Representing and Retrieving Video Shots in Human-Centric Brain Imaging Space

Junwei Han, *Member, IEEE*, Xiang Ji, Xintao Hu, Dajiang Zhu, Kaiming Li, Xi Jiang, Guangbin Cui, Lei Guo, and Tianming Liu, *Senior Member, IEEE*

***Abstract*— Meaningful representation and effective retrieval of video shots in a large-scale database has been a profound challenge for the image/video processing and computer vision communities. A great deal of effort has been devoted to the extraction of low-level visual features, such as color, shape, texture, and motion for characterizing and retrieving video shots. However, the accuracy of these feature descriptors is still far from satisfaction due to the well-known semantic gap. In order to alleviate the problem, this paper investigates a novel methodology of representing and retrieving video shots using human-centric high-level features derived in brain imaging space (BIS) where brain responses to natural stimulus of video watching can be explored and interpreted. At first, our recently developed dense individualized and common connectivity-based cortical landmarks (DICCCOL) system is employed to locate large-scale functional brain networks and their regions of interests (ROIs) that are involved in the comprehension of video stimulus. Then, functional connectivities between various functional ROI pairs are utilized as BIS features to characterize the brain's comprehension of video semantics. Then an effective feature selection procedure is applied to learn the most relevant features while removing redundancy, which results in the formation of the final BIS features. Afterwards, a mapping from low-level visual features to high-level semantic features in the BIS is built via the Gaussian process regression (GPR) algorithm, and a manifold structure is then inferred, in which video key frames are represented by the mapped feature vectors in the BIS. Finally, the manifold-ranking algorithm concerning the relationship among all data is applied to measure the similarity between key frames of video shots. Experimental results on the TRECVID 2005 dataset demonstrate the superiority of the proposed work in comparison with traditional methods.**

***Index Terms*— Brain imaging space, functional magnetic resonance imaging, Gaussian process regression, video shot retrieval.**

I. Introduction

With the explosive growth of digital video data, efficient management and retrieval of large-scale video databases have become increasingly important in recent years. A typical content-based video retrieval system involves a series of key techniques such as video structural analysis, feature representation, and similarity measurement. A video generally consists of sequences or stories, which are composed of numerous scenes [1]. Scenes can be further parsed into a set of shots. As the physical basic layer in video, a shot contains a number of frames describing a continuous action. Shot-based retrieval serves as the basis for video retrieval [2], [3], where two crucial issues need to be solved: how to meaningfully represent shots and how to measure similarity between shots. Most existing approaches extract key frames from each shot and measure the similarity between shots based on their key frames. Key frames are still images which can reasonably represent the content of shots in an abstracted manner.

The capability of current methodologies of video representation and retrieval is rather limited due to the insurmountable gap between low-level features used by machines and high-level semantics perceived by the human brain's cognitive systems. To alleviate this problem, first, many researchers attempted to design sophisticated feature descriptors that are more accurate and richer to describe visual content such as SIFT [4], SURF [5], and Bag-of-Words (BoW) [6]. The second line of research is to design biologically plausible features that can mimic human vision perception mechanisms, for example, cortex-like features and visual attention-based features [7], [8]. The third school of approaches applied supervised learning algorithms to select the most relevant or discriminative features from the feature bank with the motivation of keeping the human in the loop. In these methods, human subjects can manually offer classification labels, preferences, and ranks to visual data [9]. Therefore, the semantic gap may be narrowed by taking the advantage of human's guidance. Nevertheless, this type of human guidance generally can only provide subjective, rough, and sometimes ambiguous and incomplete

Manuscript received February 24, 2012; revised August 6, 2012 and March 18, 2013; accepted March 18, 2013. Date of publication April 4, 2013; date of current version May 16, 2013. The work of T. Liu was supported in part by the NIH Career Award EB 006878, Award NIH R01 HL087923-03S2, and Award NIH R01 DA033393, the NSF CAREER Award IIS-1149260, and The University of Georgia Start-Up Research Funding. The work of J. Han and X. Hu was supported in part by the National Science Foundation of China under Grants 61005018, 91120005, 61103061, NPU-FFR-JC20120237, and NCET-10-0079, and the Post-Doctoral Foundation of China under Grants 20110490174 and 2012T50819. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu.

J. Han, X. Ji, X. Hu, and L. Guo are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com; xiangji123@gmail.com; xintao.hu@gmail.com; lguo@nwpu.edu.cn).

D. Zhu, X. Jiang, and T. Liu are with the Department of Computer Science, The University of Georgia, Athens, GA 30602-8001 USA (e-mail: dajiang.zhu@gmail.com; superjx2318@gmail.com; tianming.liu@gmail.com).

K. Li is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Department of Computer Science, University of Georgia, Athens, GA 30602-8001 USA (e-mail: likaiming@gmail.com).

G. Cui is with the Department of Radiology, Tangdu Hospital, The Fourth Military Medical University, Xi'an 710032, China (e-mail: cgbtd@yahoo.com.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2256919

information, which may lead to a weak learning performance. Moreover, traditional computational methods rely on data mining technique and totally ignore the human brain behaviors in visual understanding, and thus are incapable of leveraging the intrinsic mechanisms of human brain perception and cognition. The limited space in this paper does not allow us to extensively review existing work. A systemic summary of techniques used in video retrieval can be found in a recent survey [10].

Essentially, human brains are the end users and evaluators of video content and representation, and quantitative modeling of the interactions between video streams and the brain's responses can provide meaningful guidance for video representation and retrieval. In recognition of this potential, in recent years, exploring the functional interactions among brain networks under the natural stimulus of watching video and applying neuroscience principles to deeply understand semantics of visual content have become a newly arising trend, which is potentially promising to offer a superior mechanism to bridge the semantic gaps. For instance, the work of [11], [12] proposed an electroencephalography (EEG)-based brain machine interface system to achieve image annotation and retrieval where human brain responses while viewing images captured by an EEG and image visual features are combined. In [13], Kapoor *et al*. adopted the Pyramid Match Kernel to integrated image visual features and EEG responses from various subject brains for image object categorization.

However, in-vivo EEG usually collects brain signals via electrodes placed around human scalp, resulting in limited spatial resolutions and inability to capture the full-length comprehensive semantics of brain responses to videos. In contrast, functional magnetic resonance imaging (fMRI) is a powerful tool to probe and monitor human full-brain activity for cognition. For example, the milestone work in [14] has demonstrated that there are high temporal correlations between relevant fMRI signals and semantic content in the movie stream, which has provided strong evidence that fMRI time series data is potentially able to model the functional interaction between the human brain and multimedia information. In [15], Walther *et al*. performed pattern analysis using fMRI data to study which set of regions of the brain can differentiate natural scene categories. The work of [16] explored the decoding method based on correlations between visual stimuli and fMRI activity in early visual areas for image identification. Recently, Hu *et al*. [17], [18] developed a video classification model by correlating fMRI-derived brain responses and low-level features by using the PCA-CCA algorithm. Li *et al*. [19] built a human-centric video summarization framework via the optimization of attentional models by using fMRI data as the benchmark.

In general, fMRI-derived functional brain activity in response to video stimuli can quantitatively, objectively, and effectively reflect the brain's comprehension of video content. In BIS, we are able to look into the functional interactions among relevant brain networks involved in video comprehension, and thus derive a wealth of high-level semantics. This motivates us to develop a generalized human-centric video representation and retrieval framework based on brain cognition related features inferred in BIS. In this paper, we extensively extended our preliminary work in [20] and designed a novel computational framework as illustrated in Fig. 1. The proposed computational framework is composed of two components: video shot representation and video shot retrieval. To represent video shots in the BIS, we firstly randomly select a subset of video shots from a large-scale video database and use them as the natural stimulus for fMRI scanning when the human subjects are watching video streams. Afterwards, 358 consistent and dense brain ROIs defined by our DICCCOL system [21] are located in each subject's brain via our brain ROI prediction methods [22]. The relevant fMRI signals associated with 358 ROIs are then acquired. Subsequently, functional connectivities between various DICCCOL ROI pairs are utilized as BIS features to represent the brain's comprehension of video semantics. An effective two-stage feature selection procedure is applied to derive the most relevant features while removing redundancy, which results in the formation of the finally obtained BIS features. Meanwhile, the feature selection provides a data-driven scheme to identify DICCCOL ROIs that are most relevant to video comprehension. Since fMRI scanning is quite expensive and time-consuming, we can only acquire a relatively small amount of fMRI datasets for predictive model learning. However, we can yield plenty of low-level visual features easily. Thus, we aim to build a mapping from low-level visual features to high-level fMRI-derived semantic features by using the GPR [23], [24], given a small number of scanned fMRI datasets in the training stage. Essentially, the mapping realizes the identification and selection of visual features that most correlate with the human cognition of video understanding. The learned mapping can be regarded as a primitive form of "mind-reading," which predicts BIS features for any video shots without fMRI data. In this way, each shot in the database can be represented in BIS. In the component of video shot retrieval, a manifold structure is inferred where video key frames are represented in BIS. Given a query of video, the manifold-ranking [25], [26] algorithm concerning the relationship among all data on the manifold is applied to measure the similarity between key frames and rank the shots.

The work in this paper is a substantial extension of our preliminary study in [20] and there are several major novelties and differences as follows. First, we combined the powerful diffusion tensor imaging (DTI) data with fMRI brain imaging data to accurately map and annotate the large-scale functional networks that are potentially involved in the perception and cognition of video clips. This brain imaging method significantly improved the quantitative measurement of functional brain responses so that our high-level brain imaging features could be much more comprehensive and systematic. Second, this work proposes a generalized methodology for representing video shots in the BIS, whereas the study in [20] only considered a special case of two classes of videos. To build the generalized methodology, the feature selection is an essential component and we develop a two-stage approach in this paper, whereas the feature selection was not explored in [20]. Third, in [20], 30 ROIs identified by task-based fMRI were used to measure the brain responses to video stimulus. In this paper, we propose a novel data-driven approach, named

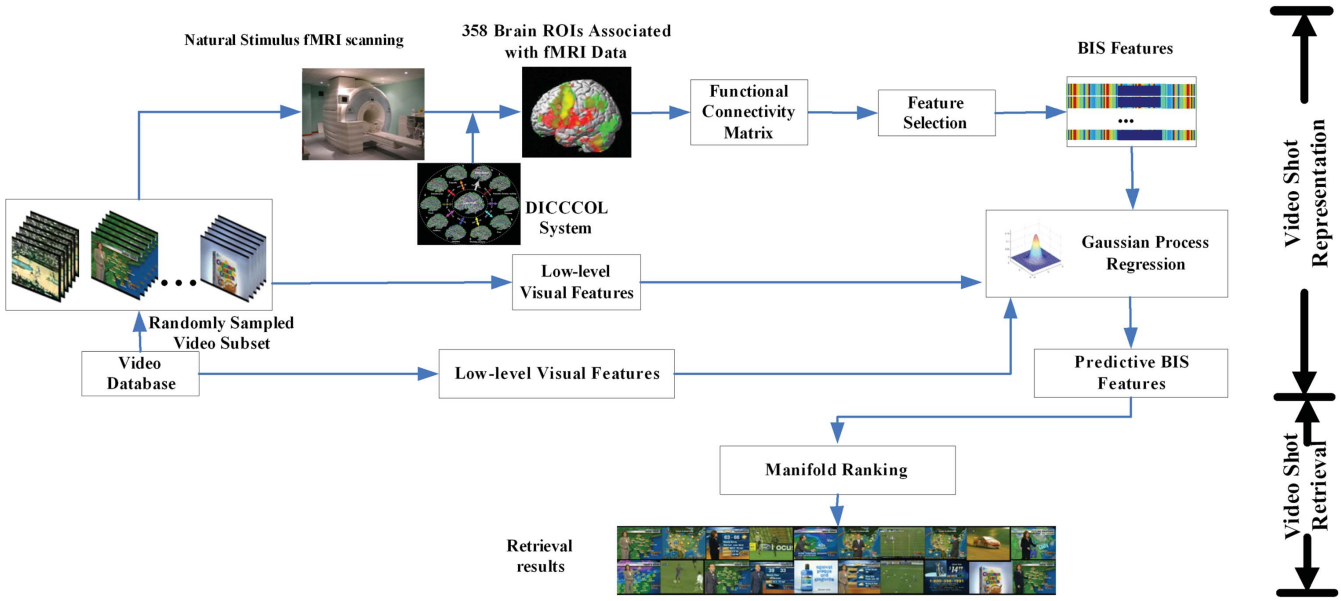


Fig. 1. The overall architecture of the proposed framework.

DICCCOL as mentioned previously, to systematically identify and locate large-scale ROIs involved in video comprehension. Specifically, we apply the DICCCOL system to collect 358 dense and consistent ROIs and then use feature selection algorithms to select a set of pairwise functional connectivities to measure the distinct brain responses involved in the comprehension of different video categories. Comparing with [20], this work avoids large-scale task-based fMRI acquisition, which is very time-consuming and expensive. More importantly, the DICCCOLs can provide more comprehensive, accurate, and reliable ROIs to measure the brain responses during video comprehension. Therefore, this paper can further improve the performance of [20], which will be demonstrated in Section IV-A (Fig. 5). Fourth, in this paper, we explore and present more in-depth theoretic analysis, provide a more extensive discussion of related literatures, and perform more comprehensive experimental evaluations.

The rest of this paper is organized as follows. Section II describes video key frame representation in the BIS. Section III introduces video shot retrieval by using manifold-ranking approaches. The experiments and results are reported in Section IV. Finally, the conclusions are drawn in Section V.

II. Representing Key Frames in the BIS

A. Brain ROI Identification and Localization

In principle, brain function is realized via large-scale structural and functional connectivities [27]–[29]. The functional connectivities and interactions among relevant brain networks reflect the brain's comprehension of video stimuli [27]–[29]. In particular, the working memory [30], vision [31], language [32], emotion [33], semantics [34], attention [35], and motor [36] brain networks are among the most relevant functional systems that are involved in the comprehension of natural movies.

In the functional brain imaging field, task-based fMRI has been widely regarded and used as a benchmark approach to localizing functionally-specialized brain regions. As a result, a large amount of fMRI tasks have been designed and published in the fMRI community [37] to map functional brain networks and their ROIs. However, it is impractical to acquire large-scale task-based fMRI data for the same group of subjects due to the cost and time constraints. In addition, the human brain's responses to the natural stimulus of watching videos could be very complex and involve many different functional networks such as visual, auditory, emotion, working memory, attention, language, and many others. It is infeasible to localize relevant large-scale functional networks involved in the comprehension of movie watching via traditional task-based fMRI analysis.

Recently, we developed and validated a novel data-driven strategy [21], [38] that identified 358 consistent and corresponding structural landmarks in multiple brains (colored bubbles in Fig. 2), in which each identified landmark was optimized to possess maximal group-wise consistency of DTI-derived fiber connection patterns [21], [22], [38]. The neuroscience foundation is that each brain's cytoarchitectonic area has a unique set of extrinsic inputs and outputs, named the "connectional fingerprint" in [39], which principally determines the functions that each brain area could perform. This close relationship between structural connection pattern and brain function has been confirmed and replicated in a variety of recent studies in the literature [39] and our own works in [21], [22], [38]. This set of 358 structural brain landmarks is named DICCCOL and has been replicated in four separate healthy populations [21]. Importantly, this set of 358 ROIs can be accurately and reliably predicted in an individual subject based only on DTI data [22], demonstrating the remarkable reproducibility and predictability of DICCCOLs. The DICCCOL system has already been released online at: http://dicccol.cs.uga.edu for additional evaluation.

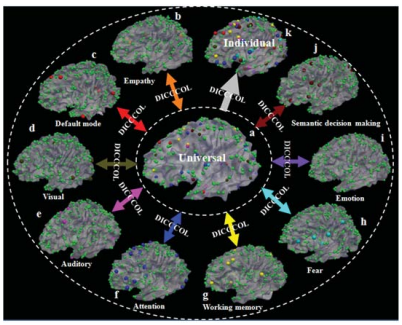


Fig. 2. Nine different functionally-specialized brain networks ((b)-(j)) identified from different fMRI datasets are integrated into the same universal brain reference system (a) via the 358 DICCCOL ROIs. Then, the functionally-labeled landmarks in the universal space can be predicted in each individual brain with DTI data such that DICCCOL ROIs and their functional identities can be transferred to a local coordinate system (k).

In addition, we have used six different multimodal DTI and fMRI datasets (including both task-based fMRI and resting state fMRI) to functionally label these DICCCOL ROIs [21] (as shown in Fig. 2). In our current collection of DICCCOL ROIs [21], there are 95 ROIs that have already been functionally annotated into the brain regions in nine functional networks including working memory, visual, auditory, semantics, attention, emotion, fear, empathy, and default mode networks. DICCCOL enables us to measure functional interactions among large-scale relevant brain networks (particularly those involved in video processing and comprehension [17]–[20]) without the need to perform large-scale costly task-based fMRI scans on the participating subjects.

This paper proposes to employ our DICCCOL system to localize large-scale relevant functional networks involved in the comprehension of movie watching as follows. First, we predict the set of 358 dense and consistent DICCCOL landmarks that provide a common and individualized brain reference system in each participant's brain based on DTI data. Afterwards, the data-driven method of feature selection (Section II-C) is adopted to infer the brain's functional interaction patterns from natural stimulus fMRI data, which can simultaneously identify and localize the most relevant DICCCOL brain ROIs involved in video comprehension.

B. FMRI Data Acquisition and ROI Prediction

To explore brain comprehension to video contents, we developed an experimental paradigm to perform fMRI scanning when human subjects were watching video stimulus. Four healthy young adults were recruited at The University of Georgia (UGA) under IRB approvals to participate in this study. MRI data was acquired in a GE 3T Signa HDx MRI system using an 8-channel head coil at the UGA. The multimodal DTI and fMRI scans were performed in three separate scan sessions for each participating subject. DTI scans were performed for each participant to localize their DICCCOL ROIs. DTI data was acquired using the isotropic spatial resolution 2 mm $\times$ 2 mm $\times$ 2 mm; parameters were: TR $=$ 15.5 s, TE $=$ min-full, b-value $=$ 1000 for 30 DWIs and 3 B0 volumes. DTI data preprocessing includes skull removal, motion correction and eddy current correction [38]. T1-weighted structural MRI data with 1 mm $\times$ 1 mm $\times$ 1 mm isotropic resolution was acquired.

To perform natural stimulus fMRI scanning, 51 video shots including 20 sports, 19 weather reports and 12 commercial were randomly selected from the TRECVID 2005 database and were composed into 8 clips. Each clip is about 11 minutes long. These clips were presented to the four subjects during fMRI scan via MRI-compatible goggles. The scan parameters are as follows: 30 axial slices, matrix size 64 $\times$ 64, 4 mm slice thickness, 220 mm FOV, TR $=$ 1.5 s, TE $=$ 25 ms, ASSET $=$ 2. The strict synchronization between movie viewing and fMRI scan was achieved via the E-prime software [40].

The preprocessing of fMRI data includes skull removal, motion correction, spatial smoothing, temporal prewhitening, slice time correction, and global drift removal. The brain ROI prediction approach in [22] was used to localize the 358 DICCCOLs in the scanned subjects with DTI data. Then, natural stimulus fMRI signals were extracted for each of these 358 DICCCOLs after linearly transforming the ROIs to the fMRI image space. Afterwards, the principal component analysis (PCA) was applied on the multiple fMRI time series within each ROI for extracting a representative fMRI signal [38]. The eigenvector corresponding to the largest eigenvalue was defined as the representative fMRI signal for this ROI.

C. Feature Extraction in the BIS

In the neuroscience field, functional connectivities [41] among relevant brain ROIs have been widely used to reflect meaningful interactions within brain networks. In this paper, we also adopt the functional connectivities of brain ROIs within each video shot time interval as the high-level semantic features in the BIS to model and describe the human brain's responses to natural stimulus of watching video streams. Typically, the functional connectivity between two ROIs is measured as the Pearson correlation coefficient between their fMRI time series. As a result, for each scanned video shot, we constructed a 358 $\times$ 358 connectivity matrix. Fig. 3 shows two randomly selected examples of the connectivity matrix corresponding to two different types of video shots (sports and commercial). As can be seen, the connectivity patterns for the two shots are quite different. For example, the functional interactions are globally much stronger when watching the commercial shot compared with that when watching the sport shot. This observation is reasonable given the fact that the content in commercial videos is typically much more complex for participant to comprehend, in comparison with that in sport videos.

In our previous paper [20], we used a functional connectivity matrix built on 30 functional brain ROIs to form the BIS and represent video shots. However, video is typically a synthesis of visual, aural, textual information. The functional brain mechanism for video comprehension may be far beyond what 30 brain ROIs can account for. In this paper, we employ a data-driven approach to select relevant fMRI response features from 358 consistent and dense ROIs that can cover the whole

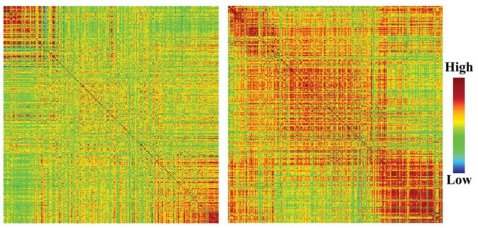


Fig. 3. Two randomly selected examples of functional connectivity matrices upon 358 brain ROIs from one subject for two video shots. (a) A sport shot. (b) A commercial shot.

brain to represent video shots. We expect that an effective feature selection procedure can pick up the most relevant brain networks and their ROIs that are involved in video comprehension. If we directly use all elements in the 358×358 connectivity matrix as the high-level features in the BIS, the size of the feature vector is 63 903 dimensions, which is apparently not a concise and informative representation. Our strategy is to use the 358 ROIs to construct an over-complete ROI set. Then, we employ supervised feature selection upon the over-complete set to seek the most relevant features that are able to differentiate various video classes while removing redundant features. The selected features finally form the BIS and are used as the feature vector to represent video shots. At the same time, this feature selection procedure provides a data-driven method to automatically identify brain ROIs that are most involved in video perception and cognition. It can overcome the shortcoming of traditional methods that apply task-based fMRI to derive brain activations and identify ROIs under the hypothesis that brain networks activated by tasks can represent the complex content of video streams.

In general, the objectives of feature selection are twofold. One is to maximize feature relevance. The other is to minimize feature redundancy. It is worth noting that most current feature selection algorithms are incapable of achieving the above two objectives upon a very large dimensional feature set (for example, 63 903) in practice. Accordingly, this paper implements a two-stage procedure to fulfill the relevance maximization and the redundancy minimization of feature selection separately, which has been demonstrated to be effective for our study by experiments (Section IV-B).

The first stage of feature selection mainly attempts to achieve relevance maximization. In this paper, we adopt a statistical test of Analysis of Variance (ANOVA) [42]. It treats each dimensional feature as an independent variable and determines each dimensional feature's relevance individually by evaluating its correlation with the target class. Given a dimensional feature with the sample $f_{i,j}$ for the ith class and jth data point, the core idea is to perform the F-test [42], [43] to assess whether the expected values of this dimensional feature within several target classes differ from each other, which can be formulated as [42], [43]:

$$F = \frac{\text{between-class variability}}{\text{within-class variability}} \tag{1}$$

where

$$\text{within-class variability} = \sum_{i=1}^{C}\sum_{j=1}^{B_i}(f_{i,j} - E_i)^2/(\sum_{i=1}^{C} B_i - C), \tag{2}$$

E_i is the ith class mean

$$\text{between-class variability} = \sum_{i=1}^{C} B_i(E_i - \bar{E})^2/(C-1),$$

$$\bar{E} = \frac{1}{\sum_i B_i}\sum_{i=1}^{C}\sum_{j=1}^{B_i} f_{i,j} \tag{3}$$

Here, C is the number of classes and B_i is the number of samples in the ith class. Essentially, a feature resulting in significant differences between classes indicates a trend that it has a high impact on classification, which is called the feature relevance. Then, the results of F are tested for statistical significance or p-value [42], [44], where the p-value is the estimated probability of rejecting the null hypothesis of a study question when that hypothesis is true. If the p-value of a feature is small, it implies there is strong evidence that the differences between classes are big, i.e. the feature is relevant. Therefore, we achieve the feature relevance maximization by comparing the p-value against a significance level γ [42], [43]. The set of features whose p-values are less than γ is considered as the most relevant features and then is selected to be processed at the second stage. The implementation details of one-way ANOVA can be found in [42].

Since the first stage of feature selection evaluates each dimensional feature individually, it cannot handle feature redundancy. This problem is left to be tackled at the second stage. In the proposed work, the Correlation-based Feature Selection (CFS) algorithm [45] is adopted as the second stage of feature selection because of its good performance. It is a heuristic method for evaluating the worth of a subset of features by taking into account feature-class and feature-feature correlations simultaneously. The hypothesis behind the heuristic can be simply described as: good feature subsets consist of features of highly correlating with the class while uncorrelating with each other. Given a feature subset S consisting of k features, its CFS can be calculated by [45]:

$$\text{Merit}(S) = \frac{k \cdot \overline{\text{Cor}(f, CL)}}{\sqrt{k + k(k-1)\overline{\text{Cor}(f, f)}}} \tag{4}$$

Here, $f \in S$ is a feature and CL denotes a class. $\overline{\text{Cor}(f, CL)}$ and $\overline{\text{Cor}(f, f)}$ are the mean feature-class correlation and the mean feature-feature correlation, which can be computed according to symmetrical uncertainty [45]. By following [45], the feature subset S^* with the maximal merit can be efficiently found and is selected as the final features. The maximal merit often corresponds to the minimal feature-feature correlation. In this way, the feature redundancy minimization is achieved.

D. Mapping from Visual Space to BIS via GPR

As mentioned before, the acquisition of high-level features in the BIS via fMRI scanning under natural stimulus is

Algorithm 1 TGP Regression to Map Features from Visual Space to BIS [24]

Input: N training data represented by visual features as: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, with $\mathbf{X}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d1})$ and represented by BIS features as: $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$, with $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,d2})$;
A test data with only visual features, $\mathbf{X}_{N+1}$.

Output: Predicted BIS features of the test data, $\mathbf{y}_{N+1}$.

Step1: Compute the kernel matrix $\mathbf{K}_X$ and $\mathbf{K}_Y$, where

$$(\mathbf{K}_X)_{i,j} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = \mathbf{K}_X(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\theta_x \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) + \lambda_x \delta_{i,j}.$$

Here, $\theta_x \geq 0$ is the kernel width parameter, $\lambda_x \geq 0$ is the noise variance, and $\delta_{i,j}$ is the Kronecker delta function. $\mathbf{K}_Y$ is defined in the same way;

Step2: For $\mathbf{x}_{N+1}$, predict $\mathbf{y}_{N+1}$ by optimizing the following function [24]:

$$F(\mathbf{y}_{N+1}) = \mathbf{K}_Y(\mathbf{y}_{N+1}, \mathbf{y}_{N+1}) - 2\mu^T \mathbf{K}_Y^{\mathbf{y}_{N+1}} - \nu \log[\mathbf{K}_Y(\mathbf{y}_{N+1}, \mathbf{y}_{N+1}) - (\mathbf{K}_Y^{\mathbf{y}_{N+1}})^T \mathbf{K}_Y^{-1} \mathbf{K}_Y^{\mathbf{y}_{N+1}}]$$

where $\mu = \mathbf{K}_X^{-1} \mathbf{K}_X^{\mathbf{x}_{N+1}}$ and $\nu = \mathbf{K}_X(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) - (\mathbf{K}_X^{\mathbf{x}_{N+1}})^T \mathbf{K}_X^{-1} \mathbf{K}_X^{\mathbf{x}_{N+1}}$. $\mathbf{K}_X^{\mathbf{x}_{N+1}}$ is a $N \times 1$ column vector with $(\mathbf{K}_X^{\mathbf{x}_{N+1}})_i = \mathbf{K}_X(\mathbf{x}_i, \mathbf{x}_{N+1})$;

Return: $\mathbf{y}_{N+1}$.

expensive and costly indeed. It is impractical to carry out fMRI scan for all video shots of a large-scale video collection. On the other hand, machines can produce plenty of visual features easily by current image processing and computer vision techniques. Therefore, we propose to learn a mapping from the visual features to the semantic features in the BIS using some training data, which results in predictions of semantic features given corresponding visual features. The mapping from the visual features to the fMRI-derived semantic features can be regarded as a primitive form of "mind-reading" that can find low-level features strongly correlating to brain behavior in differentiating video shots. The learning of this mapping can be mathematically formulated as a linear regression problem:

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}, \mathbf{w}) + \varepsilon \\ f(\mathbf{x}, \mathbf{w}) &= \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) \end{aligned} \tag{5}$$

where $\mathbf{x}$ is the low-level visual feature vector, $\mathbf{y}$ is the high-level BIS feature vector, $\phi(\mathbf{x})$ is the basis functions, and ε is Gaussian noises. The weight vector $\mathbf{w}$ is the objective of learning. GPR allows a simple analytical treatment of exact Bayesian inference, is powerful for modeling nonlinear dependencies, and has been demonstrated to achieve good performance. It is therefore used in our framework for learning the mapping. As described in [23], [24], the GPR can be efficiently implemented by defining the kernel function instead of directly choosing basis functions. However, the standard GPR algorithm [23] mainly aims at the prediction of a single output and ignores the correlations among output components. Recently, an improved GPR algorithm called Twin Gaussian Process (TGP) [24] was presented to remove these drawbacks. In our work, we utilized TGP to implement the regression. The details of TGP were provided in [24] whereas this paper only described the basic implementation process in Algorithm 1.

III. Retrieval Video Shots in BIS Using Manifold-Ranking

Recently, manifold learning has been successfully applied to image retrieval tasks [25], [46], which work under the assumption that manifold structure is more powerful than traditional Euclidean structure to represent data. Instead of assuming that the image space is a Euclidean space and estimating similarity between images based on their Euclidean distance, the works [25], [46] weakly assume that the image space is a Riemannian manifold embedded in the feature space, which is called image manifold. Then, these works focus on discovering the intrinsic geometrical structure of the image manifold and estimating the similarity between images based on the geodesic distance on the image manifold. In this paper, we basically follow the idea of [25], [46] whereas we explore the video manifold in BIS rather than in low-level feature space or the space based on limited and subjective user interactions.

A. Geometrical Structure of Video Manifold in BIS

We suppose that the key frame $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,D})$ be a point in BIS, where D is the dimensionality of BIS feature.

1. For each point $\mathbf{y}_i$, find its K nearest neighbors based on Euclidean distance; Connect any two points with an edge if they are neighbors.
2. Define the affinity matrix $\mathbf{M}$ whose element is $\mathbf{M}_{i,j} = \prod_{n=1}^{D} \exp[-(y_{i,n} - y_{j,n})^2/(2 \times \sigma^2)]$ if there is an edge linking $\mathbf{y}_i$ and $\mathbf{y}_j$. Note that self-reinforcement should be avoided, thus let $\mathbf{M}_{i,i} = 0$.

Repeating the above steps results in constructing a graph, which models the local geometrical structure of the video manifold. The geodesic distances between all pairs of video shots on the video manifold are defined as the shortest-path distances on the constructed graph.

B. Performing Retrieval in BIS Using Manifold-Ranking

We adopted a manifold-ranking algorithm [26] to perform retrieval, which measures the similarity between the query and the shots in the database via examining the relationship of all data upon the intrinsic global manifold structure in the BIS, instead of the traditional way of using local pair-wise Euclidean distances based on low-level features. Its basic idea is to spread a positive ranking score assigned to the query to its nearby unlabeled neighbors on the manifold structure until a global stable stage is achieved. The details of the algorithm are summarized below.

Algorithm 2 Manifold-Ranking for Retrieving Video Shots

Input:	The query key frame set q, geometrical structure of video manifold in BIS including N points each representing a key frame in the database.
Output:	Ranking score vector $\mathbf{r} = (r_1, \ldots, r_N)$ in which r_i denotes the ranking score of ith key frame to q.
Step 1:	Normalize $\mathbf{M}$ in Section III-A by $\mathbf{U} = \mathbf{P}^{-1/2}\mathbf{M}\mathbf{P}^{-1/2}$ where $\mathbf{P}$ is a diagonal matrix with $\mathbf{P}_{i,i} = \sum_j \mathbf{M}_{i,j}$;
Step 2:	The theorem in [47] can guarantee that $\{r_i\}$ converges to $\mathbf{r} = \beta(1-\alpha\mathbf{U})^{-1}\mathbf{L}$. Here $\beta = 1-\alpha$, $\mathbf{L} = [l_1, \ldots, l_N]^T$ is a binary vector that can indicate whether a key frame is a query or not, where $l_i = 1$ if ith point is a query, and $l_i = 0$ otherwise.
Return:	$\mathbf{r} = (r_1, \ldots, r_N)$.

IV. Experiments

The NIST TRECVID is a common, well-known, and publicly available video dataset [48]. It has been widely used as the benchmark for evaluating tasks such as video retrieval, video shot boundary detection, video summarization, video instance search, video copy detection, and so on, since it is large, diverse, and contains full-length video sequences. In this paper, we constructed our evaluations on the TRECVID 2005 video streams [48], which are mainly selected from the television news and NASA science programming. As summarized by [49], these data is categorized into 7 concepts such as politics, finance, science, sports, entertainment, weather, and commercial/advertisement. TRECVID provides annotations for three concepts of videos, including sports, weather, and commercial, which were thus collected as the test data in this paper. TRECVID 2005 had already provided key frames for each video shot with good accuracy. Therefore, we used the key frames provided by TRECVID 2005 in our evaluations. As reported in [48], these key frames are extracted by a group at Dublin City University using an automatic approach. Specifically, the shot boundaries are first automatically detected by using a system developed in [50]. Then, the I-Frame nearest to the middle frame of the shot boundary is selected as a key frame.

Totally, 581 sports, 383 weather, and 343 commercial video shots from TRECVID 2005 were adopted to evaluate the proposed work. This data was randomly split into the training set and the testing set. The training set consists of 51 video shots and was utilized as the natural stimuli presented to subjects for fMRI brain imaging. This set of training data was also applied to feature selection and GPR training. The rest of 1256 video shots construct the testing data. It is clear that the training data set is much smaller than the testing data set because fMRI scanning is expensive and time-consuming.

For each key frame of video shots, its original BIS feature vector is 63 903 dimensional. It becomes 5613 dimensional after using ANOVA and eventually becomes 65 dimensional after performing CFS. In recent years, the model of bag of visual words (BoW) [6] has become one of most successful methods to represent images according to local features, which was also adopted as the low-level visual features in our work. In this work, firstly, a set of SIFT [4] descriptors each being 128 dimensionality were extracted from each key frame. Then, the K-means clustering algorithm was applied to cluster all SIFT descriptors to a number of visual words. At last, each key frame was characterized by a feature vector to reflect the probabilistic distribution of those words. Considering that our BIS feature is 65 dimensional and the size of our training set is relatively small, in order to obtain a GPR model with good performance, we need the size of visual low-level feature vector is comparable to the size of BIS feature vector. Hence, in our implementation, the number of words in BoW was set to 65.

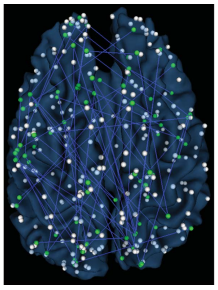


Fig. 4. Localization visualization of identified functional connections for video comprehension and discrimination in DICCCOL system on a cortical surface. The spheres (both green and white) are the DICCCOL landmarks. The green ones are the landmarks involved in the indentified function connections.

A. Mapping of Brain Networks and ROIs Involved in Video Comprehension and Discrimination

This work used the DICCCOL system and feature selection to identify a set of brain ROIs that are involved in video comprehension and discrimination. In our work, 65 features from the functional connectivity matrix were selected, which are associated with 93 brain ROIs in the DICCCOL system. Fig. 4 visualizes the identified functional connections and brain ROIs overlaid on a cortical surface. It is apparent that a large portion of the whole brain, including the visual, auditory, language, attention, emotion, working memory and motor systems, are involved in video comprehension, as we expected.

We also labeled the functions of those 93 selected ROIs into different brain networks according to the DICCCOL system's functional annotations. The top 10 functional networks that have the largest percentages of those selected ROIs are listed in Table I. It turns out that the selected functional networks are quite reasonable and meaningful given current neuroscience knowledge. For instance, the attention, speech, semantics, emotion, execution, and working memory systems are among the most relevant brain networks in video comprehension. This result suggests that video comprehension involves the functional interaction of large-scale brain networks, as demonstrated by the widespread brain ROIs in Fig. 4, and that our feature selection can pick up the relevant brain networks, as listed in Table I. Therefore, our experimental and computational framework can extract meaningful and descriptive

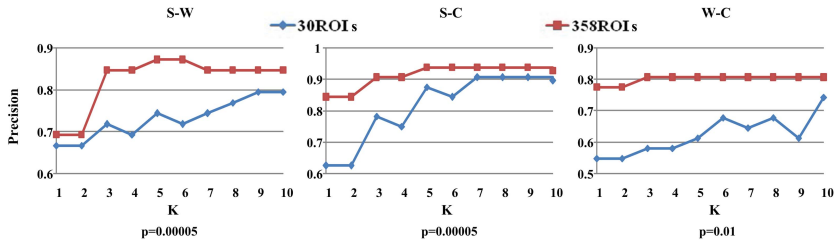


Fig. 5. The comparison of KNN-based classification using features of brain responses from the 30 ROIs and the 358 ROIs. p is the p-value of the t-test in feature selection.

TABLE I
TOP TEN BRAIN NETWORDS INVOLVED IN VIDEO COMPREHENSION AND DISCRIMINATION

#	Brain networks	Percentage (%)
1	Attention	9.15%
2	Execution.speech	7.19 %
3	Language.semantics	7.19 %
4	Emotion	6.54%
5	Language.speech	6.54%
6	Memory.explicit	6.54%
7	Execution	4.58%
8	UGA.emotion	3.92%
9	Cognition	3.27%
10	Memory.working	3.27%

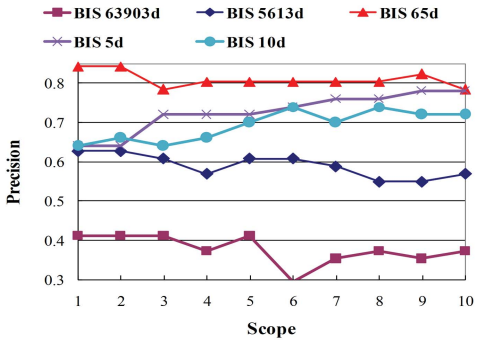


Fig. 6. Evaluations of the proposed feature selection for BIS features using K-NN classifier. (the numbers following "BIS" indicate the numbers of feature dimensions.)

BIS features from the fMRI data in order to characterize and describe the brain's responses to video shots.

This paper proposes to use our DICCCOL system to systematically localize functional brain networks, based on which the brain responses during free viewing of video stream are quantified. Then, feature selection is used to identify distinct brain responses during the comprehension of video stream in different categories. Traditionally, task-based fMRI is a standard method to localize functionally-specialized ROIs, which was used in our previous work [18], [20]. Comparing with task-based fMRI, the proposed method can provide large-scale, accurate, and reliable brain ROIs and thus offer effective measurement of brain responses involved in video comprehension. To demonstrate this point, we followed the experiment designs in [18], [20] and constructed three two-class classification experiments including Sports VS Weather Report (S-W), Sports VS Commercial (S-C), and Weather Report VS Commercial (W-C) to compare the performance by using 30 ROIs [18], [20] with that by using 358 ROIs. The experiments were performed on 51 training videos. First, the most relevant elements in the connectivity matrices were selected by a two-tailed t-test [20]. Then, we used the KNN classifier and the leave-one-out strategy to perform the classification. The results are summarized in Fig. 5. As can be seen, substantially higher classification accuracies were achieved by using 358 ROIs, which demonstrates the proposed method of brain ROIs localization is superior.

B. Evaluation of BIS Feature

We conducted three experiments of video classification to evaluate the performance of the proposed BIS features. It is worth noting that we used K-nearest neighbor (K-NN) classifiers in our test because it is a naïve classifier, which has no inherent feature selection mechanism inside unlike other advanced classifiers such as SVM. The classification performance by using simple classifiers is more likely to reflect the effectiveness of the proposed video representation in BIS.

The first experiment was to test the supervised two-stage approach of feature selection used in this paper. We adopted those 51 video shots which have BIS features and used the leave-one-out cross-validation to accomplish the classification. Fig. 6 shows the classification comparison results based on the original BIS features, selected features after the first stage (ANOVA), and selected features after the second stage (CFS), respectively. It can be clearly seen that our feature selection procedure is effective and achieves much better results. The classification accuracy by using 63 903 dimensional BIS features can be improved by more than 40% by using selected 65 dimensional features. It is worth noting that our experimental results do not imply that the classification accuracy will become higher as we use smaller number of features. We also show the classification accuracy using 5 and 10 dimensional features to demonstrate this point in Fig. 6.

In this work, a two-stage of approach consisting of ANOVA and CFS (AC) were utilized to select features to form the BIS. To further evaluate the effectiveness of AC, we designed two experiments to compare it with other classical feature selection algorithms such as Gentle Adaboost (GA) [51], Fisher Score (FS) [52], [53], Relief (RE) [53], [54], and Fuzzy Entropy (FE) [55]. The GA was selected here since it is simple to implement, robust, and has been shown to outperform other boosting variants experimentally. The first experiment used

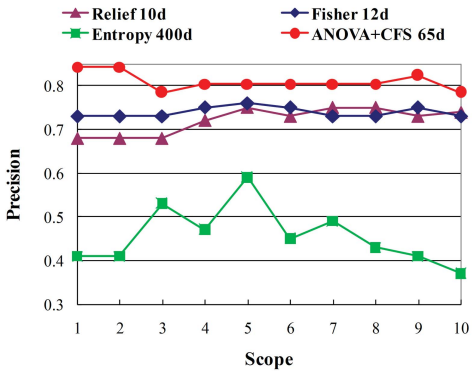


Fig. 7. Comparisons of classification using various feature selection algorithms in the BIS.

TABLE II
TIME COSTS OF AC IN DIFFERENT SETTINGS (FD INDICATES THE FEATURE DIMENSIONALITY)

$p(\times 10^{-3})$	FD After ANOVA	Time Cost of ANOVA (min)	FD After CFS	Time Cost of CFS (min)
4.1	4792	0.5103	62	8.4138
4.2	4889	0.5142	62	8.6974
4.3	4990	0.5090	62	9.0913
4.4	5079	0.5168	65	9.9737
4.5	5146	0.5087	62	9.8319
4.6	5234	0.5123	66	11.0924
4.7	5336	0.5090	65	11.5336
4.8	5440	0.5140	65	11.7384
4.9	5527	0.5088	65	12.9674
5	5613	0.5130	65	15.0000

those 51 video shots and the leave-one-out cross-validation to accomplish the classification task in the BIS. For FS, RE, and FE, we utilized K-NN as classifier. For GA, we utilized its original strong classifier consisting of a number of selected weak classifiers as the classifier, where each weak classifier corresponds to a selected feature. The classification accuracies using GA by varying the number of weak classifiers (the number of selected features), T, are listed as follows: 78.43% (T = 5), 78.43% (T = 10), 80.39% (T = 15), 76.47% (T = 20). The average accuracy is 78.43%. Fig. 7 shows comparison results of AC, FS, FE, and RE. As can be seen, FE did not obtain good performance. AC improves FS and RE by about 5%. AC is slightly better than GA with the improvement of 2.5%. However, the computational complexity of AC is much lower than that of GA according to our experiments. By using Matlab code, the running time of GA (T = 15) is 3 h on a Duo Core 2.93 GHZ machine with 2GB RAM, whereas the running time of AC is 15 min. Moreover, we list the time costs of AC in different settings by varying p-value in Table II. As can be seen, the time cost of ANOVA is relatively small and constant across various settings whereas the time cost of CFS is considerably affected by the number of input features.

In the second experiment, we used SVM classifier to measure the classification performance using various feature selection approaches. The classification accuracies of AC, FS, FE, and RE are 82.35%, 76.47%, 76.47%, and 41.18%, respectively. The results of these two experiments were basically similar and can demonstrate that AC used in this paper is effective and efficient for selecting informative features from fMRI data.

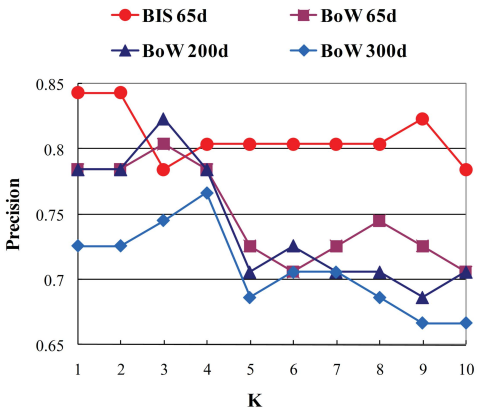


Fig. 8. Comparisons of classification in BIS and in low-level feature space. (Numbers following "BoW" indicate the numbers of feature dimensions).

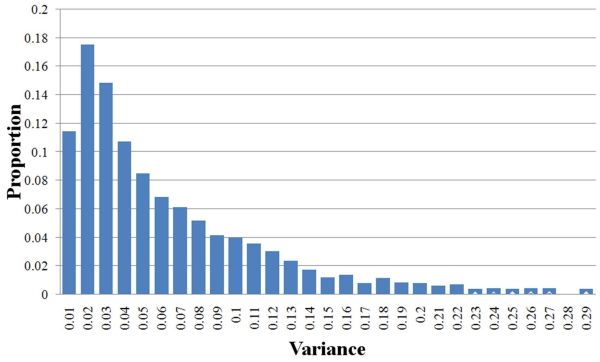


Fig. 9. The statistical consistency of identified functional connectivities across subjects.

The third experiment was constructed to compare the classification performance of BIS features and state-of-the-art low-level visual features. We also used those 51 video shots with fMRI data. Fig. 8 displays the comparison results. As can be seen, the classification performance in the BIS is better than that using BoW. Averagely, the improvement is 7.7%.

C. Evaluation of Model Consistency Across Subjects

We performed two experiments to assess if the obtained result is consistent across various subjects. fMRI scanning data of three subjects when watching those 51 videos were used. As described in Section IV-A, 65 pair-wise functional connectivities involved in video comprehension were identified via DICCCOL and feature selection. The first experiment was to evaluate the consistency of identified ROIs and connectivity across subjects. For each of 51 shots, we collected 65 pair-wise connectivities for each subject. We then calculated the variance of each element across three subjects. Totally, 65×51 variances were obtained and the distribution is illustrated in Fig. 9. The statistical result reflects that those identified ROIs and connectivities by using the proposed scheme have consistent responses to video stimulus across various subjects.

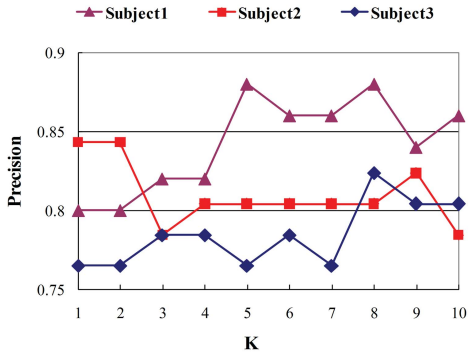


Fig. 10. Classification in the BIS using the proposed approach across various subjects.

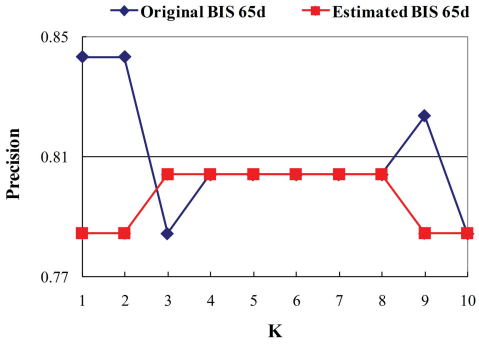


Fig. 11. Classification comparison by using the original BIS features and the estimated BIS features.

The second experiment was designed to test the consistency of performance of feature selection and classification across various subjects. We applied the leave-one-out cross-validation and K-NN classifier to compute the classification performance. For each subject, we repeated the feature selection to obtain BIS features and classification using these BIS features. Fig. 10 shows the classification results for three subjects. As can be seen, the classification performance associated with each subject is generally pretty good and performance variance across subjects is minor, which demonstrate the feature selection and classification proposed in this paper are relatively robust to different subjects.

D. Evaluation of GPR

Two experiments were designed and performed to evaluate the GPR algorithm. In the first experiment, we compared the classification capability by using estimated BIS features and by using original BIS features. In our test data, 51 video shots have fMRI data and thus BIS features. We call these BIS features "original BIS features." We trained a mapping from the BoW features to the BIS features by using the GPR algorithm described in Section II-D upon these 51 video shots. These BIS features are called "estimated BIS features." Subsequently, we applied the original BIS features and the estimated BIS features to classify video shots by using the K-NN algorithm, respectively. The classification comparison is shown in Fig. 11. As can be seen from the results, the classification precisions by using the original BIS features and the estimated BIS features are quite close. Their precision difference is only about 0.0176. This result suggests that the GPR algorithm can effective map the two features spaces of BoW and BIS features, and exhibits remarkable predictability power. In this sense, the GPR serves as a bridge that links the two feature spaces (BoW and BIS), enabling us to apply the learned model on video shots without fMRI scans. This capability is critically important to apply the proposed work of representing and retrieving video shots in BIS in large-scale real-world video shot databases.

The second experiment aims to measure the average classification precision using the estimated BIS features. We suppose $lable_i$ be the classified label for the ith video shot by using K-NN and the original BIS features, and $\overline{lable_i}$ be the classified label for the ith video shot by using K-NN and the "estimated" BIS features. The classification using the "estimated" BIS features is correct if $lable_i = \overline{lable_i}$. The average classification precision is then defined as the correction rate for all the data. We also used the K-NN classifier. The overall average classification by varying K from 1 to 10 is about 0.89. The above two aspects of quantitative experimental results together demonstrate that the estimation capability of GPR algorithm is superior, which can effectively bridge the two feature spaces of BoW and BIS and enables us to perform large-scale video shot retrieval in the BIS space.

E. Evaluation of Video Retrieval

Similar to most retrieval systems [25], [46], our retrieval is also based on the query-by-example paradigm. Given a query video represented by key frames, the similarity between every video in the database and the query is measured. Then, all videos in the database are sorted in descending order of similarity. Finally, a number of video results that are most similar to the query are returned. Our evaluation is to compare those returned results with the ground truth data and quantitatively compute the retrieval accuracy. Specifically, this paper constructed the video retrieval experiment on our testing dataset that consists of 1256 video shots from TRECVID 2005. The GPR model was learned by using training video set as described in Section II-D. It was applied to build the mapping from low-level visual space to BIS for each testing video data. The manifold structure was generated in BIS for 1256 videos as described in Section III-A. Afterwards, given a video query, the similarity between every video in the database and the query was measured upon the BIS manifold using the manifold ranking algorithm as described in Section III-B. Following [25], [46], the retrieval accuracy was calculated as: Precision = (the number of relevant results retrieved in top V returns)/V. In our experiment, we adopted every one of 1256 videos in our database as the query to perform the retrieval, which results in 1256 retrieval sessions totally. The average retrieval accuracy of those 1256 retrieval sessions was used to measure the retrieval performance of the proposed method. For comparison, we also created the manifold structure and retrieved video shots in the low-level visual space using BoW features. Fig. 12 shows three sets of retrieval examples by using the proposed BIS features and low-level BoW features, respectively. In these retrieval examples, 10 results that are

Fig. 12. Three examples of retrieving video shots using BIS features and BoW features, respectively. In each example, the top-left image shows the key frame of the query video. The retrieved results by using BIS features are shown in the top two rows and the retrieved results by using BoW features are shown in the bottom two rows. The images labeled by the red boxes are irrelevant examples.

most similar to the query from the database were returned. It can be easily seen that the retrieved results in BIS are more similar to the corresponding query than those in the visual low-level feature space. Additionally, it is worth noting that some correct results retrieved in BIS are visually different from the query (especially in the third example). This observation can highlight the difference between the proposed work and traditional work and demonstrate that understanding videos in BIS can capture the semantics of brain cognition. Quantitatively, Fig. 13 presents the performance by comparing the proposed retrieval in BIS with the retrieval in low-level visual space described by BoW features. Here, the dimensionality of BIS features is 65 and the dimensionalities of BoW features are 65, 200, 300, respectively. As can be seen, our method can improve the accuracy of the traditional method by about 20%, which is considered substantial.

In the above experiment, we built the BIS via mapping from BoW feature, and demonstrated that using BIS feature significantly improved the retrieval performance of using BoW feature. To further test the effectiveness of the derived BIS features, we constructed another experiment which compared the proposed approach using BIS features with the state-of-the-art method presented in TRECVID 2011. Currently, the popular solution for video retrieval/indexing in TRECVID 2011 is to combine global, local, and motion features to yield a more powerful visual representation for video shots, and then calculate the similarity between videos using the

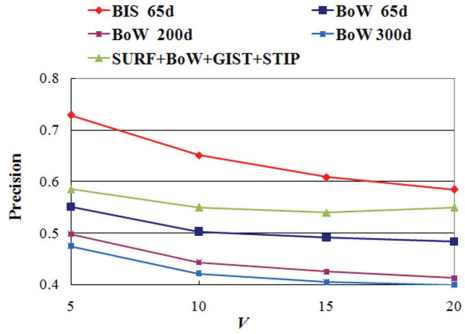


Fig. 13. Performance comparisons of video retrieval in the BIS and in low-level visual space.

combined video representation. As reported in [56], [57], the combination of different features generally can obtain better performance. Therefore, it is reasonable to regard the method based on the combined features as the state-of-the-art approach. Based on reports of TRECVID 2011 [56]–[58], we selected a number of features that were mostly used and also achieved better retrieval performance experimentally to form the video representation. To be specific, SURF [5], [57] was selected as the global feature. SIFT based BoW [6], [56]–[58] and GIST [57], [58] were selected as the local feature. STIP [57]–[59] was selected as the motion feature. The dimensionalities of SURF, BoW, GIST, and STIP features in our experiments are 9, 65, 512, and 65, respectively. We compared two retrieval systems by using the proposed BIS features and the combined low-level features (SURF + BoW + GIST + STIP), respectively, and both using manifold ranking algorithm as the similarity measure. For each of those two retrieval systems, 1256 retrieval sessions that used each video in the database as the query were performed and the average retrieval accuracy was calculated. Fig. 13 shows the quantitative comparison results on the test dataset. It is easy to see that state-of-the-art approach by using combined features achieved better accuracy than that using BoW features (the best one) only with the improvement of 5%. However, on average, it is still worse than the proposed approach using BIS features by 8.3%. Especially, the proposed approach has much higher performance than the approach using combined features when the number of returned results is small (V = 5 and 10). The average improvement is around 12.3%. The experimental results demonstrate that our proposed framework is effective.

In all experiments, a unified set of parameters was used. In ANOVA, the threshold γ was set to 0.005. In the calculation of the affinity matrix M, σ was set to 3. In the manifold ranking algorithm, α was set to 0.99 and $\beta = 1 - \alpha$. The proposed retrieval framework was implemented by Matlab. The computational speeds of off-line feature selection took a few tens of minutes, the GPR training took less than one minute, and the manifold construction took a few seconds. The online manifold-ranking retrieval took a few seconds.

V. Conclusion

In this paper, we have explored a novel and generalized framework to represent and retrieve video shots in BIS where human brain cognition of video semantics can be

captured and represented. Our major contributions can be summarized as follows. 1) The proposed work established the link between two research areas of brain science and video computing via fMRI technology. It developed an innovative brain-computer interface to investigate and leverage the high-level brain imaging space associated with brain behaviors in video perception and cognition, which can significantly boost video understanding. 2) We firstly employed the DICCCOOL system to generate an over-complete set of functional brain ROIs. A data-driven strategy of feature selection was then developed to select the most relevant features, which simultaneously plays an important role in identifying the appropriate ROIs regarded as being involved in video cognition from the over-complete set. In contrast, the traditional method relies on task-based fMRI and is incapable of describing and representing the complicated video content comprehension. The proposed data-driven strategy is more systematic and comprehensive. 3) A computational model was developed to build the mapping from the low-level visual space to the high-level BIS where the maximal correlations between them can be achieved. The computational model can alleviate the burden of lacking fMRI scanning data in the application stage. 4) The manifold-ranking algorithm was applied to retrieve videos represented by BIS features. Evaluations on a benchmark database have demonstrated the effectiveness of the proposed work.

In future, we will improve the proposed work in three aspects. First, more types of BIS features reflecting the brain's comprehension of video stimuli, e.g., functional interactions among cortical and subcortical regions, will be derived. Second, more categories of videos will be used to perform large-scale natural stimulus fMRI scanning and construct a broader BIS. Finally, other alternative computational learning techniques will be investigated. We envision that the combination of functional brain imaging and multimedia research will offer novel perspectives on both fields and advance our understanding of their interactions.

REFERENCES

[1] N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," *IEEE Multimedia*, vol. 9, no. 3, pp. 42–55, Jul.–Sep. 2002.

[2] X. Gao, X. Li, J. Feng, and D. Tao, "Shot-based video retrieval with optical flow tensor and HMMs," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 140–147, Jan. 2009.

[3] Y. Peng, C. Ngo, and J. Xiao, "OM-based video shot retrieval by one-to-one matching," *Multimedia Tools Appl.*, vol. 34, no. 2, pp. 249–266, Aug. 2007.

[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2006, pp. 346–359.

[6] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 524–531.

[7] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[8] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.

[9] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[10] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[11] J. Wang, E. Pohlmeyer, B. Hanna, Y. Jiang, P. Sajda, and S. Chang, "Brain state decoding for rapid image retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 945–954.

[12] P. Sajda, L. Parra, C. Christoforou, B. Hanna, C. Bahlmann, J. Wang, E. Pohlmeyer, J. Dmochowski, and S. Chang, "In a blink of an eye and a switch of a transistor: Cortically coupled computer vision," *Proc. IEEE*, vol. 98, no. 3, pp. 462–478, Mar. 2010.

[13] A. Kapoor, P. Shenoy, and D. Tan, "Combining brain computer interfaces with vision for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, Mar. 2004.

[15] D. Walther, E. Caddigan, F. Li, and D. Beck, "Natural scene categories revealed in distributed patterns of activity in the human brain," *J. Neurosci.*, vol. 29, no. 34, pp. 10573–10581, Aug. 2009.

[16] K. Kay, T. Naselaris, R. Prenger, and J. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 20, pp. 352–355, Mar. 2008.

[17] X. Hu, F. Deng, K. Li, T. Zhang, H. Chen, X. Jiang, J. Lv, D. Zhu, C. Faraco, D. Zhang, A. Mesbah, J. Han, X. Hua, L. Xie, S. Miller, L. Guo, and T. Liu, "Bridging low-level features and high-level semantics via fMRI brain imaging for video classification," in *Proc. ACM Conf. Multimedia*, 2010, pp. 451–460.

[18] X. Hu, K. Li, J. Han, X. Hua, L. Guo, T. Liu, "Bridging semantic gaps via functional brain imaging," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 314–325, Apr. 2012.

[19] K. Li, T. Zhang, X. Hu, D. Zhu, H. Chen, X. Jiang, F. Deng, J. Lv, C. Faraco, D. Zhang, A. Mesbah, J. Han, L. Lu, X. Hua, L. Guo, S. Miller, and T. Liu, "Human-friendly attention models for video summarization," in *Proc. 12th ACM Conf. Multimodal Inter.*, 2010, pp. 171–178.

[20] X. Ji, J. Han, X. Hu, K. Li, F. Deng, J. Fang, L. Guo, and T. Liu, "Retrieving video shots in semantic brain imaging space using manifold-ranking," in *Proc. IEEE Conf. Image Process.*, Sep. 2011, pp. 3633–3636.

[21] D. Zhu, K. Li, L. Guo, X. Jiang, T. Zhang, D. Zhang, H. Chen, F. Deng, C. Faraco, C. Jin, C. Wee, Y. Yuan, P. Lv, Y. Yin, X. Hu, L. Duan, X. Hu, J. Han, L. Wang, D. Shen, L. Miller, L. Li, and T. Liu, *DICCCOL: Dense Individualized and Common Connectivity-Based Cortical Landmarks*. London, U.K.: Oxford Univ. Press, Apr. 2012.

[22] T. Zhang, L. Guo, K. Li, C. Jing, Y. Yin, D. Zhu, G. Cui, L. Li, and T. Liu, *Predicting Functional Cortical ROIs Via DTI-Derived Fiber Shape Models*. London, U.K.: Oxford Univ. Press, 2011.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[24] L. Bo and C. Sminchisescu, "Twin gaussian processes for structured prediction," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 28–52, Mar. 2010.

[25] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Generalized manifold-ranking-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3170–3177, Oct. 2006.

[26] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *Proc. 18th Ann. Conf. Neural Inf. Process. Syst.*, 2003, pp. 169–176.

[27] K. Friston, "Modalities, modes, and models in functional neuroimaging," *Science*, vol. 326, no. 5951, pp. 399–403, Oct. 2009.

[28] M. Lynall, D. Bassett, R. Kerwin, P. McKenna, M. Kitzbichler, U. Muller, and E. Bullmore, "Functional connectivity and brain networks in schizophrenia," *J. Neurosci.*, vol. 30, no. 28, pp. 9477–87, Jul. 2010.

[29] P. Hagmann, L. Cammoun, X. Gigandet, S. Gerhard, P. Grant, V. Wedeen, R. Meuli, J. Thiran, C. Honey, and O. Sporns, "MR connectomics: Principles and challenges," *J. Neurosci. Methods*, vol. 194, no. 1, pp. 34–45, Dec. 2010.

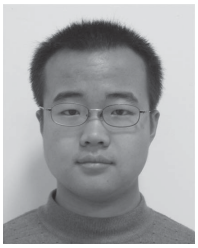
[30] A. Baddeley, "Working memory: Looking back and looking forward," *Nature Rev. Neurosci.*, vol. 4, no. 10, pp. 829–39, Oct. 2003.

[31] J. Singer and G. Kreiman, "Toward unmasking the dynamics of visual perception," *Neuron*, vol. 64, no. 4, pp. 446–447, Nov. 2009.

[32] M. Beauchamp, K. Lee, B. Argall, and A. Martin, "Integration of auditory and visual information about objects in superior temporal sulcus," *Neuron*, vol. 41, no. 5, pp. 809–823, Mar. 2004.
[33] D. Sabatinelli, P. J. Lang, M. M. Bradley, V. D. Costa, and A. Keil, "The timing of emotional discrimination in human amygdala and ventral visual cortex," *J. Neurosci.* vol. 29, no. 47, pp. 14864–14868, Nov. 2009.
[34] J. Rinnea, M. Laineb, J. Hiltunenc, and T. Erkinjuntti, "Semantic decision making in early probable AD: A PET activation study," *Cognitive Brain Research*, vol. 18, no 1, pp. 89–96, Dec. 2003.
[35] L. Ethridge, S. Brahmbhatt, Y. Gao, J. E. McDowell, and B. Clementz, "Consider the context: Blocked versus interleaved presentation of antisaccade trials," *Psychophysiology*, vol. 46, no. 5, pp. 1100–1107, Sep. 2009.
[36] J. Meier, T. Aflalo, S. Kastner, and M. Graziano, "Complex organization of human primary motor cortex: A high-resolution fMRI study," *J. Neurophys.*, vol. 100, no. 4, pp. 1800–1812, Oct. 2008.
[37] A. Laird, S. Eickhoff, F. Kurth, P. Fox, A. Uecker, J. Turner, J. Robinson, J. Lancaster, and P. Fox, "ALE meta-analysis workflows via the BrainMap database: Progress towards a probabilistic functional brain atlas," *Neuroinformatics*, vol. 3, no. 23, pp. 1–11, Apr. 2009.
[38] D. Zhu, K. Li, C. Faraco, F. Deng, D. Zhang, X. Jiang, H. Chen, L. Guo, L. Miller, and T. Liu, "Optimization of functional brain ROIs via maximization of consistency of structural connectivity profiles," *Neuro Image*, vol. 59, no. 2, pp. 1382–1393, Jan. 2012.
[39] R. Passingham, K. Stephan, and R. Kötter, "The anatomical basis of functional localization in the cortex," *Nat. Rev. Neurosci.*, vol. 3, no. 8, pp. 606–616, Aug. 2002.
[40] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime Reference Guide*. Pittsburgh, PA, USA: Psychology Software Tools, Inc. 2002.
[41] B. Horwitz, "The elusive concept of brain connectivity," *Neuroimage*, vol. 19, no. 2, pp. 466–470, 2003.
[42] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. Washington, DC, USA: Government Printing Office, 1964.
[43] R. Lomax and D. Hahs-Vaughn, *Statistical Concepts: A Second Course*, 3rd ed. New York, NY, USA: Academic, 2007.
[44] M. J. Schervish, "P values: What they are and what they are not," *Amer. Stat.*, vol. 50, no. 3, pp. 203–206, 1996.
[45] M. Hall and L. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proc. FLAIRS Conf.*, 1999, pp. 235–239.
[46] X. He, W. Ma, and H. J. Zhang, "Learning an image manifold for retrieval," in *Proc. 12th ACM Conf. Multimedia*, Oct. 2004, pp. 17–23.
[47] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Neural Inf. Process. Syst.*, 2003, pp. 321–328.
[48] *TRECVID Database*. (2005) [Online]. Available: http://www-nlpir.nist.gov/projects/tv2005/
[49] M. Naphade, L. Kennedy, J. Kender, S. Chang, J. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for TRECVID 2005," Dept. IBM Res. Division, Columbia Univ., Yorktown Heights, NY, USA, Tech. Rep. IBM-10598, May 2005.
[50] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *Proc. Text REtrieval Conf.*, 2004, pp. 184–196.
[51] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Stat.*, vol. 28, no. 2, pp. 337–374, 2000.
[52] R. Duda, P. Hart, and D. Stork. *Pattern Classication*, 2nd ed. New York, NY, USA: Wiley, 2001.
[53] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research-ASU feature selection repository," Dept. Comput. Sci. Eng., Arizona State Univ., Glendale, AZ, USA, Tech. Rep., 2010.
[54] I. Kononenko, "Estimating attributes: Analysis and extension of RELIEF," in *Proc. Eur. Conf. Mach. Learn.*, Apr. 1994, pp. 171–182.
[55] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4600–4607, Apr. 2011.
[56] H. Cheng, A. Tamrakar, S. Ali, Q. Yu, O. Javed, J. Liu, A. Divakaran, H. Sawhney, A. Hauptman, M. Shah, S. Bhattacharya, M. Witbrock, J. Curis, G. Friedland, R. Mertens, T. Darrell, R. Manmatha, and J. Allan, "Team SRI-Sarnoff's AURORA system @ TRECVID 2011," in *Proc. NIST TRECVID Workshop*, 2011, pp. 1–3.
[57] D. Scott, J. Guo, and A. Smeaton, "TRECVid 2011 experiments at Dublin City University," in *Proc. NIST TRECVID Workshop*, 2011, pp. 1–9.
[58] L. Cao, S. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, A. Natsev, and J. Smith, "IBM research and Columbia University TRECVID-2011 multimedia event detection (MED) system," in *Proc. NIST TRECVID Workshop*, 2011, pp. 1–14.
[59] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.

Junwei Han (M'10) received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2003.

He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.



Xiang Ji received the M.S. degree from the Northwestern Polytechnical University, China, in 2011, where he is currently pursuing the Ph.D. degree.

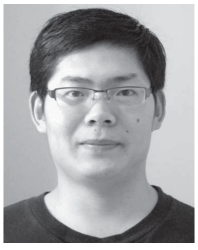
His research interests include computer vision and multimedia processing.

Xintao Hu received the M.S. and Ph.D. degrees from the Northwestern Polytechnical University, China, in 2005 and 2011, respectively.

He is currently a postdoctoral researcher at the School of Automation at NWPU. His research interests include computational brain imaging and its application in computer vision.

Dajiang Zhu was born in Beijing, China, in 1978. He received the B.S. degree in computer science from Shanghai Jiaotong University in 2001. Currently he is a Ph.D. student at the University of Georgia, Athens, GA, USA.

His research interests include MRI imaging analysis, connectome, and the relationship between brain structure and function.



Kaiming Li received the Ph.D. degree from Northwestern Polytechnical University in 2012.

He is currently a postdoctoral fellow at Georgia Tech and Emory University. His research focuses on brain imaging analysis in scientific research and clinical applications.

Xi Jiang received the B.S. degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2009. He is currently working toward the Ph.D. degree at the University of Georgia, Athens, GA, USA.

His current research interests include brain imaging, resting state fMRI, gyral/sulcal function analysis, and fMRI in multimedia (audio/video).

Guangbin Cui received the Ph.D. degree from Fourth Military Medical University, Xi'an, Shannxi, China, in 2004.

He is currently an Associate Professor at Fourth Military Medical University, China. His research interests include magnetic resonance imaging.

Lei Guo received the Ph.D. degree from Xidian University, Xi'an, China, in 1994.

He is currently a Professor at Northwestern Polytechnical University, China. His research interests include computer vision, pattern recognition, and medical image processing.

Tianming Liu (SM'08) received the Ph.D. degree from Shanghai Jiaotong University.

He is currently an Assistant Professor (Associate Professor with Tenure effective August, 2013) of computer science at the University of Georgia (UGA), Athens, GA, USA. He is also an affiliated faculty member at the UGA Bioimaging Research Center, the UGA Biomedical and Health Sciences Institute, the UGA Institute of Bioinformatics, and the UGA College of Engineering. His research focuses on biomedical image analysis, and he has published over 120 peer-reviewed papers in this area.

Before he moved to UGA, he was a faculty member at Weill Medical College of Cornell University (Assistant Professor, 2007-2008) and Harvard Medical School (Instructor, 2005-2007). He was a postdoctoral researcher at the University of Pennsylvania (2002-2004) and Harvard Medical School (2004-2005).

Mr. Liu was a recipient of the Microsoft Fellowship Award (2000-2002), the NIH K01 Career Award (2007-2012), and the NSF CAREER Award (2012-2017).