Sparsity Constrained fMRI Decoding of Visual Saliency in Naturalistic Video Streams

Xintao Hu, Cheng Lv, Gong Cheng, Jinglei Lv, Lei Guo, Junwei Han, Tianming Liu

Abstract

Naturalistic stimuli such as video watching have been increasingly used in recent functional magnetic resonance imaging (fMRI) encoding and decoding studies since they can provide real, complex and dynamic information that the human brain has to process in everyday life. In this paper, we propose a sparsity-constrained fMRI decoding model to explore whether bottom-up visual saliency in continuous and naturalistic video streams can be effectively decoded by brain activities recorded by fMRI, and to examine whether sparsity constraints can improve the performance of visual saliency decoding model. Specifically, we use a biologically-plausible computational model to quantify the visual saliency in the video streams used as stimuli, and adopt a sparse representation scheme to learn the atomic fMRI signal dictionaries that are representative of the patterns of whole-brain fMRI signals. Sparse representation also serves as a unified scheme that links the learned atomic dictionary with the quantified video saliency. Our experimental results demonstrate that the temporal visual saliency information in naturalistic video stream can be well decoded and the sparse constraints can improve the performance of fMRI decoding models, compared with conventional independent component analysis (ICA).

Index terms: fMRI encoding, visual saliency, naturalistic stimuli, sparsity constraints.

X. Hu, C. Lv, G. Cheng, J. Lv, J. Han and L. Guo are with the School of Automation, Northwestern Polytechnical University, Xi'an, 710072, P.R. China. (e-mail: {xhu, gcheng, lguo, jhan}@nwpu.edu.cn, chenggong1119@gmail.com, lledu2008@gmail.com). T. Liu is with Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, 30602, GA, USA. (e-mail: tliu@cs.uga.edu).

I. INTRODUCTION

In recent years, functional brain mapping using functional magnetic resonance imaging (fMRI) under naturalistic stimuli such as video watching (e.g., [1-5]) and music listening (e.g., [6-9]) has received increasing attention. It has been argued that the neuronal responses evoked by naturalistic stimuli are stronger than those at rest or in controlled laboratory conditions using repeated artificial stimuli [10, 11]. More importantly, the complex and dynamic stimuli are likely to engage not only more distinct activation in a wide range of functionally specialized areas [1, 12], but also a broader set of inter-regional functional connections [13].

Compared with conventional fMRI paradigms, there is no straight forward correspondence between the naturalistic stimuli and any specific cognitive functions of the human brain [3]. This makes it difficult to use hypothesis-based analysis methods (e.g., general linear model, GLM) that involve fitting fMRI signals with predictors representing specific experimental conditions or processes. In the literature, researchers have proposed two schools of methods to address this challenge. The first group includes data-driven methods such as Independent Component Analysis (ICA) [1, 13, 14] and Inter-Subject Correlation (ISC) analysis [2, 11, 15]. ICA aims at a blind separation of independent sources [16-18]. It is based on the intrinsic structure of the input data, and does not need any "a priori" specifications of the external stimuli. ISC is based on the assumption that the blood oxygen level dependent (BOLD) responses of different subjects to the same video stimuli are similar [15]. In general, these data-driven methods are predominant in current naturalistic stimuli fMRI analysis. However, one of the limitations of current data-driven studies is the lack of quantitative correlation between measured brain responses and external stimuli, which would largely limit the power of fMRI in exploring the functional brain activities during natural viewing conditions.

The second group of methods tried to quantitatively model the input predictors in external stimuli either by participants' rating [14] or employing biologically-plausible computational models [3, 19, 20]. For instance, in the study in [14], the intensity of the predictors related to color, faces, language and human body were quantitatively ranked by the participants into no percept (1), moderate (2), medium (3) and intense percept (4) for in a movie clip, and followed by a standard GLM-based analysis. In the work in [3], the authors used biologically-plausible visual saliency [21, 22] and auditory saliency computational models [23] to extract quantitative predictors associated with color, intensity, orientation, motion in video, as well as intensity, frequency contrast, temporal contrast and orientation in audio. Those studies demonstrate the promising opportunities of integrating computational video and audio analysis models to advance naturalistic stimuli fMRI based brain studies.

In parallel, multivariate approaches such as multi-voxel pattern analysis (MVPA) has become increasingly popular in linking external stimuli and brain responses [24]. MVPA typically employs pattern classification techniques to decode the cognitive variables from multi-voxel patterns of activity to characterize how cognitive states are represented in the brain, and the structure of the underlying neural codes with increased sensitivity [25]. However, the vast fMRI data poses a key challenge of dimension reduction (feature selection or voxel selection) in MVPA, which is an essential preliminary step to mine the fMRI data effectively [25-27]. Despite the successful application of ICA in fMRI feature extraction, growing interests have been directed to sparsity constrained approaches (e.g. [28-33]). It has been recognized that sparse population coding of a set of neurons is more effective than independent exploration, that is, a sparse set of neurons encode specific concepts rather than responding to the input stimuli independently [34]. At the same time, significant amount of research efforts from the machine learning and pattern recognition fields have been devoted to sparse representations of signals and patterns (e.g. [35-39]). Sparse representation aims at learning an atomic dictionary representing the basis functions contained in the data, and seeking the sparsest linear combination of the basis functions to reconstruct the data [39]. Compared to methods based on orthonormal transforms (e.g., Fourier) or direct time domain processing (e.g., wavelet) with predefined and fixed basis, sparse representation explores the intrinsic distribution of the data and pursuits a set of representative samples. Thus, it usually offers better performance for efficient signal modeling [36], and there have been several studies using sparse representation for fMRI analysis [31, 40-43].

In this paper, we present a sparsity constrained fMRI decoding of visual saliency in naturalistic video streams to explore whether bottom-up visual saliency in continuous and naturalistic video streams can be effectively decoded by brain activities recorded by fMRI, and to examine whether sparsity constraints can improve the performance of visual saliency decoding model. Similar to the study in[3], we use the biologically-plausible computational visual saliency model [44] to quantitatively measure the video saliency curves to capture the temporal information in naturalistic video streams. Afterwards, we adopt a sparse representation algorithm [38] to learn a set of basis BOLD signal patterns (a BOLD code-book) from the fMRI data acquired when the participants watched the video clips. Then, the computationally derived video saliency curves are reconstructed using the learned code-book under the same sparse representation scheme. At last, we evaluate our approach by comparing it with an ICA-based method.

The novelties of our study are twofold. First, we propose to use sparse representation and dictionary learning algorithm to naturally decompose an fMRI signal into a data-driven discovered basis functions that are representative of the patterns of whole-brain fMRI signals. Second, we use the learned atomic fMRI dictionary as a unified scheme to link whole-brain fMRI patterns with temporal saliency information extracted from video stimuli using a computational model. In the following sections, a series of experiments will be designed and conducted to demonstrate the effectiveness of these novel methodologies.

II. MATERIALS

A. System Overview

The overview of our study is illustrated in Fig.1. In the first step (panel (a)), the video saliency curves of the video streams are measured via a biological-plausible computational video saliency model proposed in [44]. In the second step (panel (b)), those video clips were used as naturalistic stimuli and presented to the participants, and fMRI data were acquired. After preprocessing, the sparse representation algorithm proposed in [38] was applied to learn an atomic dictionary for each fMRI data set (acquired when a specific subject watched a single video clip). In the dictionary, each atom corresponds to a representative fMRI signal pattern. In addition, each atom corresponds to a spatial map (see Fig.4 for examples), which describes the contribution of the atom when the trained code book is used to reconstruct whole-brain fMRI signals. In the last step (panel (c)), the trained atomic dictionary is used to reconstruct the saliency curves derived from video clips under a sparsity constrain [38]. Group-wise analysis is finally conducted to infer the consistent functional subdivisions for video saliency curve reconstruction.

[Fig. 1 here]

B. Video Stimuli

The TRECVID database is a widely used benchmark video database in multimedia analysis field. In TRECVID 2005, the Large Scale Concept Ontology for Multimedia (LSCOM) group selected a light scale concept ontology including politics, business, science/technology, sports, entertainment, weather report, and commercial/advertisement to describe the high-level semantic of the video samples [45]. In this study, we randomly selected 32 video shots which are in the semantic categories of sports and weather report (16 for sports, 16 for weather report). Those video shots were then composed into 4 clips in an interleaved semantic label fashion. Each of the video clips is about 10 minutes long.

C. Data Acquisition and Pre-processing

Four healthy university students from The University of Georgia were recruited in this study under IRB approval. All the participants were right-handed and with normal sight. No participant reported head trauma, claustrophobia, was treatment-seeking or any implants or non-removable metal contraindicated in MRI.

The video clips were presented to the 4 subjects for fMRI brain imaging. The video clips were presented to the subjects using a MRI-compatible audio-video delivery system when the subjects were lying in the scanner. Brain images were acquired using a GE 3T Signa MRI system (GE Healthcare, Milwaukee, WI) with an 8-channel head coil at the Bioimaging Research Center of The University of Georgia. The fMRI scanning parameters are as follows: 30 axial slices, matrix size 64×64, voxel size = 3.44×3.44mm in plane, 4mm slice thickness without space between slices, 220mm^2 FOV, TR=1.5s, TE=25ms, ASSET=2. The number of the volumes of the fMRI data is 413. *In total, 16 fMRI data (four subjects were watching 4 video clips) were acquired.* The strict synchronization between media viewing and fMRI scan is achieved via the E-prime software.

The pre-processing of fMRI data included skull removal, motion correction, spatial smoothing with an 8mm full-width at half-maximum (FWHM) Gaussian kernel, temporal prewhitening with autoregressive model AR(1), slice time correction, and global drift removal. The fMRI time-series were further high-pass filtered at 128s [3].

III. VIDEO SALIENCY MODELING

The bottom-up visual saliency maps [21, 22, 44] for naturalistic images/videos is one of the most successful biological-plausible computational models in the computer vision field. In general, visual saliency detection aims at quantitatively predicting attended locations in an image by mimicking the visual selection mechanism of the human vision system [21]. Visual saliency models have yielded fruitful productions in computer vision applications, and also provided promising means for human brain studies [3, 19, 20].

A typical image visual saliency model may be organized into three stages including extracting local multi-channel discontinuities feature vectors such as intensity, color and orientation at locations over the image plane; building an “activation map” (or maps) using the extracted feature vectors; and normalizing and combining the activation map(s) from multi-channel into a single master saliency map [44]. Beside those static spatial features recruited in image visual saliency models, a video saliency model usually adopts dynamic features such as motion and flicker to capture the temporal information contained in the video stream [44]. In addition, the surprise model [46] for quantifying the visual saliency of continuous video stream is typically adopted to estimate the true saliency map.

In this study, we use the graph-based visual saliency (GBVS) model proposed in [44] to build the visual saliency curve. The GBVS model exploits the computational power, topological structure and parallel nature of graph algorithms to facilitate the efficient saliency computation. The related software is available at: http://www.klab.caltech.edu/~harel/share/gbvs.php. In the descriptor extraction stage, the static image descriptors and the dynamic video descriptors employed in GBVS are similar to the classic visual saliency models [21, 22]. The static image descriptors include intensity feature (1 channel of on/off contrast); color feature (2 channels corresponding to red/green and blue/yellow contrast); orientation feature (4 channels corresponding to contrasts at [0°, 45°, 90° and 135°]). The dynamic video descriptors include motion feature (4 channels corresponding to contrasts at [0°, 45°, 90° and 135°]) and flicker feature (1 channel of on/off contrast). In the stage of activation map building, GBVS defines Markov chains over various descriptor maps, and treats the equilibrium distribution over map locations as activation or saliency values [44]. To integrate with the surprise model [46], the activation map of the previous frame is used as prior information when calculating the activation map for the current frame [44]. In the normalization stage, GBVS model defines another Markov chain to concentrate mass on activation maps. It has been reported that this normalization strategy experimentally behaves favorably compared to the standard approaches such as "DoG" (Difference of Gaussian) and "NL" nonlinear interactions [44].

For an input video clip, the GBVS model provides the master saliency maps for each frame as the final output. Following the method that converted the high-dimensional master saliency maps into regressors for the SPM design matrix described in [3], we construct the video saliency curve for the video clip. Briefly, we calculate the mean value of the master saliency map over the vertical and horizontal spatial dimensions for each frame. This step results in a high-dimensional video saliency curve, whose dimension is the same as the number of frames in the video clip. Afterwards, the high-dimensional video saliency curve is down-sampled by averaging over each fMRI repetition time (1.5s), resulting in a video saliency curve which is with the same temporal dimension (the number of volumes) of the acquired fMRI data. At last, the video saliency curve is convolved with the canonical hemodynamic response function (HRF) implemented in SPM [3, 47]. In general, the video saliency curve quantitatively describes the participants' attention directed to the video clip during passive free watching.

IV. SPARSE ENCODING OF VIDEO SALIENCY VIA SPARSE CODING

In this section, we first introduce basic sparse representation theory, and then the learning of the atomic dictionary for a set of fMRI signals. Then the spatial patterns associated with the atoms in the learned dictionary are examined. Afterwards, we introduce the reconstruction of video saliency curve using the learned fMRI atomic dictionary. At last, the strategy for group-wise statistic analysis is detailed.

A. Sparse Representation Theory and Dictionary Learning

The primary goal of classical sparse representation and dictionary learning algorithms is to model data vectors as sparse linear combinations of basis elements [48, 49]. Following the notations in [38, 48], the sparse representation and dictionary learning can be formulated as:

$$\underset{\mathbf{D}\in\mathbb{R}^{m\times k},\boldsymbol{\alpha}\in\mathbb{R}^{k\times n}}{\arg\min} \sum_{i=1}^{n}\left(\frac{1}{2}\left\|\mathbf{x}_i-\mathbf{D}\boldsymbol{\alpha}_i\right\|_2^2+\lambda\left\|\boldsymbol{\alpha}_i\right\|_1\right) \tag{1}$$

where $\mathbf{X}=[\mathbf{x}_1,...,\mathbf{x}_n]$ in $\mathbb{R}^{m\times n}$ is a finite training set of signals with *n* samples and each sample is represented by an *m*-dimensional feature vector. Usually *n* is large whereas *m* is relatively small. $\mathbf{D}\in\mathbb{R}^{m\times k}$ is the dictionary, and each column representing a basis element vector. $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_n]$ in $\mathbb{R}^{k\times n}$ are the decomposition coefficients. In general, $k << n$. In Eq. (1), the first term counts for the signal reconstruction accuracy while the second term counts for the sparsity of the reconstruction coefficients. λ is a regularization parameter to achieve the balance between the signal reconstruction accuracy and the sparsity of the coefficient. Eq. (1) can be rewritten as a matrix factorization problem with a sparsity penalty:

$$\underset{\mathbf{D}\in \mathrm{R}^{m\times k},\boldsymbol{\alpha}\in \mathrm{R}^{k\times n}}{\arg\min} \frac{1}{2}\left\|\mathbf{X}-\mathbf{D}\boldsymbol{\alpha}\right\|_F^2+\lambda\left\|\boldsymbol{\alpha}\right\|_{1,1} \tag{2}$$

where $\left\|\boldsymbol{\alpha}\right\|_{1,1}$ denotes the L_1 norm of the coefficient matrix $\boldsymbol{\alpha}$. In addition, a common constrain that L_2 norm of $\left(\boldsymbol{d}_i\right)_{i=1}^{k}$ does not exceed 1 is introduced to prevent $\mathbf{D}$ from having arbitrarily large values[38], i.e.,

$$\sqrt{\boldsymbol{d}_i^T\boldsymbol{d}_i}\le 1,\quad i=1,\ldots,k \tag{3}$$

B. FMRI Atomic Dictionary Learning

The basic assumption in sparse representation based whole-brain fMRI analysis is that each raw BOLD signal is composed of multiple components, and sparse representation and dictionary learning can naturally decompose an fMRI signal into a data-driven, discovered basis functions that are representative of the patterns of whole-brain fMRI signals. Given the superiority of sparse representation in complex signal modeling compared with predefined and fixed basis [36], it was expected that sparse representation would yield improved mining of the representative BOLD signal patterns in naturalistic stimuli fMRI data.

The sparsity constrained whole-brain fMRI signal analysis is illustrated in Fig. 2. After preprocessing, the fMRI signals for all the brain tissue voxels in a specific fMRI data (acquired when a specific subject watched a specific video clip) are aggregated to form a full data matrix **X**. Each column in **X** is the fMRI signal vector for a single voxel. Thus, **X** is in the dimension of *t*×*n*, where *t* is the length (metric in volume) of the fMRI signal and *n* is the total number of voxels. The atomic dictionary **D** and the associated coefficient matrix **α** are estimated according to Eq. (2) using an online dictionary learning algorithm [38]. *In brief, the online* *dictionary learning algorithm is based on stochastic approximations. It processes one sample at a time (or a mini-batch), and uses second-order information of the cost function to efficiently solve the dictionary learning by sequentially minimizing a quadratic local surrogate of the expected cost. The dictionary learning algorithm consists of a sequence of iterative updates of **D**. In each iteration, it draws one training sample at a time, and alternates classical sparse coding steps for computing the decomposition* $\boldsymbol{\alpha}_t$ *of* $\boldsymbol{X}_t$ *over the dictionary* $\boldsymbol{D}_{t-1}$ *obtained at the previous iteration, with dictionary update steps where the new dictionary* $\boldsymbol{D}_t$ *is computed by minimizing over the cost function. The sparse coding step is solved by a homotopy method [50]. The updating of the dictionary is based on block-coordinate descent with warm restarts, which is parameter free and does not require any learning rate tuning. It is effective to use the value of* $\boldsymbol{D}_{t-1}$ *as a warm restart for computing* $\boldsymbol{D}_t$ *after a few iterations, and a single iteration is sufficient to achieve convergence of the dictionary update step [38].*

After dictionary learning, the fMRI signal matrix **X** can be represented by a learned dictionary matrix **D** and a sparse coefficient matrix **α**.

[Fig. 2 here]

Each atom in the dictionary corresponds to a representative fMRI signal pattern. Each dimension of the coefficient vector associated with a specific atom (the corresponding row in **α**) indicates the contribution of this atom to the reconstruction of the fMRI signal for a specific voxel in the brain image. With the assistance of the preserved voxel index, we project each coefficient vector in the coefficient matrix back to the fMRI image space, which results in a coefficient map associated with that atom. The spatial pattern has the same spatial dimension as that of the input fMRI. We use a classic *t*-statistic analysis as same as that used in the classic GLM based analysis, which converts the coefficient map of an atom to a T-statistic map, to assess the significance level of the contribution of the atom in fMRI signal sparse reconstruction. For the ease of group analysis, the T-statistic maps are normalized to the standard MNI brain template using the nonlinear registration implemented in FSL with the help of the acquired T1 structural image.

C. Sparse Encoding of Video Saliency Curves

fMRI decoding models are typically used to explore the relationship between functional brain responses and the external stimuli. The introduction of sparsity constraints can significantly decrease the complexity of the decoding models and increase the interpretability of the results [32, 36]. *In our study, the learned atomic fMRI dictionary contains the representative brain activity (BOLD signal) patterns with largely reduced dimensions of the original fMRI data, and the corresponding spatial maps reveal the functional subdivision where the brain activities represented by the atoms originate. We borrow the hypothesis in brain decoding studies that if the combination of the brain activities from some brain regions can reconstruct the profile of the external stimuli, then those brain regions may play critical roles in the perception of the external stimuli. Thus, by reconstructing the video saliency curve using the learned fMRI dictionary under sparsity constraint and inspecting into the contribution of the atoms in video saliency curve reconstruction, it is feasible to explore the functional subdivisions engaged in visual saliency perception during natural viewing conditions. In this context, sparse representation serves as a bridge that directly links the representative brain activities with the video saliency curve.* In this study, we use the learned atomic fMRI dictionary to reconstruct the video saliency curve under a sparse representation scheme, which is the same as that in fMRI signal reconstruction in the dictionary learning stage formulated in Eq. (2). The Pearson correlation coefficient between the original video saliency curve and the reconstructed video saliency curve is used as a metric to evaluate the performance of sparse representation in our study. Under the sparse representation scheme, a few atoms have nonzero coefficients in the estimated coefficient vector. This subset of atoms is considered to have contribution to the reconstruction of the video saliency curves. Similarly, the T-value associated with the atoms in the video curve reconstruction is calculated using the classic T-statistic analysis.

D. Group-wise Statistical Analysis

In our study, the fMRI atomic dictionary training and the video saliency curve reconstruction are performed for each individual fMRI data separately, which is similar with that in an ICA-based study for functional brain mapping under movie watching. Thus, the correspondences of the atoms across different fMRI atomic dictionaries are still missing. To solve this problem, we use the representative 20 functional brain networks achieved in a resting state fMRI study [51] as the structural spatial guidance. Specifically, the spatial overlap rates between the T-statistic map associated with a given atom in a dictionary and the 20 representative resting state networks (RSNs) are calculated. The atom is then labeled by the RSN that has the largest overlap rate. With the established correspondences between the atoms across different dictionaries, the significance or contribution of the atoms in video saliency reconstruction are assessed by using a conventional group-wise statistical analysis.

V. RESULTS

A. Atomic fMRI Dictionary Learning

The preset parameters in sparse representation include the size of the dictionary (number of atoms in the dictionary, m), and the balance between sparsity and the residual error of fMRI signal reconstruction (λ). In our experiments, they are experimentally set as follows: m=200, λ=0.01.

To evaluate the performance of sparse representation of the fMRI signals, we calculate the residual errors, as well as the Pearson correlation coefficients between the original fMRI signals and the reconstructed signals. The Pearson correlation coefficients and the residual error for one randomly selected fMRI data are shown in Fig. 3(a) and Fig. 3(b), respectively. It is seen that the reconstructed fMRI signals are well correlated with the original fMRI signals, and the residual errors are relatively small. Further inspection of the distribution of the Pearson correlation coefficient shows that 98.45% of the voxels are with the correlation over 0.8. We can see that the fMRI signals for the voxels in gray matter can be better reconstructed compared with those in white matter and cerebrospinal fluid (CSF). In Fig. 3(c), we show the fMRI signal reconstruction for two exemplar voxels A and B. The location of voxel A and B is referred to Fig. 3(a). Voxel A is in the primary visual cortex and voxel B is in white matter. In Fig. 3(c), the brown curves are the original fMRI signals and the blue curves are the reconstructed fMRI signals. As can be seen in Fig. 3(c), the fMRI signals can be almost fully reconstructed for voxel A, and relatively well reconstructed for voxel B.

[Fig. 3 here]

In comparison, the Pearson correlation coefficient map and the residual error map for the ICA-based decomposition of the same fMRI data are shown in Fig. 4(a) and Fig. 4(b), respectively. We can see from Fig. 3 and Fig. 4 that the fMRI signal reconstruction using sparse representation is much better when compared with ICA. More specifically, none of the voxels is with the correlation over 0.8. In comparison, 98.45% of the voxels are with the correlation over 0.8 in sparse representation based fMRI signal reconstruction.

[Fig. 4 here]

In Fig. 5, we show some exemplar spatial maps *and the corresponding basis of brain responses* of the atoms in a dictionary. The spatial maps include the primary auditory cortex (Fig. 5(a)), the primary visual cortex (Fig. 5(b)), the visual motion perception cortex (Fig. 5(c)), the salience network (Fig. 5(d)), the posterior default mode network (Fig. 5(e)),the somatosensory cortex (Fig. 5(f)), the motor cortex (Fig. 5(g)), and the cerebellum (Fig. 5(h)), the putamen areas (Fig. 5(i)), and the ventricle (Fig. 5(j)). Quantitatively, we calculated the percentage of the voxels in the spatial maps shown in Fig. 5 falling into the representative RSN templates in [51]. For example, 89.64% of the voxels in Fig. 5(a) fall into the primary auditory cortex, 94.18% of the voxels in Fig. 5(b) belong to the primary visual cortex, 95.15% of the voxels in Fig. 5(c) fall into the visual motion perception cortex, 95.12% of the voxels in Fig. 5(d) fall into the executive control network, 88.63% of the voxels in Fig. 5(e) fall into the default mode network, and 91.44% of the voxels in Fig. 5(f) fall into the somatosensory cortex.

[Fig. 5 here]

In the ICA-based naturalistic stimuli fMRI study [1], the authors proposed two metrics for component selection in single session fMRI analysis, including (i) a plausible distribution of the hottest voxels, i.e., how well the hottest voxels are spatially clustered, or spatial aggregativity; (ii) bilaterality. By visual inspection, the exemplar spatial maps shown in Fig. 5, which are among the most representative RSNs, have good performance regarding the abovementioned two metrics.

B. Spare Encoding of Video Saliency Curve

Fig. 6 shows an example of the original video saliency curve (blue) and the reconstructed video saliency curve (brown) in a single fMRI data via sparse representation. The Pearson correlation coefficient between the reconstructed video saliency curve and the original video saliency curve is 0.6837, indicating that the video saliency curve derived from the naturalistic video stimulus by computation video saliency model can be well reconstructed from the learned atomic fMRI dictionary.

[Fig. 6 here]

More results on video saliency curve reconstruction are reported in Fig. 7 and Table 1. For each video, we reconstruct the video saliency curve by the learned atomic dictionary of the 4 subjects independently. The Pearson correlation coefficients between the reconstructed video saliency curves and the original video saliency curves are calculated. In comparison, the performance of video saliency curve reconstruction using ICA components is also reported. For the purpose of fair comparison, the video curve reconstruction using ICA components is regulated by the same sparsity constrain as that in sparse representation. From the reconstruction performance we can conclude: (i) the 4 video saliency curves can be relatively well reconstructed from the learned atomic fMRI dictionary under the sparse representation scheme. The highest and the lowest Pearson correlation between the reconstructed and the original video saliency curve is 0.7509 and 0.5208, respectively. Considering the noises in video saliency curve modeling, the performance of video saliency curve reconstruction via sparse representation is remarkable. (ii) The reconstruction performance of sparse representation is significantly higher than that of ICA. We performed a one-tail *t*-test on the concatenated correlation vectors. The *p*-value is 6.7×10^{-6}. (iii) Despite the discrepancies of performance in video saliency curve reconstruction between sparse representation based and ICA-based methods, the trends of the variation across videos are similar. For example, both sparse representation and ICA achieve the best video saliency curve reconstruction for the second video sample, while the worst performance for the third video sample.

[Fig. 7 here]

[Table 1 here]

C. Impact of Dictionary Size in Dictionary Learning

Parameter selection in applications of the sparse representation algorithm is still an open problem. No method has been proposed to find a set of theoretically optimal parameters.

Experimental and empirical parameter selections are typically used in existing studies. In our study, the selection of dictionary size is the trade-off between model complexity, fMRI signal reconstruction and video saliency curve reconstruction performance. Dictionary with small size may fail to cover the brain responses captured by the fMRI data and consequently is unable to accurately recover the original fMRI signals. However, increasing the dictionary size would substantially increase the computational cost. For fair comparisons, in our previous experiments we set the number of independent components (nIC) in ICA as same as the size of the dictionary in sparse representation.

We performed two experiments on a randomly selected fMRI data (acquired when the first subject were watching the first video clip) to demonstrate the impact of the dictionary size in sparse representation and nIC in ICA. In the experiments, we varied the dictionary size/nIC from 50 to 400 with the step length of 50. In addition, a few methods such as minimum description length (MDL) [52]have been employed for automatic selection of nIC for fMRI analysis. In the experiments, we also tested the automatically estimated nIC, which is 37, for ICA-based method and sparse representation-based method.

In the first experiment, we show how the performance of fMRI signal reconstruction changes with different dictionary sizes and different nICs in ICA. The plots of average Pearson correlation coefficient between the reconstructed fMRI signals and the original fMRI signals, and the average residual error in fMRI signal reconstruction against the dictionary size and nIC in Fig. 8(a) and Fig. 8(b), respectively. The average Pearson correlation coefficient increases while the average residual *error decreases with the increasing of dictionary size in the sparse representation-based method. In comparison, both the average Pearson correlation coefficient and the average residual error are about to be stable when nIC is beyond 150 in the ICA-based method.*

[Fig. 8 here.]

In the second experiment, we show how the performance of video saliency curve reconstruction changes with different dictionary sizes and different nICs in ICA. The performance of video saliency curve reconstruction is measured by the Pearson correlation coefficients between the reconstructed video saliency curve and the original video saliency curve. The experimental results are summarized in Fig. 9. The results show that the performance of video curve reconstruction increases with the dictionary size and nIC. However, when the dictionary size or nIC is beyond 200, the improvement is non-substantial.

[Fig. 9 here.]

From Fig. 1 and Fig. 2 we can see that increasing the dictionary size can improve the performance of fMRI signal reconstruction, however, its contribution to the improvement of video saliency curve decoding is limited. It is also seen that sparse representation-based method outperforms ICA-based method in both fMRI signal reconstruction and video saliency curve reconstruction experiments. Considering the consequent computational cost in sparse representation using larger dictionary size, we use 200 as dictionary size in our study.

D. Consistent Brain Subdivisions in Visual Saliency Encoding

In Fig. 10, we show all the spatial maps, coefficients and T-scores in the video saliency curve reconstruction for one single fMRI data. The order of the spatial maps in Fig. 10(a) is according to the descending order of the T-scores shown in Fig. 10(c). Those spatial maps cover several representative RSNs, for example, the primary visual cortex (#1, #5, #7, #15, #19), the motion perception visual cortex (#8 and #9), the auditory cortex (#2), the executive control network (#13), the somatosensory cortex (#20), the dorsal attention network (#22), the frontoparietal network (#12), white matter area (#29) and the cerebellum (#21). In comparison, the spatial maps, coefficients and T-scores of the video saliency curve reconstruction using ICA-based method for the same fMRI data are shown in Fig. 11. Several well-known cortical areas can also be seen in Fig. 11(a), for example, the primary visual cortex (#23), the primary auditory cortex (#4), part of the default mode network (#1), and white matter area (#30).

[Fig. 10 here]

[Fig. 11 here]

To evaluate the inter-subject consistency of the atomic spatial patterns involved in the video saliency curve reconstruction, we calculated the probability of the 20 representative RSNs [51] over the set of 16 fMRI data. Fig. 12 shows the results for sparse representation (Fig. 12(a)) and ICA (Fig. 12(b)) based methods. Ten RSNs show high inter-subject consistency (above 80%) in sparse representation based video saliency curve reconstruction. In contrast, the inter-subject consistency in ICA-based methods is relatively low, partly indicating the superiority of sparse representation in natural stimuli fMRI analysis compared with ICA-based method. Those atoms with inter-subject consistency below 50% are discarded in our further studies.

[Fig. 12 here]

To assess the significance of the atoms with inter-subject consistency of over 50%, we show the T-scores and groupwise T-scores of those atoms in the video saliency curve reconstruction in Fig. 13. In Fig. 13(a), the blue line shows the mean and standard deviation of the T-scores, and the brown line shows the group-wise T-scores for sparse representation based method. The sorted groupwise T-scores and corresponding RSNs indices are shown in Fig. 13(b). The results for ICA-based video saliency curve reconstruction are shown in Fig. 13(c) and Fig. 13(d). The 3D volume rendering of the average spatial maps for those atoms is shown in Fig. 14 for sparse representation based method and in Fig. 15 for ICA-based method. The order of the spatial maps is according to the significance indicated by the sorted group-wise T-scores shown in Fig .13(b) and Fig. 13(d), respectively.

[Fig. 13 here]

[Fig. 14 here]

The spatial maps shown in Fig. 14 cover a number of representative brain networks. Specifically, there are three visual related spatial maps included the dorsal V1 area (ranked No. 1), the ventral V1 (ranked No. 3) and the MT/MST areas (ranked No. 4) in both hemispheres. The involvement of V1 cortex in bottom-up visual saliency perception has been well documented in the literature. For example, Zhang et al. demonstrated that the neural activities in V1 contribute to the creation of a bottom-up visual saliency map by a study jointly using event related potential (ERP) and fMRI [53]. The role of MT/MST in visual saliency perception has not been well documented yet. However, it is reasonable to observe the contribution of MT/MST in visual saliency reconstruction in our experiments since motion and flicker information plays an important role in the computational video saliency model that generates the video saliency curves [44]. The spatial map ranked No. 2 is known as the frontoparietal network, which covers the bilateral intraparietal sulcus (IPS) and frontal eye fields (FEF). The frontoparietal network has been widely reported in the literature to be involved in targeting eye movements and allocating covert attention. Specifically, anterior IPS (aIPS) is sensitive to bottom-up attentional influences driven by stimulus salience [54], while FEF plays an important role in the control of eye movements [55, 56]. In addition, the neuronal processing of visual saliency is believed to comprise two stages: (i) the graded representation of saliency and (ii) the winner-take-all representation of the maximally salient position in the visual field [21, 57]. In an fMRI-based study in which static images were used as naturalistic stimuli and the computational visual saliency model [21, 22] was adopted to model the visual saliency in the stimuli, Bogler et al. demonstrated that the first stage correlates with the early visual cortex and the posterior IPS (pIPS), and the second stage is encoded in anterior IPS (aIPS) and FEFs [20]. In coincidence, in our study, the spatial maps related to the visual cortex (ranked No. 1, 4 and 6) and the IPS-FEF network (ranked No. 2) are among the most important ones in video saliency curve reconstruction.

In addition, Fig. 13 indicates that some other brain regions may contribute to the visual saliency perception, including the fifth-eighth spatial maps. The fifth spatial map is strongly lateralized and is known as the right-sided frontal-parietal network (or right executive network) [51]. The sixth spatial map corresponds to the somatosensory cortex. The seventh spatial map includes the primary and association auditory cortices. The eighth spatial map covers part of the anterior cingulate cortex (ACC) and the middle frontal gyrus and is known as the salience network. How those subdivisions participate in the saliency perception under dynamic and complex naturalistic video stimuli is still largely unknown, and thoroughly addressing this problem is out of the scope of this paper. Essentially, our experimental results may provide new clues for further exploration of the functional mechanism of saliency perception when the human are exposed to dynamic and naturalistic scenes. For example, previous study has demonstrated that early somatosensory cortex carries content-specific information that discriminates familiar visual object categories [58], and object-based visual attention might attribute to the involvement of the somatosensory cortex in our experimental results. In addition, one of the important functions of the salience network is related to emotion. We speculate that the involvement of the salience network in video attention may attribute to the emotion's regulation on visual attention, which has been shown by growing evidence in the literature [59].

Fig. 14 also shows that several other brain regions (networks) have limited contribution to video saliency curve reconstruction, including the tenth and the twelfth spatial maps (whose functions have not been well defined yet), the eleventh spatial map (is known as the left-sided fronto-parietal network), the thirteenth spatial map (corresponds to the posterior DMN), and the fourteenth spatial map (covers part of the cerebellum). It has been reported in [5] that the DMN can be robustly identified when exposing participants to a continuous segment of an audiovisual movie, however, its activities are with low intra-subject and inter-subject correlations, and hence are believed to be dissociated from the external stimulation.

[Fig. 15 here]

In the ICA-based method, the spatial maps shown in Fig. 15 also cover some representative brain regions (networks), including the ventral V1 (ranked No. 1), the dorsal V1 (ranked No. 2), the right-sided front-parietal network (ranked No. 3), the frontoparietal network (ranked No. 4), the auditory cortex (ranked No. 5), the somatosensory cortex (ranked No. 6), the posterior DMN (ranked No. 7), the cerebellum (ranked No. 10), the salience network (ranked No. 11), the left-sided front-parietal network (ranked No. 12). It is seen that the order of the spatial maps in ICA-based maps is similar to that in sparse representation based method to some extent. For example, both of the results in Fig. 14 and Fig. 15 showed that the primary visual cortex and the frontoparietal network have the most significant contribution to video saliency reconstruction. However, the spatial aggregativity and bilaterality of those spatial maps shown in Fig. 15 are not as good as those shown in Fig. 14 by visual inspection.

VI. DISCUSSION AND CONCLUSION

We proposed a novel sparsity-constrained fMRI decoding model to explore whether bottom-up visual saliency in continuous and naturalistic video streams can be effectively decoded by brain activities recorded by fMRI, and to examine whether sparsity constraints can improve the performance of visual saliency decoding model. The sparsity constraints in our study are two folds. First, we used a sparse representation scheme to learn representative and atomic BOLD signal patterns (code-book) for the whole-brain fMRI data with the purpose of dimension reduction. Second, the video saliency curves derived from video stream via computational video saliency model were reconstructed under sparsity constraints. Our experimental results have demonstrated that: 1) the whole-brain fMRI signals can be more accurately represented by the atomic dictionary learned by sparse representation when compared with the widely used ICA-based decomposition; 2) the spatial maps of the atoms resulted from dictionary learning are more reasonable when compared with those of the independent components in ICA-based analysis in terms of spatial aggregativity and bilaterality; 3) the video saliency curves of complex and naturalistic scenes derived via computational video saliency model can be relatively well reconstructed using the learned atomic dictionary of whole-brain fMRI signals, and the performance of video saliency curve reconstruction using sparse representation outperforms that in ICA-based analysis.

Due to the relatively small number of training samples (4 participants, 4 video clips), cross-validation of video saliency curve reconstruction has not been performed in our study. Instead, we firstly identified the atoms with high spatial consistency in video saliency curve reconstruction across sessions of fMRI, and then assessed the significance of their contribution to video saliency curve reconstruction. The atoms with significant contribution to video saliency curve reconstruction included the visual cortex (including ventral and dorsal V1, as well as V4), the frontoparietal network which covers the bilateral IPS and FEF. The results are in line with conventional visual saliency studies using fMRI. In addition, our experimental results indicate that the right-sided frontal-parietal network, the somatosensory cortex, the primary and association auditory cortices and the salience network may contribute to visual saliency perception when participants were exposed to complex and naturalistic scenes. Thorough explanation of how those subdivisions participated in the saliency perception under dynamic and complex naturalistic scenes is out of the scope of this paper. However, our experimental results may provide new clues for further visual saliency studies using functional brain imaging. *Nevertheless, the relatively small number of participants and video samples is one of the limitations in our study. The significance level of the reported experimental results might be degenerated by the relatively small number of training samples. A larger number of participants and video samples, as well as cross-validations would be beneficial for more robust and reliable explorations of video saliency perception.*

In the atomic fMRI dictionary learning, the BOLD signal of each voxel was treated as a spatially independent sample. In other words, the spatial information of the voxels has not been taken into account. In machine learning studies, some structured sparsity constraints have been successfully developed. Similar to the probabilistic ICA [60], it is expected that effective integration of spatial information can potentially improve the atomic fMRI dictionary learning in the future. *In addition, due to the high dimension of a typical fMRI data, the computational cost is usually high in fMRI dictionary learning using sparse representation algorithm. In the literature, a few efficient sparse models have been proposed, e.g. [63, 64]. In the future, we plan to use these solutions to accelerate dictionary learning.*

In addition, the setting of the parameters in sparse representation is still an open question in the literature. Similar to ICA-based analysis, the size of the dictionary determines the complexity of the sparse representation model and the integration of the spatial maps in our study. Larger dictionary size will increase the complexity of the model and decrease the integration of the spatial maps. Parameter λ is used to achieve the balance between the signal reconstruction accuracy and the complexity of the model. One of the limitations in our experiments is that those parameters were set experimentally and empirically. In the future, it is desired to use parameter optimization algorithms such as the optimization of the sparsity level based on minimum description length (MDL) [40, 52] and learning the optimal value of λ by a Bayesian approach [61], as well as performing reproducibility studies with regard to parameter settings.

In our study, the spatial maps of the learned atoms were anatomically labeled by referring to the representative ICA resting state templates [51]. Theoretically, it is practical to identify the correspondences of the atoms across different fMRI sessions when the dictionary learning was performed for each fMRI session individually. In addition, group sparse representation analysis of multi-session fMRI data may provide alternative solutions. For example, like in group-ICA approaches, the BOLD signals from multi-session fMRI data can be temporally concatenated [51], or can be analyzed similarly to the tensor-ICA algorithms [62].

Finally, we could like to conclude by predicting that sparsity constrained fMRI decoding models would play more and more important roles in elucidating the interactions between the brain's functional responses and naturalistic stimuli such as video and audio streams. Eventually, quantitative representations of such interactions based on sparse coding would significantly advance neuroscientific understanding of the brain and its comprehension of the real world, as well as benefitting the multimedia content analysis field [65].

ACKNOWLEDGEMENT

X. Hu was supported by the National Science Foundation of China (NSFC) 61103061 and 61473234, and the Fundamental Research Funds for the Central Universities 3102014JCQ01065. L. Guo was supported by NSFC 61273362 and 61333017. J. Han was supported by NFSC 91120005. T. Liu was supported by the NIH Career Award (NIH EB006878), NIH R01 DA-033393, NIH R01 AG-042599, NSF CAREER Award IIS-1149260, NSF CBET-1302089, and NSF BCS-1439051.

REFERENCES

[1] A. Bartels and S. Zeki, "The chronoarchitecture of the human brain - natural viewing conditions reveal a time-based anatomy of the brain," *Neuroimage,* vol. 22, pp. 419-433, 2004.

[2] U. Hasson, O. Furman, D. Clark, Y. Dudai, and L. Davachi, "Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding," *Neuron,* vol. 57, pp. 452-462, 2008.

[3] C. Bordier, F. Puja, and E. Macaluso, "Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging," *Neuroimage,* vol. 67, pp. 213-226, 2013.

[4] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology,* vol. 21, pp. 1641-1646, 2011.

[5] Y. Golland, S. Bentin, H. Gelbard, Y. Benjamini, R. Heller, Y. Nir, U. Hasson, and R. Malach, "Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation," *Cerebral Cortex,* vol. 17, pp. 766-777, 2007.

[6] P. Toiviainen, V. Alluri, E. Brattico, M. Wallentin, and P. Vuust, "Capturing the musical brain with Lasso: dynamic decoding of musical features from fMRI data," *Neuroimage,* vol. 88, pp. 170-180, 2013.

[7] V. Alluri, P. Toiviainen, I. P. Jaaskelainen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm," *Neuroimage,* vol. 59, pp. 3677-3689, 2012.

[8] P. Janata, J. L. Birk, J. D. Van Horn, M. Leman, B. Tillmann, and J. J. Bharucha, "The cortical topography of tonal structures underlying Western music," *Science,* vol. 298, pp. 2167-2170, 2002.

[9] J. Ylipaavalniemi, E. Savia, S. Malinen, R. Hari, R. Vigario, and S. Kaski, "Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli," *Neuroimage,* vol. 48, pp. 176-185, 2009.

[10] H. Yao, L. Shi, F. Han, H. Gao, and Y. Dan, "Rapid learning in cortical coding of visual scenes," *Nature*

neuroscience, vol. 10, pp. 772-778, 2007.

[11] U. Hasson, R. Malach, and D. J. Heeger, "Reliability of cortical activity during natural stimulation," *Trends in cognitive sciences,* vol. 14, pp. 40-48, 2010.

[12] A. Bartels, S. Zeki, and N. K. Logothetis, "Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain," *Cerebral Cortex,* vol. 18, pp. 705-717, 2008.

[13] A. Bartels and S. Zeki, "Brain dynamics during natural viewing conditions - a new guide for mapping connectivity in vivo," *Neuroimage,* vol. 24, pp. 339-349, 2005.

[14] A. Bartels and S. Zeki, "Functional brain mapping during free viewing of natural scenes," *Human brain mapping,* vol. 21, pp. 75-85, 2004.

[15] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science,* vol. 303, pp. 1634-1640, 2004.

[16] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human brain mapping,* vol. 14, pp. 140-151, 2001.

[17] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms," *Human brain mapping,* vol. 13, pp. 43-53, 2001.

[18] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI Data by Blind Separation Into Independent Spatial Components," *Human brain mapping,* vol. 6, pp. 160-188, 1998.

[19] D. Nardo, V. Santangelo, and E. Macaluso, "Stimulus-driven orienting of visuo-spatial attention in complex dynamic environments," *Neuron,* vol. 69, pp. 1015-1028, 2011.

[20] C. Bogler, S. Bode, and J.-D. Haynes, "Decoding successive computational stages of saliency processing," *Current Biology,* vol. 21, pp. 1667-1671, 2011.

[21] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, pp. 194-203, 2001.

[22] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 20, pp. 1254-1259, 1998.

[23] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology,* vol. 15, pp. 1943-1947, 2005.

[24] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science,* vol. 293, pp. 2425-2430, 2001.

[25] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in cognitive sciences,* vol. 10, pp. 424-430, 2006.

[26] P. K. Douglas, S. Harris, A. Yuille, and M. S. Cohen, "Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief," *Neuroimage,* vol. 56, pp. 544-553, 2011.

[27] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *Neuroimage,* vol. 56, pp. 400-410, 2011.

[28] H. Huttunen, T. Manninen, J.-P. Kauppi, and J. Tohka, "Mind reading with regularized multinomial logistic regression," *Machine Vision and Applications,* pp. 1-15, 2012.

[29] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning,* vol. 57, pp. 145-175, 2004.

[30] B. Ng, G. Varoquaux, J.-B. Poline, and B. Thirion, "A novel sparse graphical approach for multimodal brain

connectivity inference," presented at the Medical Image Computing and Computer-Assisted Intervention, MICCAI 2012, 2012.

[31] Y. Li, P. Namburi, Z. Yu, C. Guan, J. Feng, and Z. Gu, "Voxel selection in fMRI data analysis based on sparse representation," *Biomedical Engineering, IEEE Transactions on,* vol. 56, pp. 2439-2451, 2009.

[32] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *Neuroimage,* vol. 51, pp. 752-764, 2010.

[33] O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *Neuroimage,* vol. 42, pp. 1414-1429, 2008.

[34] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'Ardenne, W. Richter, J. D. Cohen, and J. Haxby, "Independent component analysis for brain fMRI does not select for independence," *Proceedings of the National Academy of Sciences,* vol. 106, pp. 10415-10422, 2009.

[35] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on,* vol. 52, pp. 1289-1306, 2006.

[36] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in neural information processing systems*, 2006, pp. 609-616.

[37] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 210-227, 2009.

[38] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research,* vol. 11, pp. 19-60, 2010.

[39] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE,* vol. 98, pp. 1031-1044, 2010.

[40] K. Lee, S. Tak, and J. C. Ye, "A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *Medical Imaging, IEEE Transactions on,* vol. 30, pp. 1076-1089, 2011.

[41] V. P. Oikonomou, K. Blekas, and L. Astrakas, "A sparse and spatially constrained generative regression model for fMRI data analysis," *Biomedical Engineering, IEEE Transactions on,* vol. 59, pp. 58-67, 2012.

[42] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Fast and incoherent dictionary learning algorithms with application to fMRI," *Signal, Image and Video Processing,* pp. 1-12, 2013.

[43] J. Lv, X. Li, D. Zhu, X. Jiang, X. Zhang, X. Hu, T. Zhang, L. Guo, and T. Liu, "Sparse Representation of Group-Wise FMRI Signals," presented at the Medical Image Computing and Computer-Assisted Intervention, MICCAI 2013, 2013.

[44] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," presented at the Advances in neural information processing systems, 2006.

[45] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 2006, pp. 321-330.

[46] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research,* vol. 49, pp. 1295-1306, 2009.

[47] K. J. Friston, A. P. Holmes, K. J. Worsley, J. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human brain mapping,* vol. 2, pp. 189-210, 1994.

[48] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing,* vol. 20, pp. 33-61, 1998.

[49] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics,* vol. 59, pp. 797-829, 2006.

[50] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares

problems," *Ima Journal of Numerical Analysis,* vol. 20, pp. 389-403, Jul 2000.

[51] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, and A. R. Laird, "Correspondence of the brain's functional architecture during activation and rest," *Proceedings of the National Academy of Sciences,* vol. 106, pp. 13040-13045, 2009.

[52] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465-471, 1978.

[53] X. Zhang, L. Zhaoping, T. Zhou, and F. Fang, "Neural activities in V1 create a bottom-up saliency map," *Neuron,* vol. 73, pp. 183-192, 2012.

[54] J. C. Culham and N. G. Kanwisher, "Neuroimaging of cognitive functions in human parietal cortex," *Current opinion in neurobiology,* vol. 11, pp. 157-163, 2001.

[55] J. D. Schall, "On the role of frontal eye field in guiding attention and saccades," *Vision research,* vol. 44, pp. 1453-1467, 2004.

[56] J. W. Bisley, "The neural basis of visual attention," *The Journal of physiology,* vol. 589, pp. 49-57, 2011.

[57] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence,* ed: Springer, 1987, pp. 115-141.

[58] F. W. Smith and M. A. Goodale, "Decoding Visual Object Categories in Early Somatosensory Cortex," *Cerebral Cortex,* p. doi: 10.1093/cercor/bht292, 2013.

[59] H. T. Schupp, J. Stockburger, M. Codispoti, M. Junghofer, A. I. Weike, and A. O. Hamm, "Selective visual attention to emotion," *The Journal of Neuroscience,* vol. 27, pp. 1082-1089, 2007.

[60] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *Medical Imaging, IEEE Transactions on,* vol. 23, pp. 137-152, 2004.

[61] Y. Lin and D. D. Lee, "Bayesian L1-norm sparse learning," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006),* 2006, pp. 605-608.

[62] C. F. Beckmann and S. M. Smith, "Tensorial extensions of independent component analysis for multisubject FMRI analysis," *Neuroimage,* vol. 25, pp. 294-311, 2005.

[63] C. Chen, Y. Q. Li, and J. Z. Huang, "Forest Sparsity for Multi-Channel Compressive Sensing," *Ieee Transactions on Signal Processing,* vol. 62, pp. 2803-2813, Jun 2014.

[64] J. Z. Huang, S. T. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis,* vol. 15, pp. 670-679, Oct 2011.

[65] T. Liu, X. Hu, X. Li, M. Chen, J. Han, and L. Guo, "Merging neuroimaging and multimedia: Methods, opportunities, and challenges," *IEEE Transactions on Human-Machine Systems,* vol. 44, pp. 270-280, 2014.

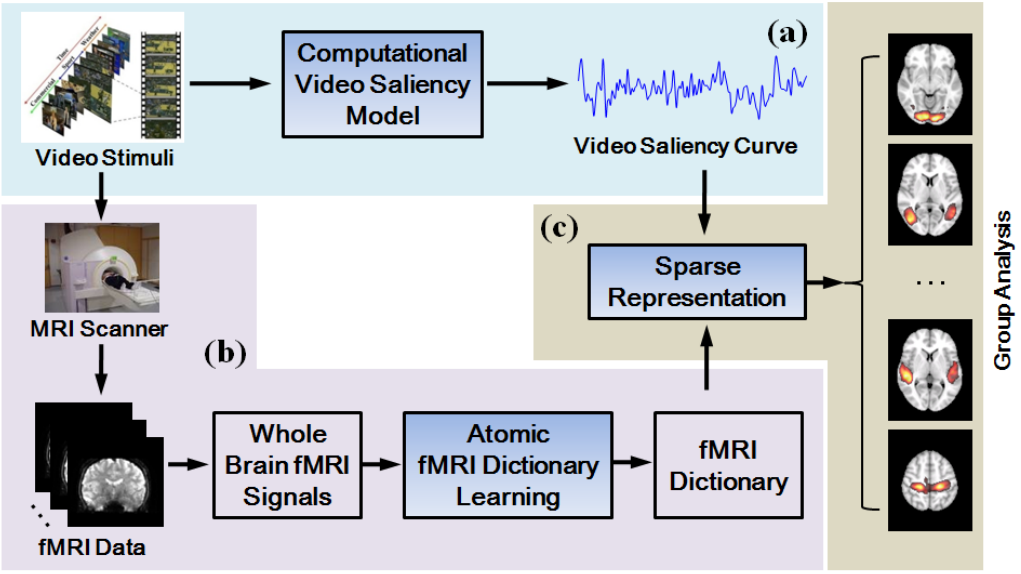


Fig. 1. The overview of our study. (a) A biological-plausible computational video saliency algorithm [44] is adopted to model the video saliency curve for the input naturalistic video stimuli. (b) A sparse representation scheme [38] is adopted to learn the atomic fMRI dictionary for the fMRI data which was acquired when the subject were watching the video stimuli. (c) The learned atomic fMRI dictionary is used to reconstruct the video saliency curve. Group-wise analysis is conducted to infer the consistent functional subdivisions for video saliency curve reconstruction.

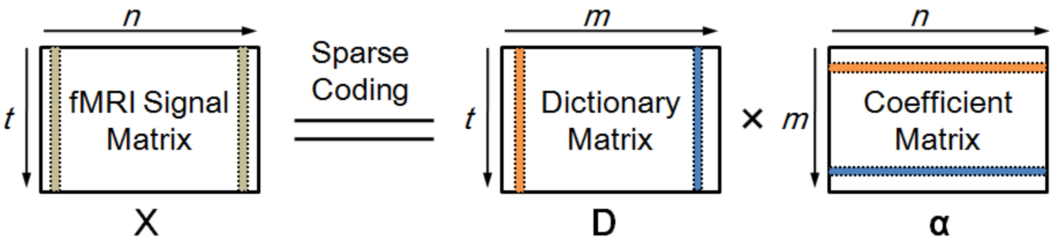


Fig. 2: Illustration of sparse representation based whole-brain fMRI signals analysis. X is the fMRI signal matrix, in which each column is the fMRI signal for a single voxel. D is the atomic fMRI dictionary, in which each column (an atom) in the dictionary corresponds to a representative fMRI signal pattern. α is the coefficient matrix, in which each row is the coefficient vector of the corresponding atom in the reconstruction of the whole-brain fMRI signals.

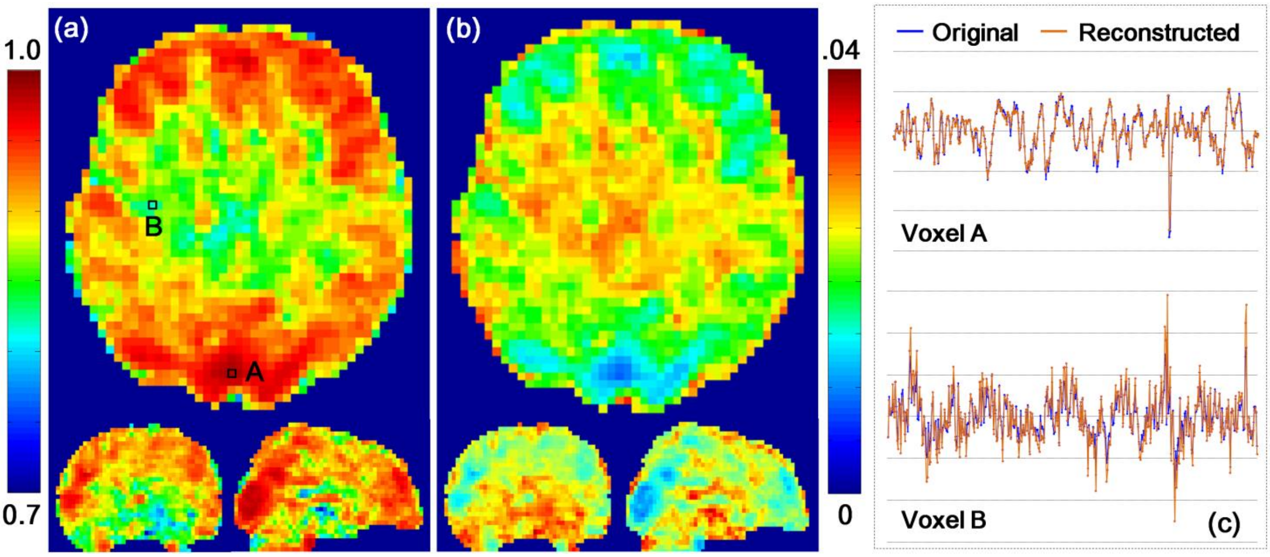


Fig. 3: The illustration of the performance of fMRI signal reconstruction using sparse representation. (a) The Pearson correlation coefficients between the reconstructed fMRI signals and the original signals. (b) The residual error in fMRI signal reconstruction. (c) Two examples for fMRI signal reconstruction. The locations of voxel A and B are referred to (a).

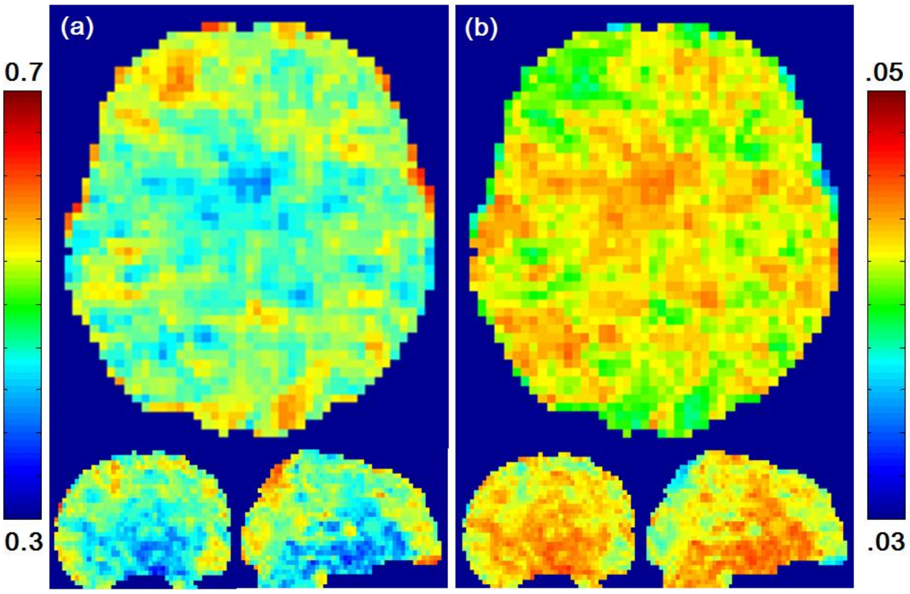


Fig. 4: The illustration of the performance of fMRI signal decomposition via ICA. (a) The Pearson correlation coefficients between the reconstructed fMRI signals and the original signals. (b) The residual error in fMRI signal decomposition.

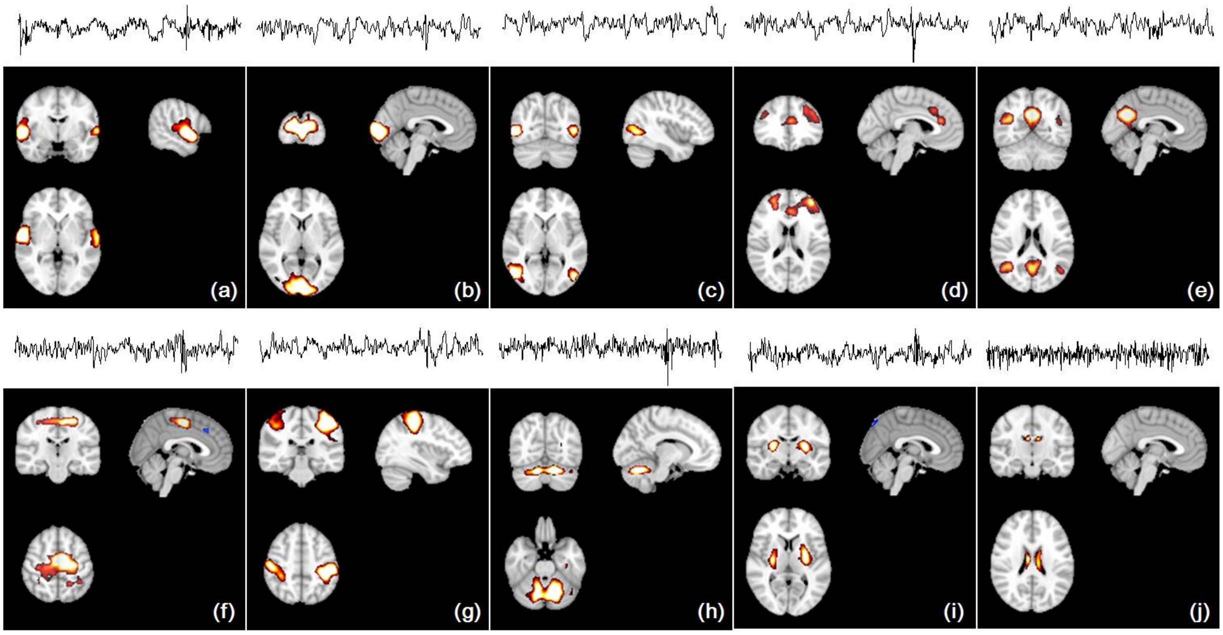


Fig. 5: Exemplar spatial maps *and the corresponding time courses* of the atoms in fMRI signal dictionary. (a) The primary auditory cortex. (b) The primary visual cortex. (c) The visual motion perception cortex. (d) The executive control network. (e) The posterior default mode network. (f) The somatosensory cortex. (g) The dorsal attention. (h) The cerebellum. (i) The putamen areas. (j) The ventricle.

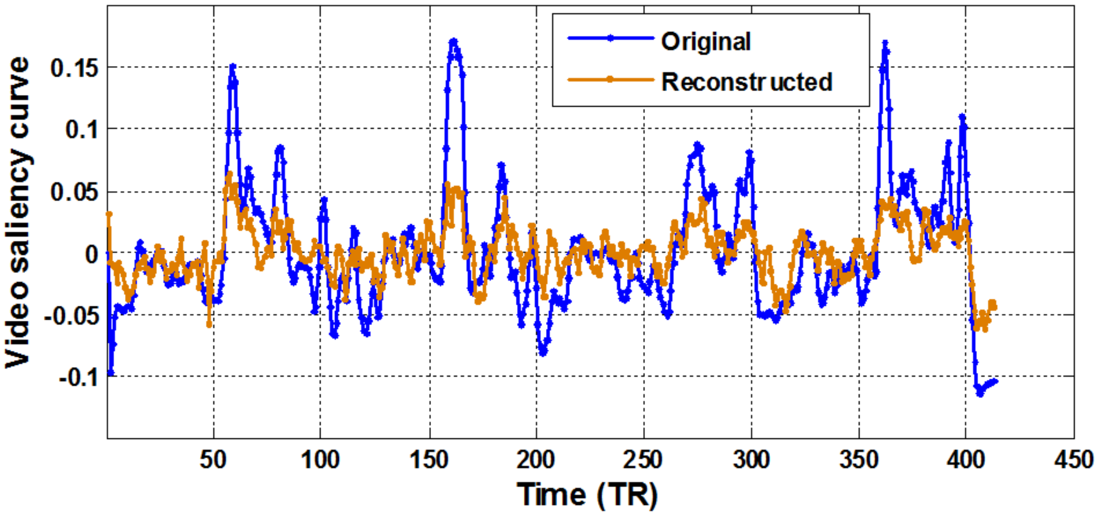


Fig. 6: An example of video saliency curve reconstruction via fMRI atomic dictionary. The Pearson correlation coefficient between the reconstructed video saliency curve and the original video saliency curve is 0.6387.

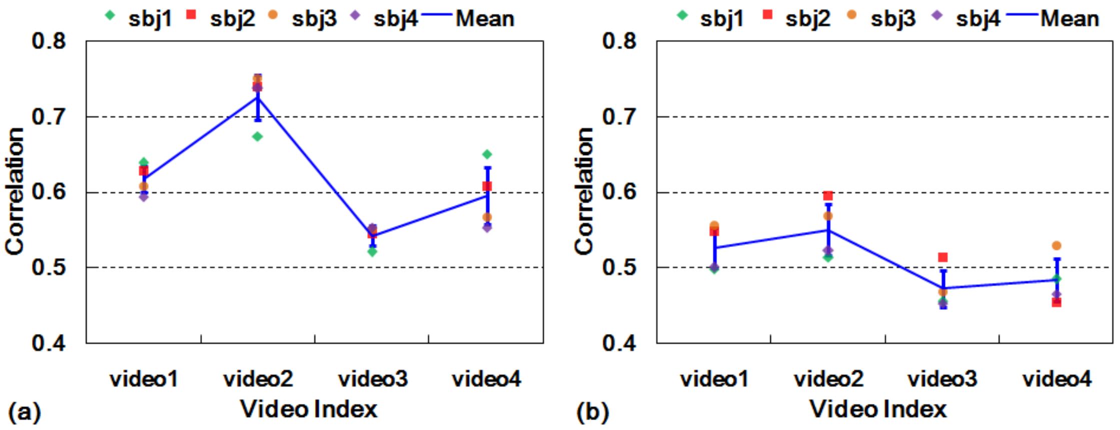
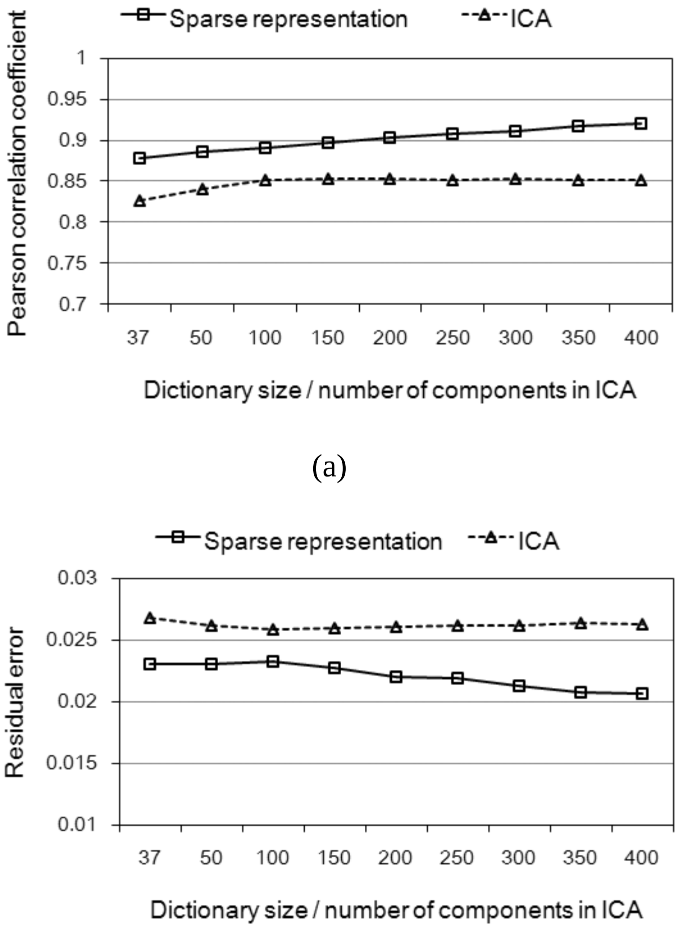


Fig. 7: Performance of video saliency curve reconstruction via sparse representation ((a)) and ICA ((b)).



(a)

(b)

Fig. 8: The impact of dictionary size in sparse representation and number of independent components in ICA on fMRI signal reconstruction. (a) The average Perason coefficient coefficients between reconstructed fMRI signals and origianal fMRI signals. (b) The average residual error in fMRI signal reconstruction.

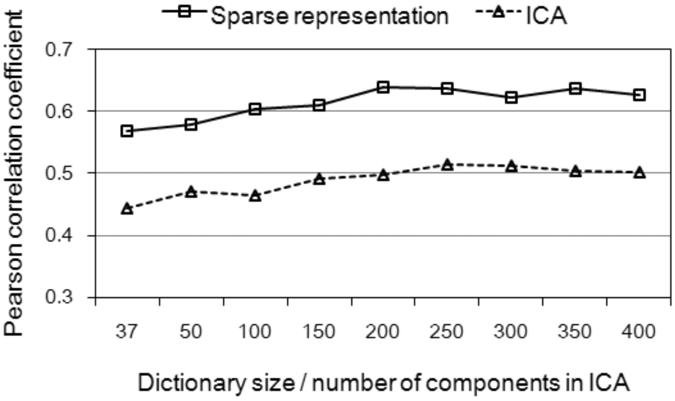


Fig. 9: The impact of dictionary size in sparse representation and number of independent components in ICA on video saliency curve reconstruction. The performance of video saliency curve reconstruction is measured by the Pearson correlation coefficients between the reconstructed video saliency curve and the original video saliency curve.

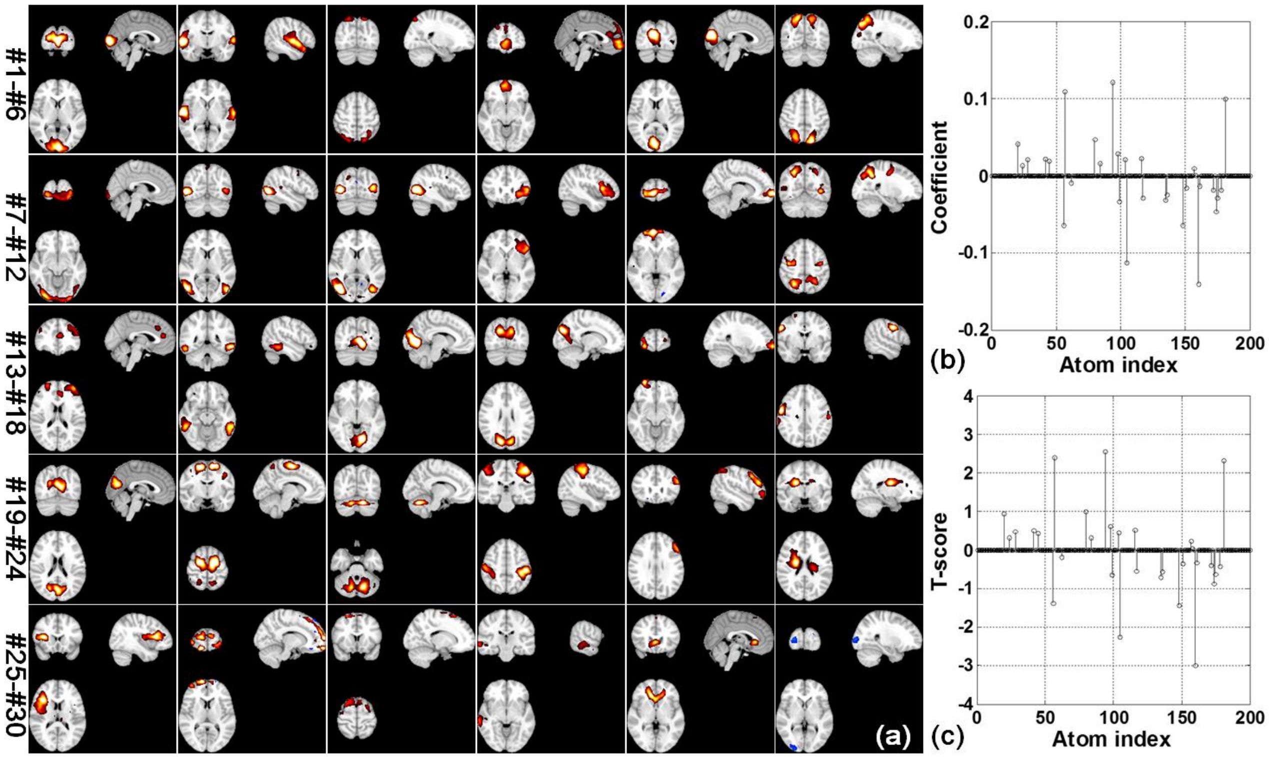


Fig. 10: Exemplar spatial maps ((a)), coefficients ((b)) and T-scores ((c)) for video curve reconstruction via the learned atomic fMRI dictionary. The order of the spatial maps in (a) is according to descend order of the T-scores showed in (c). In (a), '#-' represents the index of the spatial map. Each row contains 6 spatial maps.

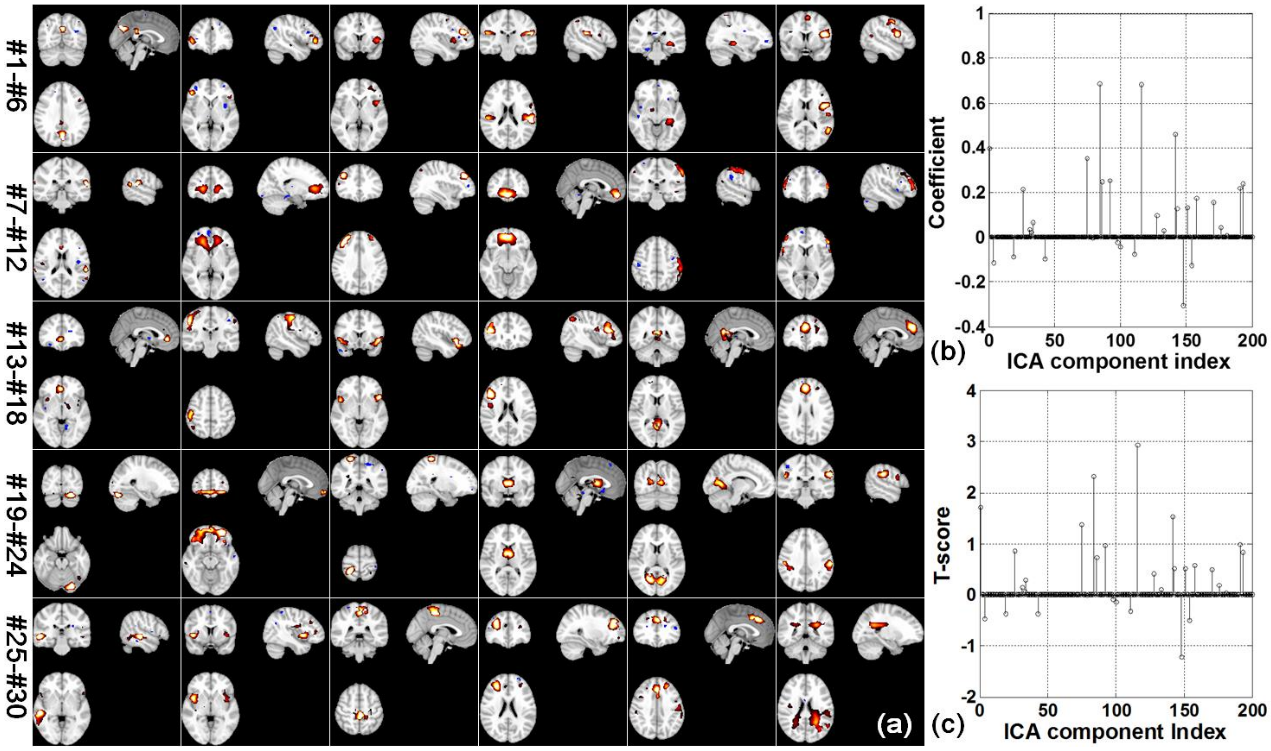


Fig. 11: Exemplar spatial maps, coefficients and T-scores for video curve reconstruction via the ICA components. The order of the spatial maps in (a) is according to descending order of the T-scores showed in (c). In (a), '#-' represents the index of the spatial map. Each row contains 6 spatial maps.

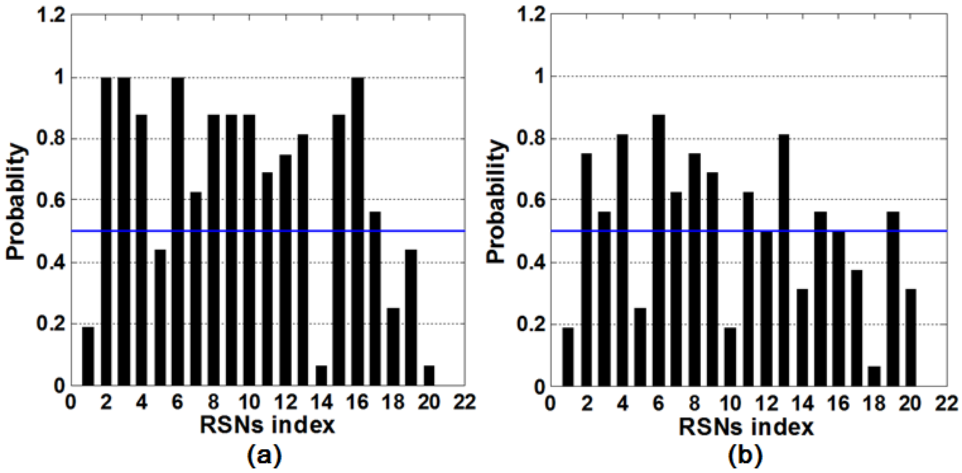


Fig. 12: The probability of the atoms' (a) and components' (b) presence in the total 16 fMRI data.

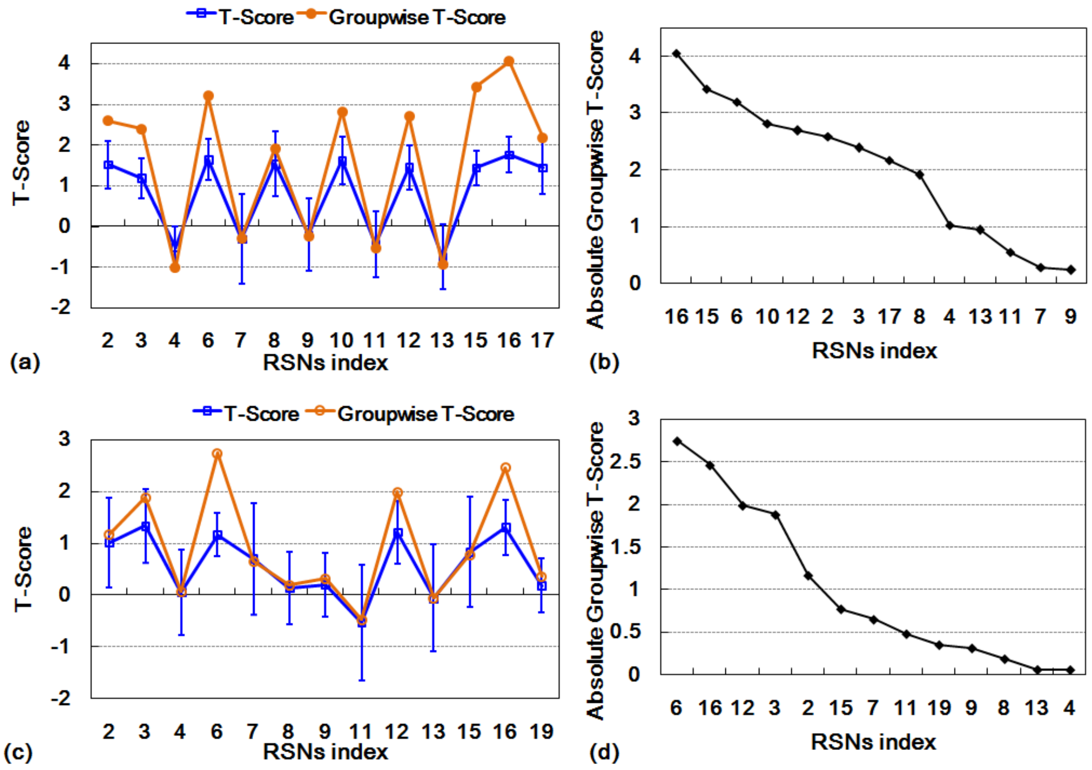


Fig. 13: (a) The T-scores and groupwise T-scores of the atoms in video saliency reconstructions using sparse representation. (b) The descending order of the absolute group-wise T-scores in (a). (c) The T-scores and group-wise T-scores of the components in video saliency reconstructions using ICA. (d) The descending order of the absolute group-wise T-scores in (c).

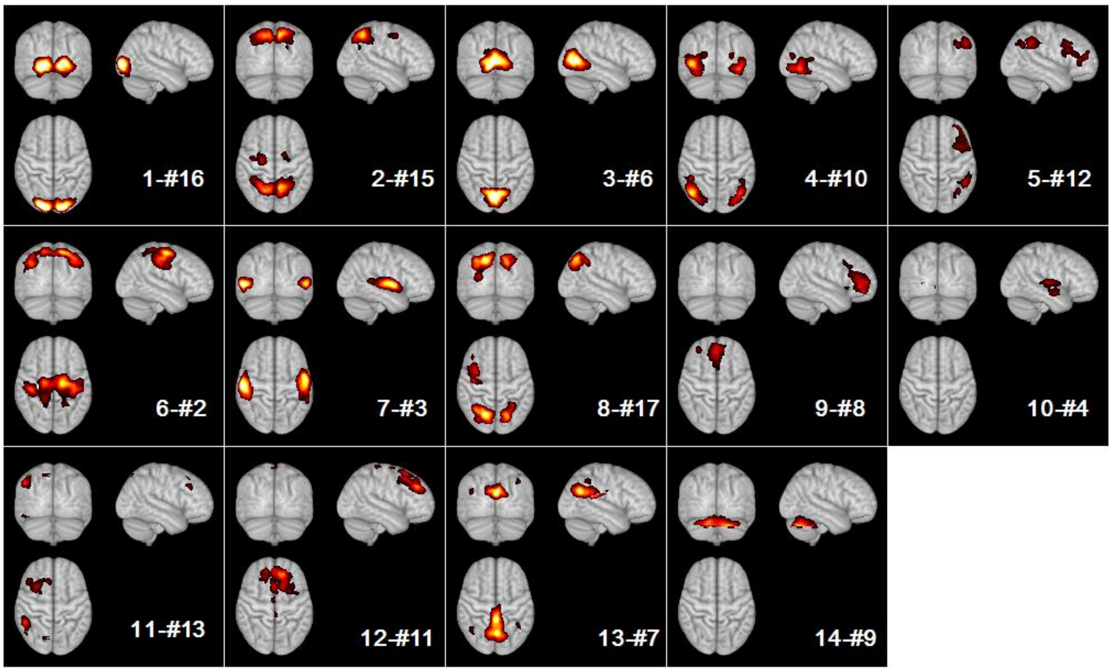


Fig. 14: The average spatial maps of the atoms in video curve reconstruction using sparse representation. The order of the spatial maps is according to the descending order of the absolute group-wise T-scores shown in Fig. 11(b).

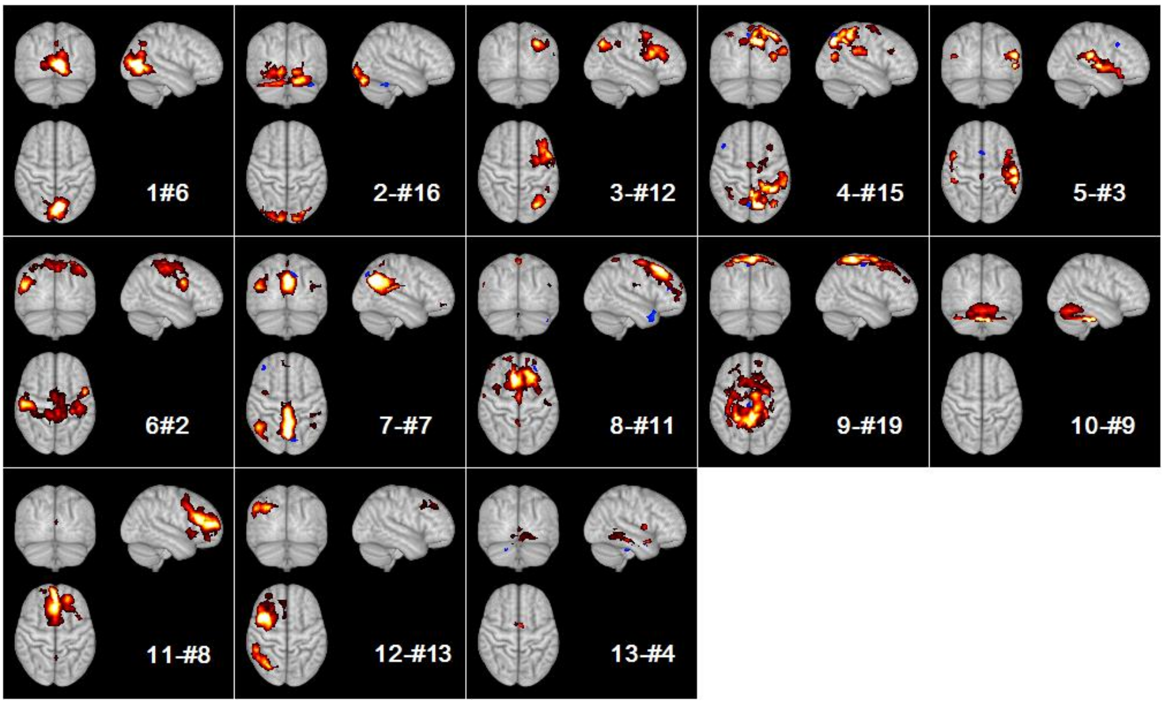


Fig. 15: The average spatial maps of the atoms in video curve reconstruction using ICA. The order of the spatial maps is according to the descending order of the absolute group-wise T-scores shown in Fig. 11(d).

Table 1: Performance of video saliency curve reconstruction. SR denotes sparse representation.

Correlation	**Video1**		**Video2**		**Video3**		**Video4**	
	SR	ICA	SR	ICA	SR	ICA	SR	ICA
Sbj1	0.6387	0.4983	0.6744	0.5137	0.5208	0.4550	0.6506	0.4847
Sbj2	0.6276	0.5490	0.7396	0.5961	0.5444	0.5131	0.6081	0.4546
Sbj3	0.6076	0.5560	0.7509	0.5691	0.5511	0.4688	0.5669	0.5286
Sbj4	0.5938	0.5005	0.7376	0.5226	0.5536	0.4528	0.5537	0.4652
Mean	0.6169	0.5260	0.7256	0.5504	0.5425	0.4724	0.5948	0.4833
Std.	0.0174	0.0267	0.0300	0.0338	0.0130	0.0243	0.0379	0.0283