# Brain State Decoding for Rapid Image Retrieval

Jun Wang[1], Eric Pohlmeyer[2], Barbara Hanna[3]
Yu-Gang Jiang[1], Paul Sajda[2], Shih-Fu Chang[1]
[1]Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
[2]Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA
[3]Meridian Vision, Princeton, NJ, 08542, USA
{jwang, yjiang, sfchang}@ee.columbia.edu
{ep2473,psajda}@columbia.edu
bhanna@meridianvision.com

## ABSTRACT

Human visual perception is able to recognize a wide range of targets under challenging conditions, but has limited throughput. Machine vision and automatic content analytics can process images at a high speed, but suffers from inadequate recognition accuracy for general target classes. In this paper, we propose a new paradigm to explore and combine the strengths of both systems. A single trial EEG-based brain machine interface (BCI) subsystem is used to detect objects of interest of arbitrary classes from an initial subset of images. The EEG detection outcomes are used as input to a graph-based pattern mining subsystem to identify, refine, and propagate the labels to retrieve relevant images from a much larger pool. The combined strategy is unique in its generality, robustness, and high throughput. It has great potential for advancing the state of the art in media retrieval applications. We have evaluated and demonstrated significant performance gains of the proposed system with multiple and diverse image classes over several data sets, including those from Internet (Caltech 101) and remote sensing images. In this paper, we will also present insights learned from the experiments and discuss future research directions.

## Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.2 [**Implementation**]: Special architectures

## General Terms

Algorithms, Design, Experimentation

## Keywords

Image Annotation and Search, Brain Computer Interface, Visual Pattern Mining, Noisy Label Refinement, EEG Signal Decoding


*MM'09,* October 19–24, 2009, Beijing, China.


## 1. INTRODUCTION

The human brain is widely considered to be the most powerful visual information processing system. The human visual system is able to get the "gist" of a scene in a few hundred milliseconds [17, 24, 33]. As a result, many efforts have been made to understand the human vision mechanism by decoding brain state from neural signals. By monitoring the neural response signals, e.g., those recorded noninvasively via electroencephalography (EEG) [3], promising results have been shown in detecting objects of interests (OOI) contained in the visual stimuli presented to subjects [2, 8, 15, 25, 27].

One of the ultimate goals for automated computer vision or media content analysis is to detect and recognize objects, scenes, people, and events in images or videos. Such capabilities, if realized, will greatly enhance the performance and utility of many applications, such as human computer interaction and visual information search. Recently, impressive progress has been reported in the literature, including advances in image feature extraction, visual matching, and object categorization. Several widely participated benchmarking efforts, such as Caltech 101 [6], PASCAL [5], ImageClef [30], and TRECVID [32], have been organized to demonstrate and evaluate the state of the art in this field. A common framework used in such efforts is to learn object models from a pool of training data, which may have been wholly or partly annotated over pre-defined object classes. Such a learning framework has been shown to be powerful. However, it is limited in its scalability to large-scale applications. One of main barriers is the dependence on the manual annotation process, which is laborious and time consuming. To overcome this, efforts have been reported using interactive annotation with relevance feedback and active learning in order to reduce the amount of the required manual input [11, 29]. Recent works have also started to explore the freely available (but imperfect) metadata associated with images on the Web [7, 6, 14, 36].

In this paper, we propose a novel framework that combines the power of brain state decoding and visual content analysis to maximize the efficiency of the image annotation and retrieval task in a completely hand free streamlined fashion. The proposed **B**rain **C**omputer **I**nterface and **V**isual **P**attern **M**ining (BCI-VPM) based image annotation system, as shown in Figure 1, consists of two critical components, EEG-based generic interest detector and graph-based salient visual pattern discovery. The EEG-based interest detector mentioned above is generic - a subject-adaptive

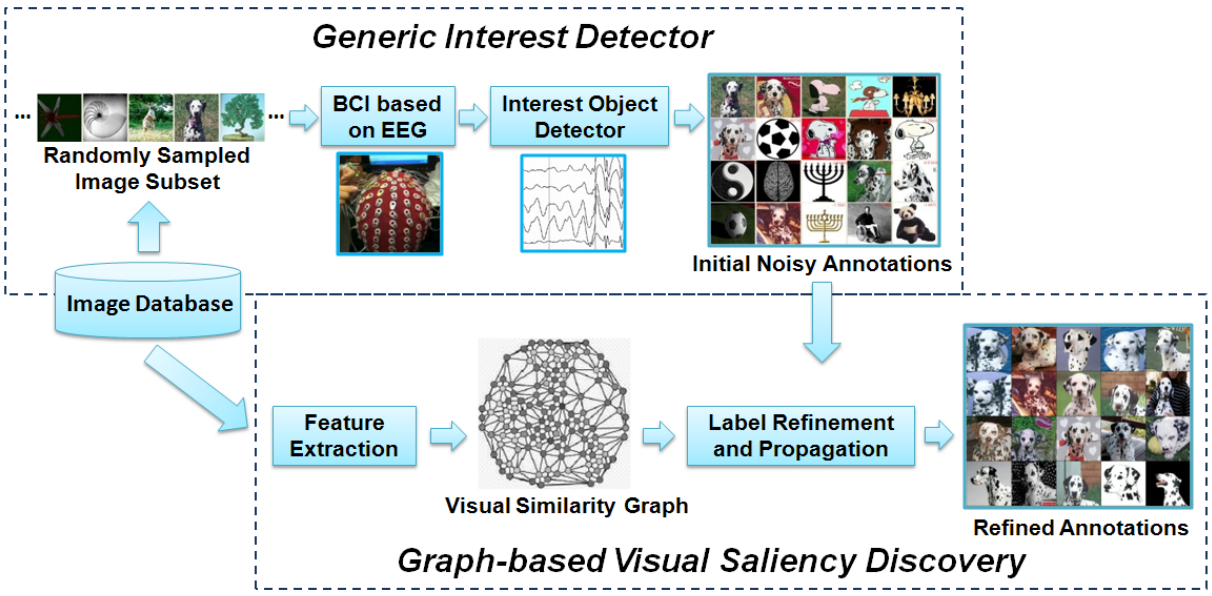


**Figure 1: System diagram of the proposed BCI-VPM image annotation system. A small subset of images is shown to users, whose EEG-based neural response signals are used to detect objects of interest that catch users' attention. The EEG scores of the subset are then refined and propagated to the entire image collection through recovering the visual consistency and discovering the salient visual pattern over a visual similarity graph.**

EEG-based detector trained over a generic class can be applied to detect any new objects of interest. Likewise, the subsequent graph visual saliency discovery is general as it does not assume any prior knowledge about image classes or data sets. Additionally, by completely freeing users from any manual operation (e.g., button pressing) in the viewing stage, we achieve the maximal throughput of annotation or retrieval from a fast stream of image sequence via a process called *rapid serial visual presentation* (RSVP) paradigm [28]. RSVP involves images being flashed to the viewer at high rates, such as 10 frames per second. EEG signal recorded from an array of transducers placed on the scalp of the subject are recorded continuously and analyzed to extract discriminative spatial-temporal features. Techniques like Hierarchical and Bilinear Discriminant Component Analysis [4] can be applied to analyze EEG signal to generate a probabilistic relevance score for each image. The image sequence ordered by the EEG scores is then given further visual content analysis to determine the salient visual patterns that catch users' attention. It utilizes a graph-based diffusion processing to recover visual consistency on a regular graph constructed from visual similarity. Finally, the unreliable EEG labels are eliminated, while the EEG scores associated with the small image subset that have undergone the EEG-based brain state decoding step are propagated to the entire database to retrieve further relevant images, including those not yet shown to users. The regular graph and the diffusion process with label correction have unique power in handling imperfect labels and sparse targets with extreme low prior.

In summary, we propose exploration of an integrated paradigm that marries the strengths of human vision and machine learning systems in a unique and synergistic way - human vision for its superb capability in detecting general objects in diverse and complex conditions and content analytics for automatic processing of large volumes of data.

We will show effectiveness of the proposed approach over images from multiple domains. The first is for the task of target detection in high-resolution satellite imagery, containing 1051 images with "*Helipad*" targets. The second set includes a total of 3798 Internet images with 62 object categories from the well-known Caltech 101 dataset. Experimental results from multiple subjects indicate a very promising performance. In the Internet image experiments, the annotation accuracy, measured in terms of average precision (AP), of one of the object classes *Dalmatian* was improved from 1.76% (random baseline) to 36.7% by EEG interest detector, and further boosted to 69.1% by the visual pattern mining process.

The paper is organized as follows. We present the proposed integrated paradigm in Section 2. The EEG interest detector through decoding brain state is summarized in Section 3. Section 4 presents the method using graph-based approach to discover salient visual pattern with uncertain or unreliable EEG labels. Test case design and evaluation results are included in Section 5. Section 6 includes reviews of works in related fields and discussion of unique contributions of this paper. Discussions and conclusions are summarized in Section 7.

## 2. SYSTEM OVERVIEW

The proposed BCI-VPM image annotation framework consists of two subsystems (as shown in Figure 1) - one using the neural signals measured by EEG to detect generic OOI present in images, while the other using graph-based visual pattern mining methods to refine the detection results from the first subsystem and propagate results to an expanded data set. We provide a brief overview here and give further details on each components in later sections.

For EEG-based OOI detection, a small subset of images (on the order of few hundred) is first randomly sampled from an image collection and presented as visual stimuli to the

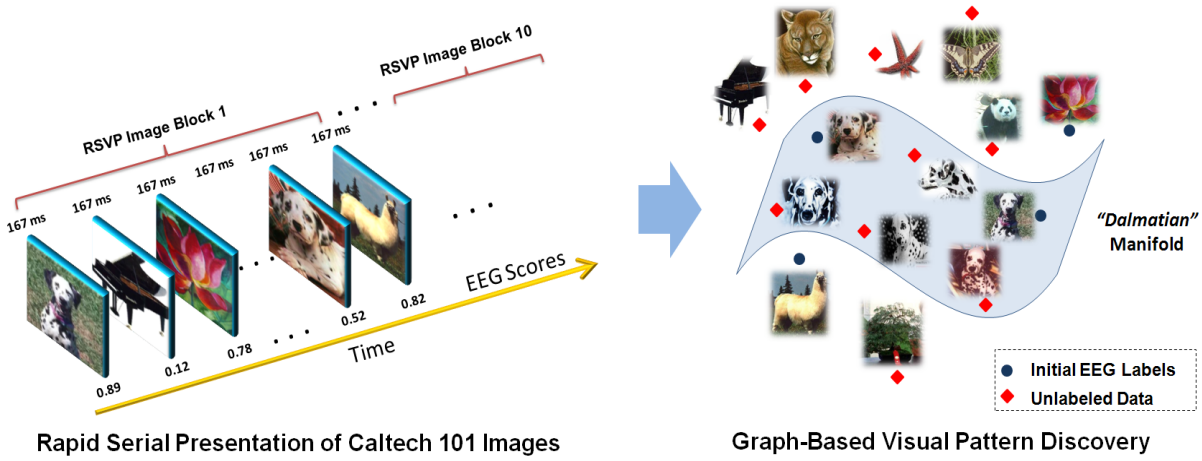


**Figure 2: Overview of the processing pipeline of the proposed BCI-VPM image annotation system. The left demonstrates the RSVP paradigm used for presenting visual stimuli to subjects. The RSVP sequence contains 1000 randomly selected images, which are further partitioned into 10 blocks and 100 images each. An image block typically is shown at $5-10$Hz and each image lasts around $100-200$ milliseconds. The right shows the process of graph based visual pattern mining. After ingesting the estimated EEG "interest" scores as initial labels, the underlying data manifold structure is explored to discover the salient visual pattern among the top EEG score ranked images to refine the initial labels, retrieve additional relevant images, and propagate labels to a much large image pool.**

subject. The selection process avoids long EEG recording sessions which may cause subject fatigue. One of our design goals is to require minimal subject participation, yielding just sufficient information for the neural state decoder and the pattern mining module to effectively infer objects that have attracted a user's attention and generate labels for all the images in the collection.

The sampled images are then presented to the subject in a sequential fashion, following a paradigm called Rapid Serial Visual Presentation (RSVP) [8], as shown in in the left part of Figure 2. The subject is instructed to focus on a fixed marker in the screen center in the first 2 seconds, then each image is shown to the subject for a fixed period of time, ranging from 100ms to 200ms each (equivalent to $5-10$ Hz). The subject may be given instructions ahead of time to look for a specific class of object or simply allowed to choose any object class to their interest on the spot. An array of EEG electrodes placed on the subject scalp are used to continuously record multiple channels of neural response signals (e.g., at 1K sampling rate). Spatio-temporal processing and discriminative analysis are performed on the post-stimulus signals to compute a score, which predicts the confidence in detecting the OOI in each image viewed by the subject. Based on the EEG scores, images are ranked to form initial results, from which top ranked results are used as pseudo positive labels and fed to the pattern discovery module for further refinement and propagation. Note that due to low signal quality and subject variations, the EEG-based OOI detector is pre-trained for each subject. However, such one-time offline training can be done very efficiently without restricting the generality of the detector. As the detector is trained to detect shifts in the user's attention as opposed to detect the recognition of a specific object, an object class uncorrelated with the test objects can be used to train the detector.

The pattern discovery subsystem starts with construction of an affinity graph, which captures the pairwise visual content similarity among nodes (corresponding to images) and the underlying subspace structures in the high dimensional space (as shown in the right part of Figure 2). Such a construction process is done offline before user interaction. The small set of pseudo positive labels generated by the EEG-based interest detector is fed to the initially unlabeled graph as assigned labels for a small number of nodes, which are used to drive the subsequent processes of label identification, refinement, and propagation. Graph based semi-supervised learning techniques [35, 36] play a critical role here since we will rely on both the initial labels and the large pool of unlabeled data points throughout the diffusion process. Finally, the propagated label predictions over the entire graph can be used to generate annotations for every single image in the collection, or to re-rank the images based on the detection scores. The top ranked results, as shown in Figure 6, 7, 8, and 12 (b), are expected to be more accurate (in terms of both precision and recall) than the baseline of using EEG-based detection alone.

## 3. GENERIC INTEREST DETECTOR VIA SINGLE TRIAL EEG DECODING

There has been increasing interest in investigating the application of BCI for image annotation and search [8, 26, 15, 1]. Motivated by humans' ability to make very rapid and accurate visual decisions in "the blink of an eye" [9], we extend the usage of the BCI based image search system in [8] to design an generic interest detector, where users are instructed to look for specific object classes in sequences of images presented using the RSVP paradigm. Examples of segments of the RSVP image sequences used in our experiments are shown in Figure 3. A crucial aspect of this particular annotation system is that we can measure brain signals, in real-time, that can be used to annotate or rank an image given a desired object class. We know from neuroimaging studies that there are neural signals that can be measured non-invasively which are related to the detection and recog-

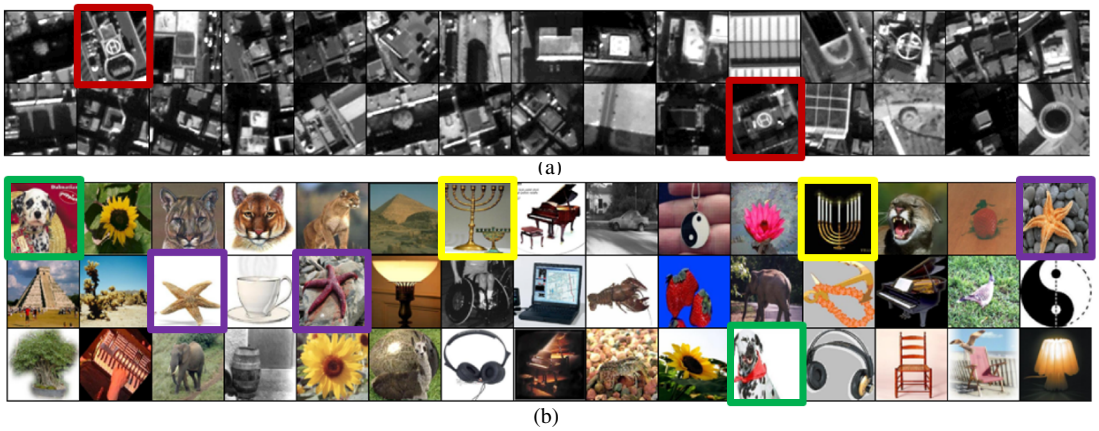


**Figure 3: Example images shown to subjects with target objects highlighted with a color bounding box. a) Satellite imagery with target "*helipad*"; b) Caltech101 images with targets "*dalmatian*", "*starfish*" and "*menorah*".**

nition of rapidly shown images [33, 17]. A very robust signal measurable from the EEG is the P300. It reflects a perceptual "orienting response" or shift of attention which can be driven by the content of the sensory input stream. Below we briefly describe the method we use to map the EEG into an "interest score" to be used for annotating the imagery.

Given the RSVP paradigm for presenting a rapid sequence images to the subject, we simultaneously record EEG, using 64 scalp electrodes (Figure 4), and map the activity to an "interest score" for each image. The interest score is meant to reflect how much of a user's attention was directed toward an image. From a neuroscience perspective it can be seen as the single trial correlate of the P300-related orienting response. The algorithm we use to decode the EEG and ultimately map it to an interest score has been described previously [8]. Briefly, our approach begins with the linear model

$$z_t = \sum\nolimits_i \alpha_i s_{it} \tag{1}$$

where $s_{it}$ represents the electrical potential measured at time $t$ for electrode $i$ on the scalp surface, while $\alpha_i$ represents the spatial weights which will be learned based on a set of training data. The goal is to combine voltages in the electrodes linearly such that the sum $z_t$ is maximally different between two conditions. The two conditions are "target of interest" vs "distracter". We also assume that this maximally discriminant activity is not constant but changes its spatial distribution within the second that follows the presentation of an image, thus we assume a stationarity time $T$ of approximately $100ms$. As a result, we find distinct optimal weight vectors, $\alpha_{ki}$ for each $100ms$ window following the presentation of the image (index $k$ labels the time window):

$$z_{kt} = \sum\nolimits_i \alpha_{ki} s_{it}\,, \quad t = (k-1)T \cdots kT \tag{2}$$

These different $z_{kt}$ are then combined in an average over time to provide the optimal discriminant activity over the entire second of data, with the result being our "interest score" $e$ for the image.

$$e = \sum\nolimits_t \sum\nolimits_k v_k z_{tk}. \tag{3}$$

For on-line implementation purposes we use the method of Fisher Linear Discriminants to train coefficients $\alpha_{ik}$ within each 100ms time window. The coefficients $v_k$ are learned using penalized logistic regression after all exemplars have been observed. Because of the two step process of first combining activity in space, and then again in time, the above EEG decoding method is termed Hierarchical Discriminant Component Analysis [26]. One advantage of such two-stage hierarchical modeling is the significant reduction of the number of model parameters that need to be learned - from about $10^5$ to $10^4$ (a 10 fold reduction). Colored scalp maps indicating spatial distribution of recorded cortical signal with different time interval are shown in Figure 4. It is important to confirm a strong correlate with the P300 attention orienting neural response, suggested by the neurological studies. Detectors built based on such single trial spatio-temporal EEG signal analysis have shown very promising results in various tasks such as people detection and image triage [8]. We will discuss the effectiveness of such a detector in detecting diverse objects such as those in Caltech 101 database [6] in Section 5.1.

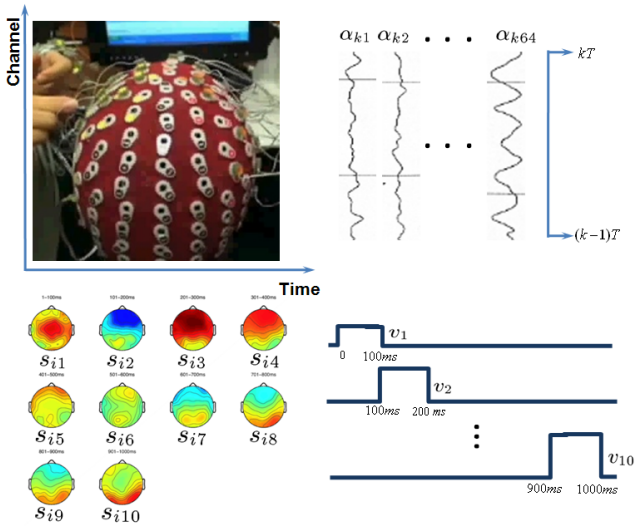


**Figure 4: Demonstration of the EEG-based generic interest detector. The scalp surface shown is monitored with 64 electrodes. The bottom left color scalp maps show the spatial distribution of the recorded cortical signal at different time intervals. The right part shows the signal decoding procedure by Hierarchical Discriminant Component Analysis.**

# 4. VISUAL PATTERN MINING WITH NOISY EEG LABELS

Assume that the generic interest detector outputs the EEG score $\mathbf{e} = \{e_1, e_2, \cdots, e_n\}$ from a RSVP sequence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ shown to the subject [1]. Usually, the top ranked images based on scores $\mathbf{e}$ do not match the desired OOI due to the noisy nature of EEG signals in practice, as shown in Figure 6, 7, 8, and 12 (a). Previous study shows that the existing semi-supervised methods cannot handle cases with extremely noisy labels [36]. In order to refine the noisy EEG scores, our method first extracts the salient image pattern and recover the visual consistency among the top ranked images. In other words, an improved interest measurement $\mathbf{f}$ is estimated using an image based representation and initial EEG scores as $\{\mathcal{X}, \mathbf{e}\} \rightarrow \mathbf{f}$. We formulate the following process of EEG label refinement and visual pattern mining.

1. Convert the image representation to a visual similarity graph $\mathcal{X} \rightarrow \mathcal{G} = \{V, E, W\}$, where vertices $V$ are the image samples $\mathcal{X}$ and the edges $E$ with weights $W$ measure the pairwise similarity of images.

2. Transfer the interest scores to pseudo EEG labels $\mathbf{e} = \{e_1, e_2, \cdots, e_n\} \rightarrow \mathbf{y} = \{y_1, y_2, \cdots, y_n\}$. In other words, a binarization function $g(\cdot)$ is applied to convert EEG scores to EEG labels as $\mathbf{y} = g(\mathbf{e})$, where $y_i \in \{1, 0\}$ and $y_i = 1$ for $e_i > \epsilon$, otherwise $y_i = 0$. The value $\epsilon$ is called interest level for discretizing the EEG scores.[2]

3. Apply the bivariate regularization framework to define the following risk function

$$E_\gamma(\mathbf{f}, \mathbf{y}) = \mathcal{Q}(\mathbf{f}, \mathbf{y}) + \gamma \mathcal{V}_\mathcal{G}(\mathbf{f}) \tag{4}$$

   which imposes the tradeoff between the smoothness measurement $\mathcal{V}_\mathcal{G}(\mathbf{f})$ of function $\mathbf{f}$ and empirical error $\mathcal{Q}(\mathbf{f}, \mathbf{y})$. Specifically, the function smoothness is evaluated over the undirected graph $\mathcal{G}$.

4. Alternatively minimize the above risk function with respect to $\mathbf{f}$ and $\mathbf{y}$ to finally achieve the optimal $\mathbf{f}^*$

$$\mathbf{f}^* = \arg\min_{\mathbf{f}, \mathbf{y}} E_\gamma(\mathbf{f}, \mathbf{y}) \tag{5}$$

In the following discussion, we follow the above procedure to detail our method for salient visual pattern mining with noisy EEG labels.

## 4.1 Image Features and Graph Construction

For the image feature extraction, we applied the widely used Bag-of-Visual-Words ($BoW$) derived from local key points, which has been shown to be effective in many applications of object and scene classification. In particular, we use the difference of Gaussian (DOG) and Harris-Affine as key point detector and SIFT as descriptor [19, 21]. To weigh the importance of a visual word to an image, a soft-assignment strategy for computing the frequency of visual words is adopted [13]. With a constructed visual vocabulary (size is 5000), the sparse representation of $BoW$ features for each image is extracted. The $\chi^2$ distance is often used for the calculation of dissimilarity $\mathbf{A_{ij}}$ between histograms of $BoW$ as:

$$A_{ij} = \sum_k \frac{(\mathbf{x}_{i,k} - \mathbf{x}_{j,k})^2}{\mathbf{x}_{i,k} + \mathbf{x}_{j,k}} \tag{6}$$

where $\mathbf{x}_{i,k}$ is the $k$th element of feature vector $\mathbf{x}_i$. We have also used global features (texture and shape feature) and Euclidean distance for part of the experiments, such as satellite images. Such features have been shown to be efficacious in previous work [34]. Starting from the distance matrix $\mathbf{A}$, the graph construction $\mathcal{X} \rightarrow \mathcal{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{W}\}$ is addressed in two steps, graph sparsification and edge weighting. Sparsity is important to ensure that a graph based algorithms remain efficient and robust to noise. The most common algorithm for recovering a sparse subgraph is the $K$ nearest neighbors algorithm ($K$NN), where each node greedily connects with its $K$ neighbors with the minimal distance. However, the $K$NN method typically produces asymmetric and irregular graphs, where the connectivity is uneven over different parts of the graph. This situation generates unreliable learning results if $\mathcal{X}$ contains very imbalanced class ratios, which occur very often in realistic image annotation settings, such as the 1.76% prior of target images in our Caltech 101 experiments. Recent investigation shows that $b$-matched graph is superior to $k$NN graph in terms of stability and accuracy for semi-supervised learning approaches [12]. Though the comparison between $k$NN and $b$-matched graph is not provided in this article due to space limitation, $K$NN graphs poorly performed in our experiments because the OOI is very infrequent in the tested image database.

With the vertex sparsified subgraph, the edge weights $W_{ij}$ are estimated by applying heat kernel function on the $\chi^2$ distance $A_{ij}$ as: $W_{ij} = \exp(-\frac{A_{ij}}{2\sigma^2})$. Realize the samples $\mathcal{X}$ might be draw unevenly, here we re-weight the similarity measure using an adaptive kernel size $\sigma$ as suggested in [10].

## 4.2 Graph based Visual Pattern Mining

Given the constructed $b$-matched graph with edge weight $\mathbf{W}$, the node degree matrix $\mathbf{D} = diag([d_1, \cdots, d_n])$ is defined as $d_i = \sum_{j=1}^n W_{ij}$. The normalized graph Laplacian is computed as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. Starting from the regularization framework in Equation 4, we formulate the following risk function:

$$E_\gamma(\mathbf{f}, \mathbf{y}) = \|\mathbf{f} - \mathbf{y}\|^2 + \gamma \|\mathbf{f}\|_\mathcal{G}^2 \tag{7}$$

where $\|\mathbf{f} - \mathbf{y}\|^2$ is the empirical square loss with pseudo EEG label $\mathbf{y} = g(\mathbf{e})$. The semi-inner product $\|\mathbf{f}\|_\mathcal{G}$ measures the function smoothness over the graph $\mathcal{G}$, which reflects the visual consistency.

$$\|\mathbf{f}\|_\mathcal{G}^2 = < \mathbf{f}, \mathbf{f} > = \mathbf{f}^\top \mathbf{L}\, \mathbf{f} \tag{8}$$

The above risk function is different from the existing regularization frameworks. For example, compared with regularization based kernel function learning in [20], and graph regularization for semi-supervised learning in [37], our problem is formed in a bivariate risk function and the empirical risk is estimated using unreliable pseudo EEG labels. Motivated by the alternating optimization approach, such as the one used in [35], we derive partial differentials with respect to $\mathbf{y}$ and $\mathbf{f}$, respectively, and iteratively update the function values to refine EEG labels.

[1] For an RSVP image sequence, the decoded EEG score vector $\mathbf{e} = \{e_1, e_2, \cdots, e_n\}$ is usually normalized as $e_i \in [0, 1], i = 1, \cdots, n$.

[2] In practice, the value of $\epsilon$ is set dynamically to achieve a fixed-number $l$ of EEG positive labels, i.e. $\sum_i y_i = l$.

**Algorithm 1 : EEG label refinement and visual pattern mining**

*Input*: initial EEG scores $\mathbf{e}$;
Graph $\mathcal{G}$ and normalized graph Laplacian $\mathbf{L}$;
Constant matrix: $\mathbf{p}_\gamma = (\mathbf{I} + \gamma\mathbf{L})^{-1}$.
*Initialization*:
The number of pseudo EEG labels $l$;
Iteration number $M = l/2$;
Initialize function $\mathbf{f}$ with EEG scores $\mathbf{f}^0 = \mathbf{e}$.
*Loop*: $\tau = 0, \cdots, M$
Convert EEG scores to labels:
$\mathbf{y}^\tau = g(\mathbf{f}^\tau)$, satisfying $\sum_i y_i^\tau = l$;
Compute gradient:
$\nabla_{\mathbf{y}^\tau} E_\gamma = \left\{ \|\mathbf{p}_\gamma - \mathbf{I}\|^2 + \gamma \mathbf{p}_\gamma^\top \mathbf{L} \mathbf{p}_\gamma \right\} \mathbf{y}^\tau$;
Update EEG label with truncated gradient:
$\mathbf{y}^{\tau+1} = \mathbf{y}^\tau - T(\nabla_{\mathbf{y}^\tau} E_\gamma)$
Recalculate interest level function:
$\mathbf{f}^{\tau+1} = \mathbf{p}_\gamma \mathbf{y}^{\tau+1}$;
*Output*: the refined EEG scores $\mathbf{f}^*$.

Since $\mathbf{f} \in \mathbb{R}^n$ is a continuous valued function, the optimal one can be derived by zeroing the partial differential $\nabla_{\mathbf{f}} E_\gamma$:

$$\begin{aligned} \frac{\partial E_\gamma}{\partial \mathbf{f}} &= \frac{\partial \mathcal{Q}(\mathbf{f}, \mathbf{y})}{\partial \mathbf{f}} + \gamma \frac{\partial \mathcal{V}(\mathbf{f})}{\partial \mathbf{f}} = 2(\mathbf{f} - \mathbf{y}) + 2\gamma \mathbf{L}\mathbf{f} = 0 \\ \Rightarrow \quad \mathbf{f}^* &= (\mathbf{I} + \gamma\mathbf{L})^{-1}\mathbf{y} = \mathbf{p}_\gamma \mathbf{y} \end{aligned} \tag{9}$$

where $\mathbf{p}_\gamma = (\mathbf{I} + \gamma\mathbf{L})^{-1}$ is a constant matrix in $\mathbb{R}^{n \times n}$ and $\mathbf{I}$ is identity matrix. Replace function $\mathbf{f}$ in Equation 7 by the optimal one $\mathbf{f}^*$ and rewrite the risk function as:

$$E_\gamma(\mathbf{y}) = \|\mathbf{p}_\gamma - \mathbf{I}\|^2 \|\mathbf{y}\|^2 + \gamma \mathbf{y}^\top \mathbf{p}_\gamma^\top \mathbf{L} \mathbf{p}_\gamma \mathbf{y} \tag{10}$$

Derive the partial differential $\nabla_{\mathbf{y}} E_\gamma$ and ignore the constant coefficient:

$$\frac{\partial E_\gamma}{\partial \mathbf{y}} \propto \left\{ \|\mathbf{p}_\gamma - \mathbf{I}\|^2 + \gamma \mathbf{p}_\gamma^\top \mathbf{L} \mathbf{p}_\gamma \right\} \mathbf{y} \tag{11}$$

Note that $\mathbf{y} \in \mathbb{B}^n$ is a binary vector representing class labels. The conventional approaches, such as zeroing $\nabla_{\mathbf{y}} E_\gamma$ or standard stochastic gradient is not appropriate for minimizing $E_\gamma$ with respect to the binary-valued variable $\mathbf{y}$. Here, we truncate the gradient $\nabla_{\mathbf{y}} E_\gamma$ and discretize it to $T(\nabla_{\mathbf{y}} E_\gamma)$:

$$T(\nabla_{y_i} E_\gamma) = \begin{cases} 1 & : \quad \nabla_{y_i} E_\gamma = \max(\nabla_{\mathbf{y}_l} E_\gamma) \\ -1 & : \quad \nabla_{y_i} E_\gamma = \min(\nabla_{\mathbf{y}_u} E_\gamma) \\ 0 & : \quad otherwise \end{cases} \tag{12}$$

where $\mathbf{y}_l, \mathbf{y}_u$ are the labeled and unlabeled parts of the label variable $\mathbf{y}$. Then the variable $\mathbf{y}$ can be updated with this truncated stochastic gradient $T(\nabla_{\mathbf{y}} E_\gamma)$:

$$\mathbf{y} \leftarrow \mathbf{y} - T(\nabla_{\mathbf{y}} E_\gamma) \tag{13}$$

Intuitively, the above updating by truncated gradient descent will remove one unreliable EEG label, and meanwhile choose the most suitable one from the remaining data as a new EEG label. Through iteratively repeating this truncated gradient descent updating, the EEG label set will be gradually refined to derive visually consistent visual pattern from the top ranked image list. Note that the gradient truncation approach is different with the method developed in [18], where the truncated gradient is applied to induce sparsity in the continuous-valued weights for online learning. We summarize the proposed method for EEG label refinement and visual pattern mining in algorithm chart 1.


**Figure 5: The summary of the experimental results on Caltech 101 dataset. The performance is evaluated in terms of average precision (AP). A total of four subjects and three OOI are tested (12 trials of RSVP presentations). The APs of random sequence, EEG detector and BCI-VPM refinement are recorded. The yellow color table cells highlight the significant improved trials (8 out of 12).**

## 5. EXPERIMENTS

We tested the developed BCI-VPM annotation system on the image data from various domains, including satellite imagery (DigiGlobe images) and Internet image collections (i.e. Caltech 101) to show the scalability and generalization. The detailed experimental setting and performance evaluation are reported below. The experimental scenario is that a user is instructed to look for a certain OOI in each presentation of an RSVP image sequence. The BCI interest detector generates probability based EEG confidence scores, which measure the relevance of the presented image to the instructed OOI. The estimated EEG score are then fed into the subsystem of visual pattern mining for refinement and propagation. During these tests, EEG data was recorded at 2048 Hz using a 64-electrode EEG recording system (Biosemi, BrainProduct, Germany) in a standard (10-20) setting.

### 5.1 Caltech 101 Object Annotation

**Data filtering:** Caltech 101 is a very challenging dataset for EEG decoding because it contains fairly common and diverse object categories with large intra class variations. Moreover, images greatly vary in both resolutions and scales. This can significantly affect the user's detection performance during the EEG signal decoding. To design a practical set of intial EEG experiments, we first filter the object categories by selecting 62 categories that provide 3798 images that have similar scales and resolutions. When displayed during the RSVP, these images are re-scaled to a size of $240 \times 240$ to achieve the desired uniformity in view.

**Experimental scenario:** As it is impractical to require a user to perceive 62 OOI simultaneously, we also narrowed

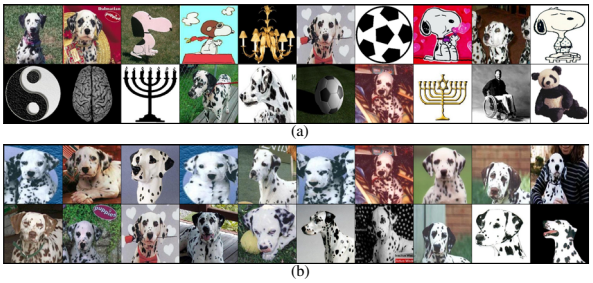

**Figure 6: The experimental results (top 20 images) of the trial from Subject A on Caltech 101 RSVP with OOI as *Dalmatian*. a) ranking by interest scores from EEG detector; b) ranking by scores after label refinement.**

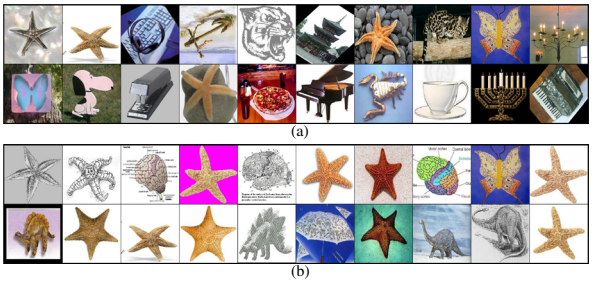

**Figure 7: The experimental results (top 20 images) of the trial from Subject B on Caltech 101 sequence with OOI as *Starfish*. a) ranking by interest scores from EEG detector; b) ranking by scores after label refinement.**

th choice of target categories that a user has to detect to one specific target each time. Specifically, users were instructed perceive *Starfish*, *Dalmatian*, and *Chandelier/Menorah* [3] for each pass of RSVP (with the target order being varied between subjects). Notice that this simplification still did not reduce much of the challenge due to the diverse object categories and sparse targets. For example, "*Starfish*" and "*Dalmatian*" only account for 2.26% and 1.76% of the data set, respectively. For the BCI interest detector training, we use two popular objects, *Soccer Ball* or *Baseball Gloves* as OOI to train the EEG-based detector. Most subjects are familiar with such objects without the need of special instruction and thus they serve as adequate common patterns that can catch user's attention. The Caltech 256 database was used to select training images to differentiate between the training and testing images.

**Image database down sampling:** Furthermore, in order to show the scalability of the proposed method, we randomly partition the 62-object database into two parts $\mathcal{X} = \{\mathcal{X}_s, \mathcal{X}_u\}$. The subset $\mathcal{X}_s$ containing 1000 images is randomly ordered to form a RSVP image sequence shown to all test subjects. The RSVP image sequences were shown in 10 blocks of 100 images each, with images shown at 6Hz within each block. The other subset $\mathcal{X}_u$ containing a total of 2798 images is treated as EEG-unlabeled samples in later label refinement stage by setting their EEG scores as $\mathbf{e}_u = \vec{\mathbf{0}}$. This strategy shows that with only partial images

[3]The experiments were initially designed to annotate the object class *Chandelier*. Realizing the ambiguity between *Chandelier* and *Menorah* due to visual similarity, we decided to treat the samples from these two object categories as the same class in these experiments.

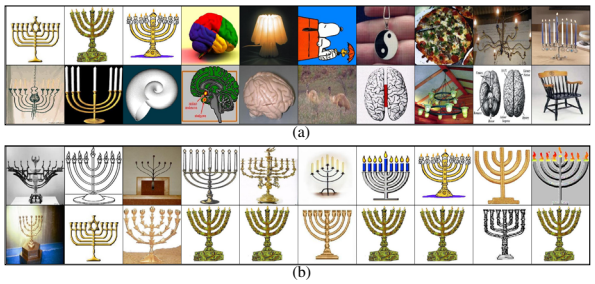

**Figure 8: The experimental results (top 20 images) of the trial from Subject C on Caltech 101 sequence with OOI as *Chandelier/Menorah*. a) ranking by interest scores from EEG detector; b) ranking by scores after label refinement.**

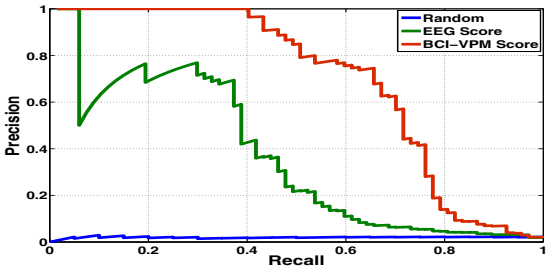


**Figure 9: The performance evaluation on Caltech101 image sequence by Precision-Recall (PR) curve (the trial of Subject A annotating "*Dalmatian*").**

processed by the BCI detector, the proposed BCI-VPM system can extend the usage of the BCI annotation to the images that are not processed by BCI interest detector. This merit is extremely critical because it provides the capability and scalability to annotating a large collection of images.

**Results:** Four subjects are drawn from undergraduate and graduate students, staff and faculty that were not digital media analysts, but were familiar with EEG work. These subjects participated in the experiments and were instructed to identify three object classes from RSVP presentations. A total of 12 trials of RSVP presentations are evaluated in terms of average precision (AP), as shown in Figure 5. AP is a performance metric commonly used in information retrieval [32]. It approximates the area under the precision-recall curve. The experiments show promising results. For example, the BCI detector achieved 33.73% and BCI-VPM label refinement further improved to 69.1% for subject A annotating *"Dalmatian"*. Among the 12 trials of tests, 8 trials achieved significant performance improvement. Even for some tough cases, where the BCI detector only obtained less than 10% AP, the label refinement process still significantly bootstraps the annotation accuracy. Figure 6, 7, and 8 showed the top 20 images from the BCI detector and BCI-VPM label refinement from three different test subjects, where significant precision gain can be observed. In Figure 9, we analyze the precision-recall(PR) curves of one of the successful case (Subject A annotating "*Dalmatian*"), which further confirms the efficiency of the proposed BCI-VPM annotation system.

However, there are some cases, where possibly due to users' misunderstanding of the OOI or some uncontrolled distractions, the BCI detector generated very poor performance, typically less than 7% AP. In those cases, the top

| Subject | Method | AP-30 | AP-60 | AP-100 | AP-ALL |
|---|---|---|---|---|---|
| **A** | EEG score | 19.76 | 15.83 | 14.9 | 30.97 |
| | BCI-VPM score | **50.19** | **32.65** | **25.46** | **37.89** |
| **B** | EEG score | 8.76 | 9.79 | 9.56 | 23.71 |
| | BCI-VPM score | **87.31** | **63.29** | **46.07** | **57.41** |
| **C** | EEG score | 12.70 | 19.58 | 16.54 | 29.62 |
| | BCI-VPM score | **90.82** | **63.30** | **41.21** | **53.66** |
| **D** | EEG score | 10.68 | 11.87 | 11.75 | 24.45 |
| | BCI-VPM score | **91.87** | **60.62** | **40.24** | **52.70** |

**Table 1: The performance comparison of annotation performance of EEG interest detector and the BCI-VPM refined EEG score in terms of average precision of top 30, 60, 100 ranked images and entire satellite image dataset (the number of pseudo EEG labels $l = 30$).**

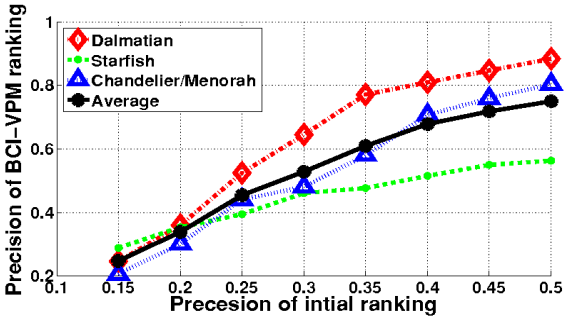


**Figure 10: Simulated evaluation of the dependency of the BCI-VPM re-ranking performance (in terms of top-20 precision) on the performance of the initial EEG detection. Individual curves for different classes (*Dalmatian*, *Chandelier/Menorah*, *Starfish*) and the average results across three categories are shown.**

ranked images do not contain a majority consistent pattern. Therefore, the subsequent label refinement process is unable to extract the salient visual patterns.

To analyze the sensitivity of the combined BCI-VPM accuracy to the quality of the front end BCI detection precision, we further evaluate the effectiveness of the BCI-VPM system with a varying number of true positive samples contained in the top images (e.g., 20) of the initial EEG-based ranking. The positive images are randomly drawn from the target category and the negative images are randomly drawn from the database. Average performances of 200 random runs per category and the average results among the three targets are shown in Figure 10. The results confirm that the combined BCI-VPM approach can effectively improve the precision by using the EEG detector alone. They also confirm the monotonic relationship between the final accuracy and the initial EEG detection accuracy. For example, the average top-20 precision is improved from 20% to 35% and from 30% to about 55%. Such performance curves can be used to determine the required accuracy for the initial EEG component given a target detection performance for the final combined system. The results can also be used to roughly measure the visual pattern complexity and the difficulty in re-ranking the three targets - *Starfish* is more difficult than *Chandelier/Menorah*, which is in turn more difficult than *Dalmatian*. Such ordering matches the intuitive expectation of the relative complexities of the object classes.

## 5.2 Target Annotation in Satellite Imagery

The other type of RSVP image sequence shown to test subjects consists of blocks of chipped images taken from satellite imagery with each chip potentially containing a target, as shown in Figure 3 (a). Among the 1051 samples in the RSVP sequence of satellite imagery, 105 images have a target "*helipad*" centered. This sequence is displayed to the subjects with a speed of 10 Hz, i.e. 10 images per second. A total of 4 subjects were tested with single trial of presentation of RSVP. The performance is evaluated in terms of average precision of the top 30, 60, and 100 and entire dataset of the BCI interest detector and the final refined results, as shown in Table 1. Compared with the Caltech 101 experiments, Table 1 shows much more consistent performance gain for all trials. The reason lies in that the targets of "*helipad*" have salient visual clue of "H" symbol, which easily attracts users' attention. In addition, the non-target image blocks are very noisy and mostly unmeaningful, which reduces level of distraction.

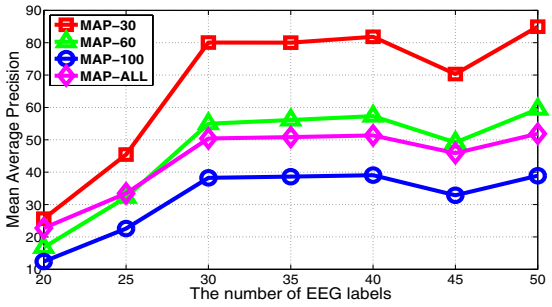


**Figure 11: The Mean Average Precision of top 30, 60, 100, and entire satellite image set using different numbers of EEG scores as initial labels.**

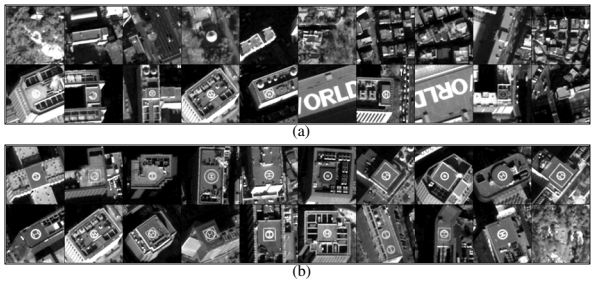


**Figure 12: The experimental results of Subject C on "*helipad*" target RSVP, showing the top 20 ranked images . a) ranking by original EEG scores; b) ranking by the BCI-VPM refined interest score.**

Since the value of $l$ is applied for truncating the EEG scores to create initial binary labels, it is necessary to evaluate the performance using different $l$. We vary the value of $l$ from 20 to 50 and evaluate the performance in terms of mean average precision (MAP) (averaging among four subjects). As shown in Figure 11, a fairly stable performance range is with the value of $l$ in $[30, 50]$. As an illustration, Figure 12 shows the test results from Subject C. The sub-figure (a) gives the top 20 image blocks by ranking the original EEG interest scores $\mathbf{e}$ and the sub-figure (b) shows the top 20 images from the ranking of the refined interest score $\mathbf{f}$.

# 6. COMPARISON WITH PRIOR WORKS AND UNIQUE CONTRIBUTION

Despite the growing interest in BCI, few works can be

| *System* | **C3Vision** [8] | **HAC** [31] | **HAC-CV**[15] | **BCI-VPM** |
|---|---|---|---|---|
| *System Structure* | pure BCI | pure BCI | hybrid BCI+CV | hybrid BCI+CV |
| *Neural Signal Trials* | single trial<br>single subject | single/multiple trials<br>single/multiple subjects | multiple trials<br>single/multiple subjects | single trial<br>single subject |
| *Object Class* | people vs. background | face vs. animal<br>animal vs. inanimate | face, animal,<br>and inanimate | general object class |
| *Target Frequency* | 2% | 25% | 50% | $\sim 2\%$ |
| *Manual Labels* | No | Yes | Yes | No |
| *Image Presentation Speed* | $5-10$ HZ | $1-2$ HZ | $1-2$ HZ | $5-10$ HZ |
| *Learning Method* | unsupervised | supervised | supervised | unsupervised |

**Table 2: Comparison of the existing BCI-based image analysis system and our proposed BCI-VPM image annotation system.**

found in using BCI for image annotation and search. We summarize the ideas of the prior works and point out the unique contributions of the work presented in this paper.

The pioneer system (called Cortically-Couple Computer Vision, C3Vision) using EEG-based neural measurement in image target detection was reported in [8]. Its RSVP visual presentation paradigm and spatio-temporal discriminant analysis approaches are extended to develop the interest detector in this paper. Compared to the current work, it focused on less diverse object classes (e.g., people vs. background) and had not been combined with the pattern discovery subsystem for label refinement and propagation.

Recently, a supervised learning approach (called Human Aided Computing, HAC) was proposed in [31] to develop an EEG-based classifier for recognition of distinct objects in images, such as face vs. no-face, or animals vs. inanimate objects. The proposed method showed performance improvement when neural response measurements from multiple trials (i.e., same images presented multiple times) and/or multiple subjects were combined. However, the technique is limited due to the requirement of predefined object classes and ground truth labels for training the object detectors. In contrast, our work focuses on robust detection using only single trial EEG signals and scalability to detection of arbitrary object classes. The only training session is done offline only once per subject using an object class that is independent of the test targets.

Computer vision component was added to the HAC method (HAC-CV) in [15] to fuse the EEG signals and the image features in the same target classifier. Multiple trials were used and improved performance was reported compared to detection using EEG signals alone. Again, the method is restricted as target classes are predefined, labeled, and trained in a supervised fashion.

Other works have also explored the use of fMRI imaging to decode brain state [22, 16, 23]. Such approaches enjoy a higher spatial resolution at the cost of lower temporal throughput. In our work, we focus on the system utilizing a non-invasive continuous recording BCI framework based on EEG. Table 2 lists the comparison between our proposed method and the prior works discussed above. The key features and unique contributions of the proposed system include:

. **Generality**: No predefined target classes and annotations are needed.

. **Robustness**: The combination of BCI and pattern discovery results in greatly improved accuracy in generic object detection.

. **High-throughput**: Only a small subset of images need to be viewed by the subject via a single trial setting. The results are automatically propagated over to the rest of images in a large collection.

In addition, our proposed system can handle rare target classes with a prior as low as 1.8% (as demonstrated in Section 5.1) at a high speed (5-10 images per second).

# 7. DISCUSSIONS AND CONCLUSIONS

In this paper we propose joint exploration of neurotechnology for brain state decoding and media content analytics for improving the performance of image retrieval systems. The brain state decoding subsystem utilizes advanced spatio-temporal analysis of neural response signals measured by EEG in a single trial setting. The content analytics component implements a graph-based diffusion process that is capable of handling rare targets, small label size, and noisy conditions. The former has unique power in recognizing generic target classes, while the latter is suitable for high-throughput processing. The proposed system explores the synergy of the two and has shown promising performance in detecting generic target classes in a high throughput fashion.

Several aspects of the system merit further investigation. First, in the current setting subjects are instructed to look for certain object classes during the image viewing session. It will be interesting to relax this and allow subjects to "lock in" on any object class fitting his/her interest on the fly without instruction. This will result in greater ambiguity in the user's understanding and perception of targets and less consistent target-specific neural responses. Second, the graph-based pattern discovery output can be used to continuously assess the quality of EEG interest detector and provide a closed-loop feedback mechanism to incrementally update the detector. The re-ranked images could also be used to adjust the order of image presentation in the RSVP viewing stream in order to improve sensitivity/specificity of the neural response signals.

# 8. ACKNOWLEDGEMENT

This work was funded by Government contract No. NBCHC080029. Aerial images are provided by DigiGlobe.

# 9. REFERENCES

[1] N. Bigdely-Shamlo, A. Vankov, R. Ramirez, and S. Makeig. Brain Activity-Based Image Classification From Rapid Serial Visual Presentation. *IEEE Trans. on NSRE*, 16(5):432–441, 2008.

[2] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig. Brain activity-based image classification from rapid serial visual presentation. *IEEE Trans. on NSRE*, 16(5):432–441, Oct. 2008.

[3] J. Donoghue. Bridging the brain to the world: a perspective on neural interface systems. *Neuron*, 60(3):511–521, 2008.

[4] M. Dyrholm, C. Christoforou, and L. Parra. Bilinear discriminant component analysis. *JMLR*, 8:1097–1111, 2007.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007.

[7] R. Fergus and P. Perona. A Visual Category Filter for Google Images. In *Proc. ECCV*, 2004.

[8] A. Gerson, L. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Trans. on NSRE*, 14(2):174–179, 2006.

[9] M. Gladwell. *Blink: The power of thinking without thinking*. Little, brown and company: Time warner book group, New York, 2005.

[10] M. Hein and M. Maier. Manifold denoising. *Proc. NIPS*, 19, 2006.

[11] T. Huang, C. Dagli, S. Rajaram, E. Chang, M. Mandel, G. Poliner, and D. Ellis. Active Learning for Interactive Multimedia Retrieval. *Proc. of the IEEE*, 96(4):648–667, 2008.

[12] T. Jebara, J. Wang, and S.-F. Chang. Graph construction and $b$-matching for semi-supervised learning. In *Proc. ICML*, 2009.

[13] Y. Jiang, C. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proc. of CIVR*, pages 494–501, 2007.

[14] Y. Jing and S. Baluja. VisualRank: Applying PageRank to Large-Scale Image Search. *IEEE Trans. on PAMI*, 12, 2008.

[15] A. Kapoor, P. Shenoy, and D. Tan. Combining brain computer interfaces with vision for object categorization. In *Proc. CVPR*, 2008.

[16] K. Kay, T. Naselaris, R. Prenger, and J. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[17] C. Keysers, D. Xiao, P. Foldiak, and D. Perrett. The speed of sight. *Journal of Cognitive Neuroscience*, 13(1):90–101, 2001.

[18] J. Langford, L. Li, and T. Zhang. Sparse Online Learning via Truncated Gradient. *JMLR*, 10:777–801, 2009.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[20] C. Micchelli and M. Pontil. Learning the kernel function via regularization. *JMLR*, 6(2):1099–1125, 2006.

[21] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[22] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175, 2004.

[23] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. Tanabe, N. Sadato, and Y. Kamitani. Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron*, 60(5):915–929, 2008.

[24] A. Oliva. Gist of the scene. In *Encyclopedia of Neurobiology of Attention*, pages 251–256, San Diego, CA, 2005. Elsevier.

[25] L. Parra, C. Christoforou, A. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. Philiastides, and P. Sajda. Spatiotemporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks. *IEEE Signal Processing Magazine*, 25(1):95–115, January 2008.

[26] L. Parra, C. Christoforou, A. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. Philiastides, and P. Sajda. Spatiotemporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks. *IEEE Signal Processing Magazine*, 25(1):95–115, January 2008.

[27] P. Poolman, R. Frank, P. Luu, S. Pederson, and D. Tucker. A single-trial analytic framework for EEG analysis and its application to target detection and classification. *NeuroImage*, 42(2):787–798, 2008.

[28] M. Potter and E. Levy. Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1):10, 1969.

[29] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. on CSVT*, 8(5):644–655, 1998.

[30] M. Sanderson and P. Clough. cross-language image retrieval track. http://imageclef.org/.

[31] P. Shenoy and D. Tan. Human-Aided Computing: Utilizing Implicit Human Processing to Classify Images. In *Proc. CHI*.

[32] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[33] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[34] J. Wang, S. F. Chang, X. Zhou, and S. T. C. Wong. Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels. In *Proc. CVPR*, 2008.

[35] J. Wang, T. Jebara, and S.-F. Chang. Graph transduction via alternating minimization. In *Proc. ICML*, 2008.

[36] J. Wang, Y.-G. Jiang, and S.-F. Chang. Label diagnosis through self tuning for web image search. In *Proc. CVPR*, 2009.

[37] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf. Ranking on data manifolds. In *Proc. NIPS*, 2004.