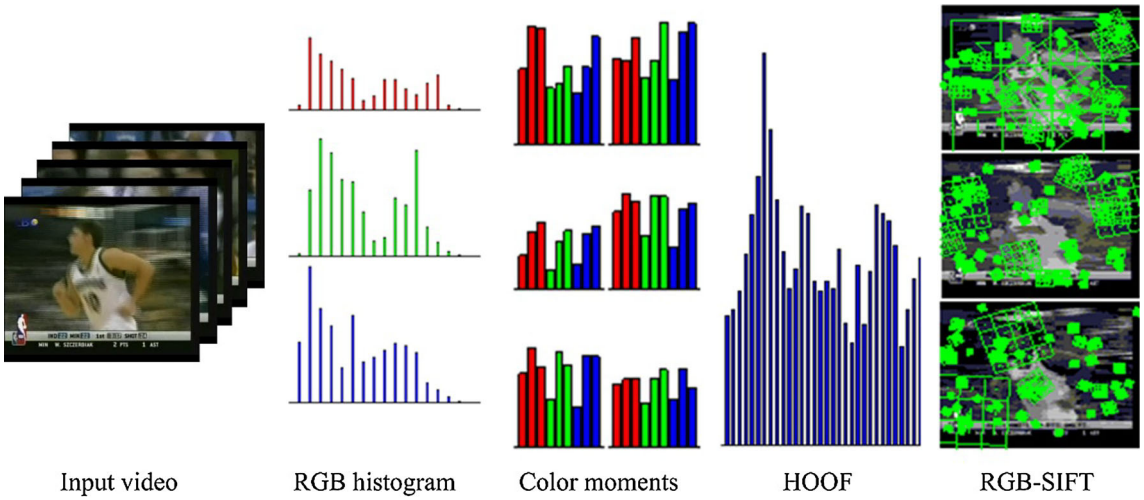

Fig. 2 The visual patterns of RGB histogram, color moments, HOOF, RGB-SIFT of an exemplar video clip. While RGB histogram and color moments describe the video color distribution, the HOOF characterizes the global motion information in the video and the RGB-SIFT characterizes the local shape and spatial information. (Color figure online)

was extracted as follows. First, optical flow was computed at every key frame. Then, optical flow vector was binned according to its primary angle and weighted based on its magnitude. The number of bins was set to 60.

TRECVID 2005 dataset provides multiple key frames for each video sample. At first, for each given key frame provided by TRECVID 2005 dataset, the above described four feature descriptors (RGB histogram, color moments, RGB-SIFT and HOOF) were calculated. Then each video sample was represented by the average of feature vectors of its all key frames. Figure 2 shows the visual patterns of each descriptor for a sample video clip. Each representation corresponds to a visual descriptor.

LSSVR-based stimuli–brain response mapping

In the current studies, the mapping between external stimuli and brain response is mainly accomplished by machine learning methods. In the early research of brain mapping, GLM models (Friston et al. 1995) were widely used to map the brain's hemodynamic responses with external stimuli due to its simplicity and effectiveness (Haxby et al. 2001; Naselaris et al. 2011; Sterzer et al. 2008; Walther et al. 2009; Mitchell et al. 2008). Additionally, researchers have explored several machine learning methods such as Gaussian Naive Bayes (GNB), SVM-based methods, and K nearest neighbor (KNN) to model the relationship between the stimuli and brain response (Naselaris et al. 2009; Mitchell et al. 2004; Walther et al. 2009). Among these methods, SVM-based approaches showed great advantages especially where there are a large number of features as the regularization in SVM-based methods help weaken the effect of noisy features which are highly correlated with each other (Pereira et al. 2009).

In our study, the least squares support vector regression algorithm (LSSVR) (Suykens and Vandewalle 1999) is adopted to solve the mapping $f(\mathbf{X} \rightarrow y)$ such that $f(\mathbf{X})$ has at most ε deviation from the actually obtained targets for all the training data, and is as flat as possible simultaneously (Suykens and Vandewalle 1999). The flatness of $f(\mathbf{X})$ ensures the superior generalizability when predicting the brain's responses from the corresponding visual features for a new video sample. The encoding model was trained for each dimension of the functional response vector independently. Denote $y = (e_{ij}^1, e_{ij}^2, \ldots, e_{ij}^n)^T$ as the set of brain's responses. e_{ij}^k is the functional connectivity between the i-th and j-th ROI in the k-th video sample. Denote $\mathbf{X} = (X_1, X_2, \ldots X_3)^T$ as the visual feature set where $X_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^T \cdot p$ is the dimensionality of the visual feature and n is the total number of training video samples. As suggested in (Naselaris et al. 2011; Pereira et al. 2009), the linear kernel was used.

Afterwards, the leave-one-out cross-validation was adopted to evaluate the performance of the trained encoding model. Each encoding model's training and testing were performed for each subject and visual feature set independently. For each video, nine sets of encoding models can be trained for the three subjects by using three visual features. Given a trained encoding model, the prediction error associated with the functional connectivity between every pair of ROIs is calculated using all video samples:

$$error_{ij} \frac{1}{n} \sum_{k=1}^{n} \left| \left(\hat{e}_{ij}^k - e_{ij}^k \right) \Big/ e_{ij}^k \right| \quad (1)$$

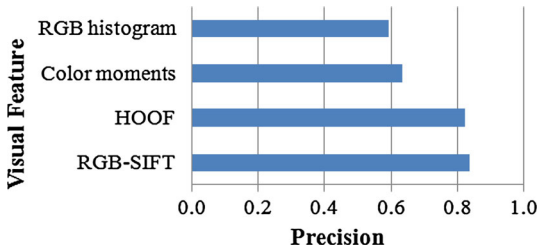

Fig. 3 Evaluation of visual descriptors in feature space. The *bar-plot* indicates the average precision of video classification in leave-one-out cross-validation using these visual features

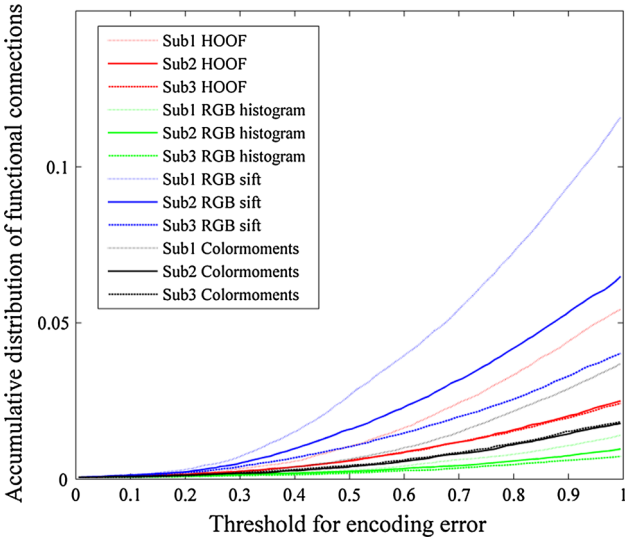

Fig. 4 Encoding accuracy results using DICCCOL: Accumulative distribution of the relatively accurate predicted connections in different feature spaces and subjects against a predefined threshold for encoding error (*error* in Eq. (1). The *x*-axis is a predefined threshold for error defined in Eq. (1) and the *y*-axis is the proportion (against all the 6,3903 functional interactions) of the functional connections with encoding error less than the error threshold

where $\hat{e}_{ij}^{k}$ is the predicted e_{ij}^{k} for the k-th video sample.

Experimental results

Evaluation of visual features in feature space

We first performed a video classification test to evaluate the distinctiveness of the proposed visual features. As for the classifiers, we adopted the K-nearest neighbor (K-NN) classifiers due to its simplicity and efficiency. The classification test was performed on those 51 video clips with visual features and the average precision in leave-one-out cross-validation was calculated. Figure 3 shows the results of different visual features which reflect their distinctiveness.

Table 1 Area under the accumulative distribution curves in Fig. 4

	RGB-SIFT	HOOF	Color moments	RGB Histogram
Subject1	0.0365	0.0220	0.0170	0.0043
Subject2	0.0213	0.0128	0.0058	0.0033
Subject3	0.0136	0.0101	0.0061	0.0026

From the Fig. 3, we can see that the RGB-SIFT perform the best, then followed by HOOF, color moments and RGB histogram respectively.

Encoding accuracy

In terms of the encoding error defined in Eq. (1), we assessed the proportion of relatively accurate predictions using different visual features and across different subjects. The accumulative histogram curves of encoding error are shown in Fig. 4. The *x*-axis in Fig. 4 is a predefined threshold for error defined in Eq. (1) and the *y*-axis is the proportion (against all the 63,903 functional interactions) of functional connections with encoding error less than the error threshold. Note that in Fig. 4, the threshold for error is only up to 100 % for the purpose of better visualization. The areas under those curves in Fig. 4 are summarized in Table 1.

Based on the results shown in Fig. 4 and Table 1, a few important points can be observed: (1) A number of the functional connections can be predicted with relatively high accuracy by the proposed encoding models. For example in the first subject, 255, 87, 124 and 74 functional connections can be predicted with error less than 20 % by RGB-SIFT, HOOF, color moments and RGB histogram, respectively. And the numbers in the second and the third subject are 328, 155, 86, 23 and 295, 174, 135, 75, respectively. (2) The number of accurately predicted functional connections is the highest by using the RGB-SIFT feature in all the three subjects, and followed by using the HOOF, color moments and RGB histogram in turn. This result may be explained by two reasons. One is that in the computer vision community it is widely accepted that RGB-SIFT which characterizes meaningful shape and spatial information of visual stimuli is more complex than other features. Thus the comprehension of RGB-SIFT may involve more brain regions and their functional interactions. The other reason is that it has also been reported in (Van De Sande et al. 2010) and validated in "Evaluation of visual features in feature space" section that the RGB-SIFT performs better than other features in recognizing objects and scenes. The distinctiveness of the feature may be an inherent factor to determine its encoding accuracy. (3) Inter-subject variation can be observed,

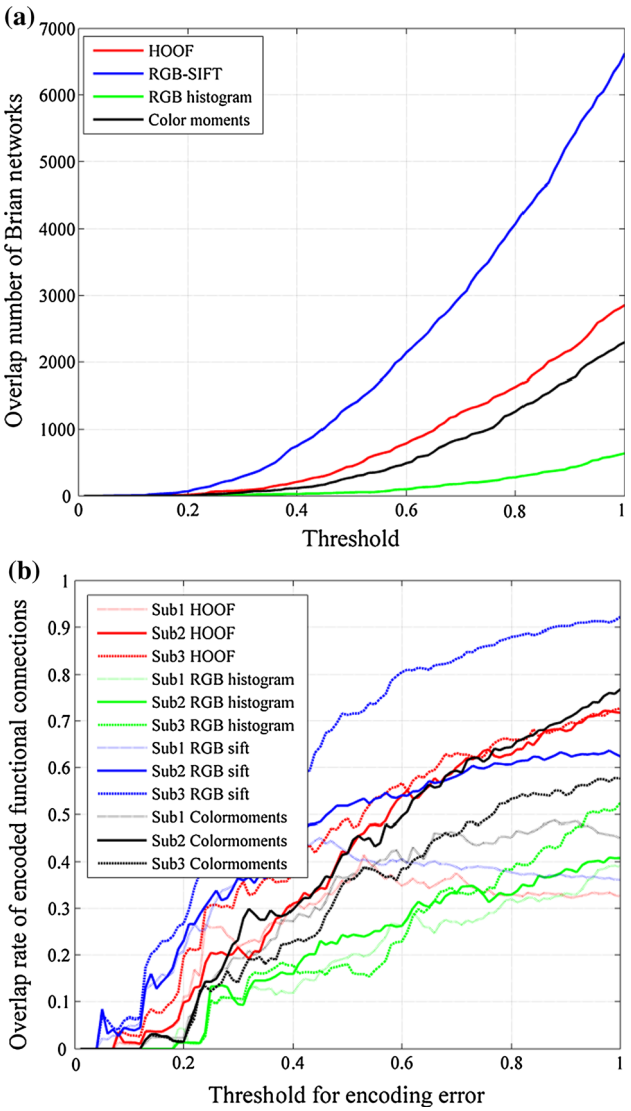

Fig. 5 Encoding consistency across different subjects. **a** The number of overlapped functional connections against a predefined threshold for *error* (Eq. 1) across the three subjects for different feature spaces. **b** The overlap ratio for different subjects and feature spaces

which may be caused by the different capabilities of the subjects in recognizing objects and scenes. We will provide more details about inter-subject consistency of the encoding model in the next subsection.

Encoding consistency across subjects

The number of correctly encoded functional connections for each visual feature is different across subject. Here, a functional connection is regarded as "correctly encoded" if its corresponding prediction error is below a predefined threshold. Then, we use the overlap ratio of correctly encoded functional connections across subjects to assess the inters-subject consistency of the encoding model for a visual feature. In this paper, we assume that two functional connections from different subjects are equivalent if both of them are in the same type of sub-network interactions. For example, two functional connections may relate to different DICCCOL ROIs in two subjects. However, if both of them are functional interactions between visual and attention system of the human brain, they are treated as "equivalent". Figure 5a shows the number of overlapped functional connections against a threshold for error across the three subjects for different feature spaces. Figure 5b shows the overlap ratio, which is calculated as the ratio between the number of overlapped functional connections and the total number of correctly encoded functional connections for a specific subject. Again, the threshold for encoding error is up to 100 % for better visualization.

From Fig. 5 we can see that the encoding model based on the RGB-SIFT shows the best inter-subject consistency followed by the ones based on HOOF, color moments and RGB histogram, especially when the encoding error is small (e.g., less than 30 %). For example, when the encoding error is less than 30 %, the average overlap rate in the four feature space is 0.3708, 0.3088, 0.1881 and 0.1020, respectively. Unlike the performance metric of encoding accuracy in subsection "Evaluation of visual features in feature space", the inter-subject consistency of the encoding models is only related to the capability of the corresponding feature set in characterizing the content of the input video stimuli. In this context, we may draw the conclusion that the RGB-SIFT outperforms HOOF, color moments and RGB histogram in describing video content from the perspective of functional brain responses prediction. Likewise, the work in (Van De Sande et al. 2010) and section "Evaluation of visual features in feature space" also demonstrated that the RGB-SIFT perform better than the other features in recognizing objects and scenes from images and videos.

Conclusion

In this paper, we proposed an fMRI encoding model to predict brain network responses to free viewing of videos. The brain responses were quantified as the functional interactions in large-scale brain networks identified by recently developed and validated DICCCOL brain landmarks localization system. The encoding model which maps the feature space to the brain response space was trained based on LSSVR. Our experimental results demonstrated that the brain network responses to video stimuli can be robustly and accurately predicted across both different feature spaces and different subjects.

Our major contributions are summarized as follows. (1) We adopted a number of representative visual features in computational vision analysis community to represent a

video sample. As mentioned before, the computational representation of stimuli in feature space is quite limited. The idea of taking advantage of visual features in computational vision community will greatly benefit the encoding and decoding studies. (2) We firstly employed the DICCCOOL system to explore the consistency of the encoding and decoding models across different subjects. The remarkable reproducibility and predictability of DICCCOLs in individual subject demonstrate great advantages of DICCCOL for inter-subject generalization. (3) Our study revealed the feasibility of using fMRI-based brain encoding techniques to evaluate visual features. Beyond neuroimaging, our results of testing visual features in encoding model are consistent with computational community which implies inherent correlations between the discriminativeness and the encoding capability of a feature.

In future, we will improve the proposed work in the following aspects. First, both the number of participants and the number of natural-stimulus video clips are relatively small. In the future, we will collect a larger scale dataset which includes more participants and uses more video clips as external stimuli, are repeat the studies proposed in this paper. Second, a number of structured visual features in computational community will be applied. Meanwhile, we will derive and test more brain response features reflecting the brain's comprehension of video stimuli. Finally, other alternative brain mapping techniques, such as sparsity constrained regression model, will be investigated and compared with the LSSVR algorithm used in this study. We believe that the combination of functional brain imaging and computational vision research will offer great benefit to both fields.

Acknowledgments J Han was supported by the National Science Foundation of China under Grant 61005018 and 91120005, NPU-FFR-JC20120237 and Program for New Century Excellent Talents in University under Grant NCET-10-0079. X Hu was supported by the National Science Foundation of China under Grant 61103061 and Program for New Century Excellent Talents in University under grant NCET-13-0472. T Liu was supported by NIH Career Award (NIH EB 006878), NSF CAREER Award (IIS-1149260), NIH R01 DA033393, NSF BME-1302089 and NIH R01 AG-042599. L Guo was supported by the National Science Foundation of China under Grants 61273362 and 61333017.

References

Amir A, Berg M, Chang S-F, Hsu W, Iyengar G, Lin C-Y et al (2003) IBM research TRECVID-2003 video retrieval system. NIST TRECVID-2003

Baker S, Scharstein D, Lewis JP, Roth S, Black MJ, Szeliski R (2011) A database and evaluation methodology for optical flow. Int J Comput Vis 92(1):1–31

Bartels A, Zeki S (2004) Functional brain mapping during free viewing of natural scenes. Hum Brain Mapp 21(2):75–85

Bartels A, Zeki S, Logothetis NK (2008) Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. Cereb Cortex 18(3):705–717

Chen M, Han J, Hu X, Jiang X, Guo L, Liu T (2014) Survey of encoding and decoding of visual stimulus via FMRI: an image analysis perspective. Brain Imaging Behav 8(1):7–23

Dumoulin SO, Wandell BA (2008) Population receptive field estimates in human visual cortex. Neuroimage 39(2):647–660

Friston KJ (2009) Modalities, modes, and models in functional neuroimaging. Science 326(5951):399

Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SCR, Frackowiak RSJ et al (1995) Analysis of fMRI time-series revisited. Neuroimage 2(1):45–53

Hagmann P, Cammoun L, Gigandet X, Gerhard S, Ellen Grant P, Wedeen V et al (2010) MR connectomics: principles and challenges. J Neurosci Methods 194(1):34–45

Han J, Ji X, Hu X, Zhu D, Li K, Jiang X et al (2013) Representing and retrieving video shots in human-centric brain imaging space. IEEE Trans Image Process 22(7):2723–2736

Hasson U, Malach R, Heeger DJ (2010) Reliability of cortical activity during natural stimulation. Trends Cogn Sci 14(1):40–48

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293(5539): 2425–2430

Haynes J-D, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7(7):523–534

Heeger DJ, Ress D (2002) What does fMRI tell us about neuronal activity? Nat Rev Neurosci 3(2):142–151

Hu X, Li K, Han J, Hua X, Guo L, Liu T (2012) Bridging the semantic gap via functional brain imaging. IEEE Trans Multimed 14(2):314–325

Kay KN, Gallant JL (2009) I can see what you see. Nat Neurosci 12(3):245

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452(7185):352–355

Laird AR, Eickhoff SB, Kurth F, Fox PM, Uecker AM, Turner JA et al. (2009). ALE meta-analysis workflows via the brainmap database: progress towards a probabilistic functional brain atlas. Front Neuroinform 3:23

Liu T (2011) A few thoughts on brain ROIs. Brain Imaging Behav 5(3):189–202

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412(6843):150–157

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

Lynall M-E, Bassett DS, Kerwin R, McKenna PJ, Kitzbichler M, Muller U et al (2010) Functional connectivity and brain networks in schizophrenia. J Neurosci 30(28):9477–9487

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M et al (2004) Learning to decode cognitive states from brain images. Mach Learn 57(1–2):145–175

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA et al (2008) Predicting human brain activity associated with the meanings of nouns. Science 320(5880): 1191–1195

Miyawaki Y, Uchida H, Yamashita O, Sato M-A, Morito Y, Tanabe HC et al (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60(5):915–929

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63(6):902–915

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56(2):400–410

Nayak N, Sethi R, Song B, Roy-Chowdhury A (2011). Motion pattern analysis for modeling and recognition of complex human activities. Guide to Video Analysis of Humans: Looking at People. New York, Springer-Verlag, 289–310

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21(19):1641–1646

Passingham RE, Stephan KE, Kotter R (2002) The anatomical basis of functional localization in the cortex. Nat Rev Neurosci 3(8):606–616

Peelen MV, Fei-Fei L, Kastner S (2009) Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460(7251):94–97

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45(1):S199–S209

Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. Science 310(5756):1963–1966

Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D (2011) Decoding brain states from fMRI connectivity graphs. Neuroimage 56(2):616–626

Schneider W, Eschman A, Zuccolotto A (2002). E-Prime reference guide: Psychology Software Tools, Incorporated

Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD (2012) Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb Cortex 22(1):158–165

Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, ACM, pp 321–330

Sterzer P, Haynes, J-D, Rees G (2008). Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. J Vis 8(15):10

Sugase-Miyamoto Y, Matsumoto N, Kawano K (2011) Role of temporal processing stages by inferior temporal neurons in facial recognition. Front Psychol 2:141

Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300

Thirion B, Pinel P, Mériaux S, Roche A, Dehaene S, Poline J-B (2007) Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. Neuroimage 35(1):105–120

Van De Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596

Walther DB, Caddigan E, Fei-Fei L, Beck DM (2009) Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci 29(34):10573–10581

Zhang T, Guo L, Li K, Jing C, Yin Y, Zhu D et al (2012) Predicting functional cortical ROIs via DTI-derived fiber shape models. Cereb Cortex 22(4):854–864

Zhu D, Li K, Faraco CC, Deng F, Zhang D, Guo L et al (2012) Optimization of functional brain ROIs via maximization of consistency of structural connectivity profiles. Neuroimage 59(2):1382–1393

Zhu D, Li K, Guo L, Jiang X, Zhang T, Zhang D et al (2013) DICCCOL: dense individualized and common connectivity-based cortical landmarks. Cereb Cortex 23(4):786–800