INVITED PAPER

In a Blink of an Eye and a Switch of a Transistor: Cortically Coupled Computer Vision

To identify "interesting" images, human observers view 10 images/sec, while electroencephalography (EEG) signals from the observers own brains are automatically decoded.

By Paul Sajda, *Senior Member IEEE*, Eric Pohlmeyer, Jun Wang, Lucas C. Parra, *Senior Member IEEE*, Christoforos Christoforou, Jacek Dmochowski, Barbara Hanna, *Member IEEE*, Claus Bahlmann, Maneesh Kumar Singh, *Member IEEE*, and Shih-Fu Chang, *Fellow IEEE*

ABSTRACT | Our society's information technology advancements have resulted in the increasingly problematic issue of information overload—i.e., we have more access to information than we can possibly process. This is nowhere more apparent than in the volume of imagery and video that we can access on a daily basis—for the general public, availability of YouTube video and Google Images, or for the image analysis professional tasked with searching security video or satellite reconnaissance. Which images to look at and how to ensure we see the images that are of most interest to us, begs the question of whether there are smart ways to triage this volume of imagery. Over the past decade, computer vision research has focused on the issue of ranking and indexing imagery. However, computer vision is limited in its ability to identify interesting imagery, particularly as "interesting" might be defined by an individual. In this paper we describe our efforts in developing brain-computer interfaces (BCIs) which synergistically integrate computer vision and human vision so as to construct a system for image triage. Our approach exploits machine learning for real-time decoding of brain signals which are recorded noninvasively via electroencephalography (EEG). The signals we decode are specific for events related to imagery attracting a user's attention. We describe two architectures we have developed for this type of cortically coupled computer vision and discuss potential applications and challenges for the future.

KEYWORDS | Brain-computer interface; computer vision; electroencephalography; image search; image triage

I. INTRODUCTION

Our visual systems are amazingly complex information processing machines. We can recognize objects at a glance, under varying poses, illuminations, and scales, and are able to rapidly learn and recognize new configurations of objects and exploit relevant context even in highly cluttered scenes. This visual information processing all happens with individual components which are extremely slow relative to state-of-the-art digital electronics—i.e., the frequency of a neuron's firing is measured in hertz whereas modern digital computers have transistors which switch at gigahertz speeds (a factor of 10^8 difference). Though there is some debate on whether the fundamental processing unit in the nervous system is the neuron or whether ensembles of neurons constitute the fundamental unit of processing, it is nonetheless widely believed that the human visual system

Manuscript received February 13, 2009; revised November 22, 2009.
Current version published March 5, 2010. The work was supported by DARPA under Contract NBCHC090029.

P. Sajda and **E. Pohlmeyer** are with the Department of Biomedical Engineering, Columbia University, New York, NY 10027 USA (e-mail: psajda@columbia.edu; ep2473@columbia.edu).

J. Wang and **S.-F. Chang** are with the Department of Electrical Engineering, Columbia University, NY 10027 USA (e-mail: jwang@ee.columbia.edu; sfchang@ee.columbia.edu).

L. C. Parra, **C. Christoforou**, and **J. Dmochowski** are with the Department of Biomedical Engineering, City College of New York, New York, NY 10031 USA (e-mail: parra@ccny.cuny.edu; cchristoforos@yahoo.com; dmochow@emt.inrs.ca).

B. Hanna is with Neuromatters, LLC, New York, NY 10038 USA (e-mail: bhanna@neuromatters.com).

C. Bahlmann and **M. K. Singh** are with Siemens Corporate Research, Princeton, NJ 08540 USA (e-mail: claus.bahlmann@siemens.com; maneesh,singh@siemens.com).

Digital Object Identifier: 10.1109/JPROC.2009.2038406

0018-9219/$26.00 ©2010 IEEE

is bestowed with its robust and general purpose processing capabilities not from the speed of its individual processing elements but from its massively parallel architecture—the brain has 10^{11} neurons and 10^{14} synapses of which the visual cortex is by far the largest area.

Since the early 1960s there have been substantial efforts directed at creating computer vision systems which possess the same information processing capabilities as the human visual system. These efforts have yielded some successes, though mostly for highly constrained problems. By far the biggest challenge has been to develop a machine capable of general purpose vision. A key property of the human visual system is its ability to learn and exploit invariances. As mentioned above, we can in most cases effortlessly recognize objects under extreme variations in scale, lighting, pose, and other variations in the object and world. Understanding how this invariance comes about and relating it to the physics of objects and projections of scenes onto our retinas (or a machine system's imager) is one of the most active areas of computer vision research. More recently the problem of invariance has been considered from a statistical perspective, with the idea that "natural scene statistics" may hold the key to how our visual systems learn and represent these invariances [1]. We are, however, many years, if not decades, away from realizing a computer vision system with general purpose visual processing capabilities.

A. The Image Triage Problem

Instead of focusing on how to build a computer vision system to emulate human vision, in this paper we consider how we might synergistically integrate computer and human vision systems to perform the task of *image triage*. Assume we start with a database of images $\mathcal{D}_0$ which is organized as an ordered set of N images, $\mathcal{D}_0 = \{I_1 \ldots I_N\}$, where N is very large. Also assume the state of the database can be characterized by a utility function, U which quantifies how "ordered" the database is with respect to the interest of the person, p, searching the data at a given time, t; $U(\mathcal{D}_0|p,t)$. For now we will assume that this utility function has large positive values when the database is ordered such that "interesting" images (given p and t) are at the front of the database. Conversely, if "interesting" images appear randomly in the database then $U \approx 0$.

The image triage problem can then be defined as finding a transformation $T(\cdot)$ which operates on $\mathcal{D}_0$, or more generally $\mathcal{D}_i$ where i indexes the database after applying the ith transform, to reorder the images so as to maximize U and minimize the cost of computing and applying the transformation $T(\cdot)$

$$\mathcal{D}_{i+1} \leftarrow T(\mathcal{D}_i|p,t) : \arg\max_{T(\cdot)}(U(T(\mathcal{D}_i)|p,t) - \lambda C(T(\mathcal{D}_i)) \tag{1}$$

where λ balances the cost of $T(\cdot)$ relative to its utility. One potential cost of computing and applying $T(\cdot)$ is time—i.e., if it takes very long to compute and apply $T(\cdot)$, then this will reduce the rate at which interesting images will be discovered and therefore reduces the overall utility of the triage.

In this paper we describe a basic set of principles we will use to construct a reordering transform which leverages the strengths of both computer (CV) and human vision (HV)—i.e., the transform will be a synergistic combination of $T_{\mathrm{CV}}(\cdot)$ and $T_{\mathrm{HV}}(\cdot)$.

B. Human Vision and "Gist"

In considering the human vision reordering transform, we will first take advantage of our ability to get the "gist" of a scene. That is, for a very brief presentation of an image, we are able to extract a tremendous amount of information which enables a general characterization of the image content. For example, if an image is flashed for 50 ms, we might be able to infer that the image contained a car, but perhaps not what model car it was. This is exactly the operational mode of the image triage problem. Given images in our database $\mathcal{D}_i$, we are using HV to obtain a general characterization of what is in the image. "Gist" processing by humans has been an active area of research [2] leading to several theories of the type of features we use to infer general scene characteristics. In our triage system we are less concerned with how we "get the gist," and more interested in whether a subject gets a particular gist within the context of a binary discrimination—i.e., if the flashed image contains a car or not, or more generally whether or not there is something "interesting" in the scene. The definition of the binary discrimination can be explicit, such as instructing the subject to look for a particular class of object or sets of objects, or implicit, such as a subject being interested in certain objects or characteristics of images, as one might be during casual browsing. The binary discrimination can also be dynamic and context dependent—i.e., depend on the previous images the subject has seen in the sequence $\mathcal{D}_i$.

To maximize the throughput of the triage and thus minimize the time it takes to apply $T(\cdot)$, we want a sequence of images having a presentation rate which is as rapid as possible while still enabling the subject to get the gist of images in the sequence. For this we use a presentation methodology termed "rapid serial visual presentation" or RSVP [3], [4]. RSVP has been well-studied in visual psychophysics and it has been shown that we can get the "gist" of what is in an image at RSVP rates of greater than 10 Hz [5]–[7]. The specifics of the RSVP paradigm we use will be outlined later in the paper.

Given RSVP presentations and the binary discrimination problem formulation, how do we detect the subject's decision—e.g., whether a given image in the sequence is of interest or not? One way is to have the subject behaviorally respond, perhaps by pressing a button or making an eye-movement. However an explicit response by subjects has several drawbacks. The first is that there is substantial trial-to-trial response time (RT) variability and at high RSVP rates this can result in errors in localizing the image to which the subject responded [8]. In addition, by dissociating the response from the decision one can potentially speed up the process and make it less taxing. Lastly, the subject may over-analyze the image, resulting in a behavioral decision threshold which is higher than one might want for the triage task. For all these reasons we monitor the subject's EEG during the RSVP and use machine learning to identify neural "components" reflective of target detection and attentional orienting events, which in turn can be used to infer the binary discrimination—i.e., the reordering transform $T(\cdot)_{\mathrm{HV}}$ will be based on EEG; $T(\cdot)_{\mathrm{EEG}}$. Our framework for decoding EEG using machine learning is described in Section II below. Several groups, including ours, have investigated image detection/classification based on EEG [8]–[13]. In contrast to this previous work, in this paper we focus on the integration of computer vision and EEG, specifically describing two architectures for the integration of the two triage transform systems, presenting results for each.

C. Organization of the Paper

The remainder of the paper is organized as follows. Section II will describe the neurological basis for the signals utilized to detect attentional shift and orienting, surrogates of the "that is interesting" response to a flashed image. We will also describe the signal processing and machine learning framework we utilize for decoding these signals and generating a probability-based "interest" score for each image. Section III describes two systems we have developed for coupling computer vision and decoding of cortical EEG signals for image triage. The first uses computer vision as a preprocessor to provide an initial reordering transform and then samples from this reordered $\mathcal{D}_i$ to generate sequences to pass through EEG-based reordering. This architecture is particularly appropriate when prior information about target/object class is known and can be incorporated into a CV model. The second method begins with the EEG-based reordering and samples from $\mathcal{D}_i$ based on the user browsing these samples. The result of the EEG-based browsing is used to reorder $\mathcal{D}_i$ and samples of the reordered database are used as exemplars in a semisupervised CV system for further reordering. This framework is most appropriate when the object of interest is unknown and examples can only be generated after the subject browses the database. We conclude the paper by discussing future technical development and ethical and human factors issues related to these types of BCI systems.

II. DECODING BRAIN STATE

There has been substantial interest, which has accelerated over the past decade, for decoding brain state. Efforts have ranged from decoding motor commands and intentions, to emotional state and cognitive workload. There has also been a variety of neural signals which have been targeted for decoding, ranging from spike trains collected via invasive recordings to hemodynamic changes measured noninvasively via fMRI [14], [15]. Our focus is on using EEG as a noninvasive measure to relate brain state to events correlated with the detection of "interesting" visual objects and images. What is the neural correlate of an "interesting" image? It is not clear that there is such a well-defined correlate. However, we do know from neuroimaging studies that there are neural signals that can be measured noninvasively which are related to the detection and recognition of rapidly shown images [5], [7], [8]. 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 [16]. Additional signals that may be indicative of a subject's attentional state are oscillatory activity often found during resting state (10 Hz oscillations known as "alpha" activity) as well as transient oscillations sometimes associated with perceptual processing (30 Hz and higher known as "gamma" activity). However, as of yet, none of these oscillatory signals have been identified in the RSVP paradigms.

A. Signal Detection via Spatio-Temporal Linear Projections

The approach we have taken for interpreting brain activity is to constrain the experimental paradigm such that we have to distinguish only among two possible brain states: $(+)$ positive examples in which the subject sees something of interest in an image, versus $(-)$ negative examples for which the image contains nothing of particular interest. The goal is not to deduce from the brain signal what the exact content is, or what the subject sees in the image. This would indeed be a difficult task given the limited spatial resolution of EEG. Instead, we aim to utilize the high temporal resolution of EEG to detect *when* an individual recognition event occurred. For individual images we aim to detect the brain signals elicited by positive examples, and distinguish them from the brain activity generated by negative example images. The task for the EEG analysis algorithm is therefore to classify the signal between two possible alternatives.

In our RSVP paradigm images are presented very rapidly with 5 to 10 images per second. To classify brain activity elicited by these images we analyze 1 second of data, recorded with multiple surface electrodes, following the presentation of an image. With 64 electrodes and approximately 100 time samples within this second, this amounts to a data vector of 6400 elements. In the specific case of image triage, we may have hundreds or thousands of images that are to be ignored and a very few, perhaps a dozen or two, which are assumed to attract the subject's attention. The goal is to identify a classification criterion in this large spatio-temporal data space using only a few known example images.

B. Hierarchical Discriminant Component Analysis

We begin by assuming that the discriminant activity, i.e., the activity that differs the most between positive and negative examples, is a deflection of the electrical potential from baseline (either positive or negative) over a number of electrodes. By averaging over electrodes with just the right coefficients (positive or negative with magnitudes corresponding to how discriminant each electrode is) we obtain a weighted average of the electrical potentials that will be used to differentiate positive from negative examples

$$y_t = \sum_i w_i x_{it}. \tag{2}$$

Here x_{it} represents the electrical potential measured at time t for electrode i on the scalp surface, while w_i represents the spatial weights which have to be chosen appropriately. The goal is to combine voltages linearly such that the sum y is maximally different between two conditions. This can be thought of as computing a neuronal current source y_t that differs most between times samples $t+$ following positive examples and the times $t-$ following negative examples, $y_{t+} > y_{t-}$.[1] There are a number of algorithms available to find some optimal coefficients w_i in such a binary linear classification problem, e.g., Fisher linear discriminants (FLDs), penalized logistic regression (PLR), or support vector machines (SVMs) [17].

In [8] we assume that these maximally discriminant current sources are not constant but change their spatial distribution within the second that follows the presentation of an image. Indeed, we assume a stationarity time T of approximately 100 ms. Therefore, we find distinct optimal weight vectors, w_{ki} for each 100 ms window following the presentation of the image (index k labels the time window)

$$y_{kt} = \sum_i w_{ki} x_{it}, \quad t = T, 2T, \ldots (k-1)T, kT. \tag{3}$$

These different current sources y_{kt} are then combined in an average over time to provide the optimal discriminant activity over the entire second of data

$$y = \sum_t \sum_k v_k y_{tk}. \tag{4}$$

For an efficient online implementation of this method we use FLD to train coefficients w_{ik} within each window of time, i.e., we seek w_{ik} such that $y_{kt+} > y_{kt-}$. The coefficients v_k are learned using logistic regression after the subject has viewed the entire training set such that $y_+ > y_-$. Because of the two step process of first combining activity in space, and then again in time, we have termed this algorithm "Hierarchical Discriminant Component Analysis" (HDCA).

Note that the first step does not average over time samples within each window. Instead, each time sample provides a separate exemplar that is used when training the FLD.[2] These multiple samples within a time window will correspond to a single exemplar image and are therefore not independent, yet, they do provide valuable information on the noise statistic: variations in the signal within the time window are assumed to reflect nondiscriminant "noise." In other words, we assume that spatial correlation in the high-frequency activity ($f > 1/T$) is shared by the low-frequency discriminant activity. In addition, by training the spatial weight separately for each window we assume that the discriminant activity is not correlated in time beyond the 100 ms time scale. Both these assumptions contribute crucially to our ability to combine thousands of dimensions optimally despite the small number of training images with known class labels [10].

An example of the activity extracted for one subject with this algorithm is shown in Fig. 1. The spatial distributions show the portion of the electrical potentials measured on the electrodes that correlates with the discriminant current sources.[3] For example, as shown, the activity measured in frontal areas 801–900 ms poststimulus presentation (shown as red in the scalp plot) strongly correlates with the classifier output y. Class-conditional histograms, computed via integrating the ten component activities, show the distribution of y_+ and y_- computed on unseen test data (fivefold cross validation). The receiver-operator characteristic (ROC) curve is computed from these histograms.

The HDCA algorithm is computationally very efficient and easy to implement in real time. It is thus the algorithm of choice for the current implementation of the C3Vision system. More recently we have developed new learning algorithms to find optimal linear weights (see Appendix). While these can yield better classification accuracy their computational cost makes them less-suitable for real-time implementations.

[1] Label "+" always indicates that the expression is evaluated with a signal x_{it} recorded following positive examples and "−" indicates the same for negative examples.

[2] For instance, we may have 50 training exemplars and 10 samples per window resulting in a possible 500 training samples for the classification algorithm that needs to find 64 spatial weighting coefficients w_{ik} for the kth window.

[3] This is also called a forward-model and is computed from w_{kt} and x_{kt} within the relevant time window (see [10], [17] for more detail).

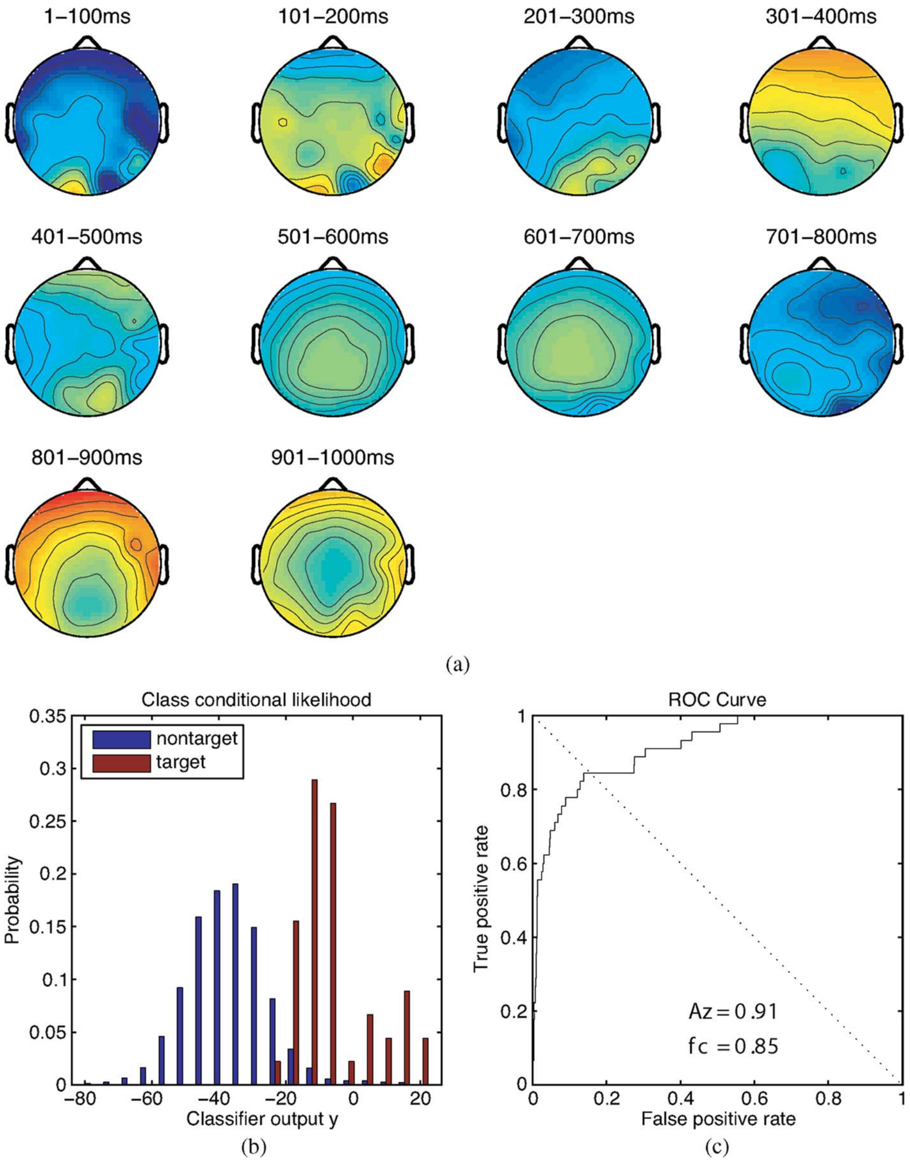


Fig. 1. ***Activity extracted by the hierarchical discriminant component analysis method. (a) Plotted is the forward model for the discriminating component at each time window, which can also be seen as the normalized correlation between the component activity in that window and the data [17]. The colormap for the scalp plot represents the normalized correlations, with red being positive and blue being negative. The series of ten spatial maps thus shows that the spatial distribution of the forward model of the discriminant activity changes across time. Activity at 300–400 ms has a spatial distribution which is characteristic of a P3f, which has been previously identified by our group and others [18], [19] during visual oddball and RSVP paradigms. In addition, the parietal activity from 500–700 ms is consistent with the P3b (or P300) indicative of attentional orienting. Other significant discriminant signal can be found at earlier and later time and often vary from subject to subject and the specifics of the experimental paradigm, e.g., presentation speed. Note that all scalp maps are on the same color scale. (b) The ten components characterized by the scalp maps above are linearly integrated to form a single classification score, which can be represented via the class-conditional histograms. The performance of the classification is established via the ROC curve which is computed from these class-conditional histograms.***

III. SYNERGISTIC COMBINATIONS OF COMPUTER AND EEG-BASED HUMAN VISION

Given the HDCA algorithm for EEG decoding, we consider how to integrate the results of EEG-based binary classifications with computer vision (CV). There are three basic modes for creating such a cortically coupled computer vision system.

- *Computer vision followed by EEG-RSVP, i.e.,* $T(\cdot)_{\text{CV}}$ *followed by* $T(\cdot)_{\text{EEG}}$: Given prior information of a target type (e.g., examples of the target class or description of the features and or context associated

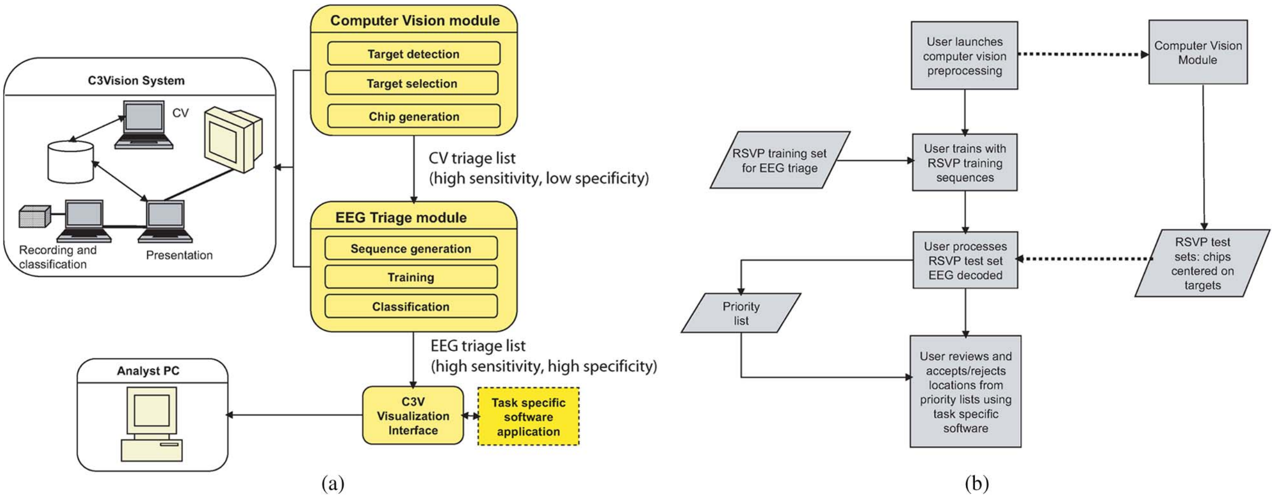

Computer Vision module
Target detection
Target selection
Chip generation
C3Vision System
CV
Recording and classification
Presentation
CV triage list
(high sensitivity, low specificity)
EEG Triage module
Sequence generation
Training
Classification
EEG triage list
(high sensitivity, high specificity)
Analyst PC
C3V Visualization Interface
Task specific software application
(a)
User launches computer vision preprocessing
Computer Vision Module
RSVP training set for EEG triage
User trains with RSVP training sequences
User processes RSVP test set EEG decoded
RSVP test sets: chips centered on targets
Priority list
User reviews and accepts/rejects locations from priority lists using task specific software
(b)

Fig. 2. ***(a) Software, hardware, and functional components of the integrated computer vision–EEG triage system. The cortically coupled computer vision system (termed C3Vision) runs on three computers and includes a computer vision module, a EEG triage module, and an interface which enables the user to visualize the results of the combined triage in a suitable visualization environment. (b) Corresponding workflow. The user launches the computer vision processing which performs an analysis of the entire image, generating detections for likely locations of targets. These detections are used to divide the large image into smaller image chips for presentation via RSVP. At the same time the user wears an EEG cap and a separate set of images is used to train the EEG classifier. The classifier learns to decode the EEG signal and generates confidence scores indicating the level of "interest" of the image chip—i.e., how much the particular chip grabbed the user's attention. Image chips generated by the computer vision module are passed to the user and EEG is decoded and scores generated for this "testing" data. The resulting scores are used to generate a priority list, which is a rank of the chip, by EEG score, together with their corresponding location in the large image. This priority list is ingested in the visualization software which, for image analysts, allows the user to interact with the imagery in a way they are are most accustomed (pan, zoom, mark objects) while also providing a toolbar to jump around the image based on the EEG-based score.***

with the target class) one can instantiate a CV model, including contextual cues, to operate on $\mathcal{D}_i$ so as to eliminate regions of very low target probability and provide an initial ordering of regions that have high target probability. In addition, CV should place potential regions of interest (ROIs) in the center of the test images. This will improve human detection performance since potential targets are then foveated when presented to the subject.[4] The top M images of the reordered $\mathcal{D}_i$ are sampled and presented to the subject for EEG-RSVP. Thus, the CV processing is tuned to produce high sensitivity and low specificity, with the EEG-RSVP mode increasing specificity while maintaining sensitivity;

- *EEG-RSVP followed by computer vision, i.e.,* $T(\cdot)_{\mathrm{EEG}}$ *followed by* $T(\cdot)_{\mathrm{CV}}$: In the absence of prior knowledge of the target type or a model of what an "interesting" image is, the EEG-RSVP is first run on samples of $\mathcal{D}_i$, the result being a reordering which ranks images based on how they attracted the subject's attention. This reordering is used to generate labels for a computer vision based learning system which, given a partial labeling of $\mathcal{D}_i$, propagates these labels and reorders the database. Thus, the EEG-RSVP is designed to identify a small number of "interesting" images which are then used by a semisupervised CV system to reorder the entire image database.
- *Tight coupling of EEG-RSVP and computer vision*; *i.e.,* $T(\cdot)_{\mathrm{CV}}$ *and* $T(\cdot)_{\mathrm{EEG}}$ *are applied in parallel and results integrated*: Both EEG-RSVP and CV are run in parallel and coupled either at the feature space level, or at the output (or confidence measure) level, leading to a single combined confidence measure that can serve as a priority indicator. As with the first coupling mode, this mode requires prior information on the target type.

These modes also potentially include feedback or multiple iterations within a closed-loop system. Below we describe two cortically coupled computer vision systems we have developed, together with results, which demonstrate the first two modes of fusion.

A. Computer Vision Followed by EEG

1) System Description: Fig. 2 illustrates the software, hardware, and functional components of our system for the triage application when objects of interest are well defined and known, a problem routinely encountered, for

[4] Our research has shown that, for the RSVP paradigm, the strength of the EEG signals falls substantially as a function of the eccentricity of the the target, thus indicating the importance of CV for centering potential targets.

example, by aerial image analysts. This system is composed of three main software modules:

- a computer vision preprocessing module that includes a target detection framework and a chipping engine;
- an EEG triage module using the RSVP paradigm to allow the user to rapidly browse through a selective set of image locations and to detect those most likely to contain an object of interest, which we call here a "target";
- a visualization interface for final confirmation by the user.

The computer vision target detection framework is a model-based framework that relies on two components:

1) *a feature dictionary*: a set of *low-level* feature extractors that—when combined—provide a general platform for describing and discriminating numerous different (aerial) object classes;
2) *grammar formulation and inference*: a formal grammar for specifying domain knowledge and describing complex objects using *mid-* and *high-level* compositions, and an algorithm that performs inference on such grammar.

While the feature dictionary can include generic, and perhaps complex, shape and feature extraction operators, our initial implementation has been specifically addressing the problem of detecting objects in aerial imagery. Thus, an expressive element for identifying an object in aerial images is its geometric shape, e.g., the shape of its edge contours, boundaries of different textures, etc. To ensure that our module remains scalable and adaptable to new objects of interest, we have employed a hierarchical approach, where shapes are approximated by a composition of shape parts. In turn, those parts can correspond to further compositions or geometric primitives. For instance, airfields in aerial images usually decompose into a set of straight lines (i.e., airstrips) and buildings (i.e., terminals), buildings into a set of lines.

In order to obtain a feature dictionary that also can help pruning regions of no interest to the human observer, we have included features that are used to assign image regions to one or more scene categories using a statistical classifier trained on small (typically ≈200 × 200 m) image patches. The categories are defined by a taxonomy of scene contexts, that discriminate between main scene contexts (i.e., terrain types such as "Forest" or "Desert") that occur in aerial scenes on a coarse level, and specific scene contexts, such as different building densities in an urban context, on a finer level. The statistical classifier and features are inspired by well-developed texture classification techniques that operate in the space of textons, where textons are identified as texture feature clusters using vector quantization of labeled data, as described by [20]. Descriptors are then constructed from histograms over the representative textons with regard to training images. This histogram representation serves as input for a statistical classifier, e.g., *k*-nearest neighbor.[5] Note that the dependency on training data is not problematic since the scene context is typically not task specific and, hence, can be precomputed for a wide problem domain and reused for unseen objects.

Our grammar and inference formulation [21], [22] is based on first order predicate logic. In the present context, predicate logic allows for: i) the specification of domain knowledge and ii) reasoning about propositions of interest. In order to properly deal with uncertainties in patterns (which are very typical in computer vision problems), the predicate logic formulation is augmented with a mathematical structure called bilattice [26]. Bilattices assume partial orders along the two axes of truth and the amount of evidence, and, by doing so, provide a formal method of inference with uncertainties. For more detail, refer to [21]–[23].

The target detection framework is applied to a subset of pixels in large aerial images and assigns detection confidences to each pixel in this selection. The pixel selection is currently defined by a uniform grid, whose density can be determined based on the image type and content. Based on a user specified threshold for the detection confidence, a list of the most likely detection candidates is generated and passed to the chipping engine. The engine then generates image chips centered on the detection candidates.

The EEG triage module receives the list of image chips and detection details from the computer vision module, that includes pixel locations and detection confidence scores, and uses this input to generate the RSVP image sequences that will be used for triage. It then performs several functions: it acquires and records the EEG signals, orchestrates the RSVP, matches the EEG recordings with the presented images, trains an underlying triage classifier using training sequences, and uses the classifier with newly generated image sequences.

The triage system currently utilizes a 64 electrode EEG recording system (ActiveTwo, Biosemi, Germany) in a standard 10–20 montage. EEG is recorded at a 2048 Hz sampling rate. While the EEG is being recorded, the RSVP display module uses a dedicated interface to display blocks of images at the specified frame rate. Blocks are typically 100 images long with only a few targets per block ($N < 5$). The frame rate is set between 5 and 10 Hz depending on the difficulty of the target detection task—i.e., each image is shown for 100–200 ms before next image is shown. The interface draws from a pool of potential target chips and a pool of "distractors." The role of the distractors is to achieve a desired prevalence of target chips, that will maintain the human observer engaged in the presentation: if the prevalence is too low or too high, the observer may not keep an adequate focus and may more easily miss detections. Given that the computer vision outputs include some false positives, the number of distractors used depends in fact on the expected number of true target chips from the computer vision module.

Currently the triage system's classification module relies on the hierarchical discriminant component analysis

[5] We thank Roberto Tron and Rene Vidal for providing an implementation of their texture classification.

algorithm described in Section II-B. The triage classification module is used in two different stages: training and actual usage with new imagery. The training typically consists of a presentation of 25–35 blocks with a set number of known targets in each block. The training sequences need not be related to the test sequences, in terms of their content, as the premise of the approach is that it detects objects of interest, but is not sensitive to the signatures of specific objects, and can therefore maintain its detection performance from one type of imagery to another. Training is a significant stage of the triage process. It not only is the vehicle to training the EEG classifier, it is also a mechanism for providing some practice to the human observer. Therefore, to help the human observer maintain his level of attention and gauge his training performance, the RSVP display module also displays feedback screens on the training progress at the end of each block.

Once the triage module has completed the triage, it generates a list of chips and their associated classification confidences, which can be used to prioritize the visualization of the corresponding image locations. The visualization interface permits the visualization of the prioritized locations in an adequate software environment for the task or user at hand. For example, for image analysts, we have developed an interface to RemoteView (Overwatch Systems, Sterling VA), an imagery exploitation software application standardly used in the GeoIntelligence community. The interface provides a toolbar comparable to a playback control allowing the analyst to jump from one prioritized location to the next. Meanwhile the analyst still retains access to all of RemoteView's functionality.

Currently, the corresponding prototype hardware implementation uses three laptops, one for the RSVP display, one for the EEG recording and classification, and one for image processing.

2) Experiments and Results: To evaluate the performance of the integrated computer vision–RSVP triage system, we have performed experiments with five subjects and using satellite electro-optical (EO) greyscale imagery. The task of each subject was to detect surface-to-air missile (SAM) sites. A 27 552 × 16 324 image was processed and chipped by the computer vision module. Based on the computer vision performance (see Fig. 3), the top 40 chips were retained for RSVP. An additional 760 distractor chips were used leading to 8 blocks of 100 chips each, all presented at

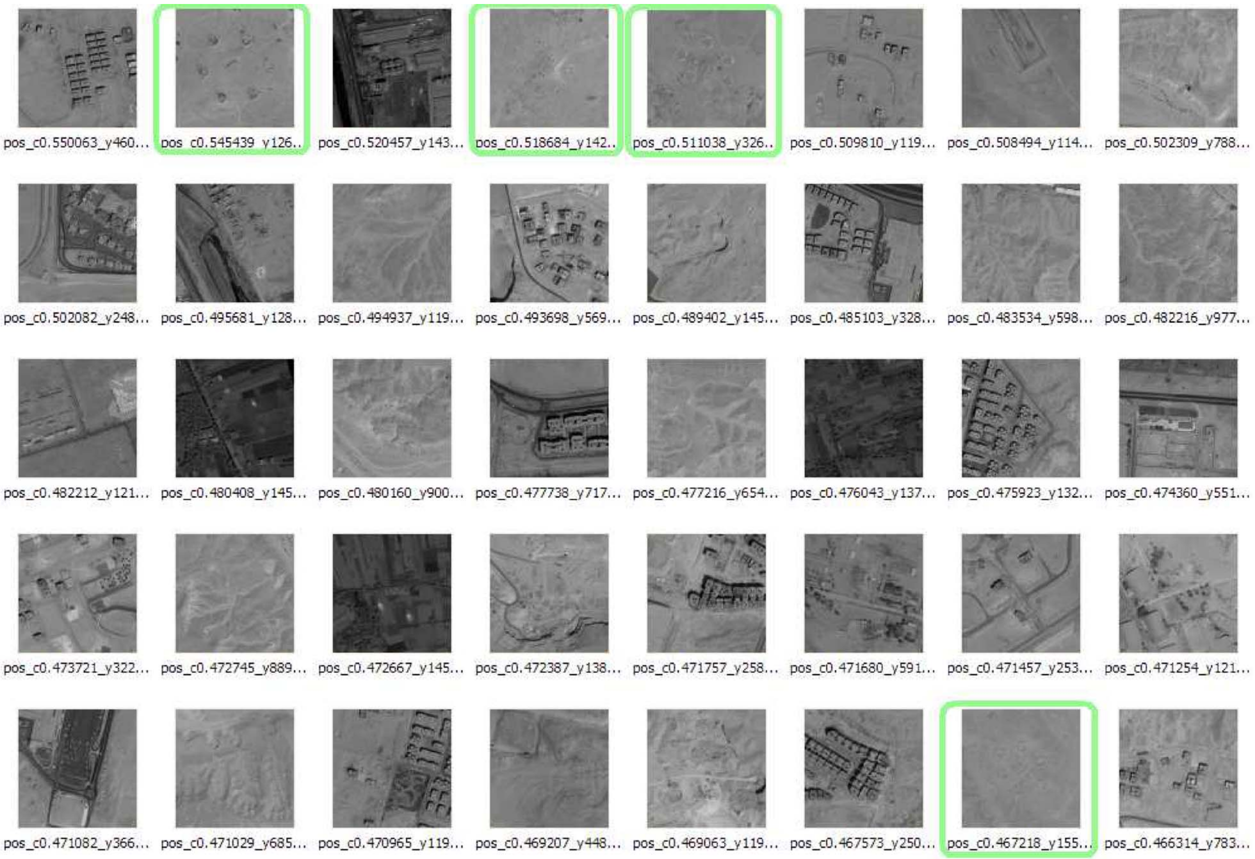


Fig. 3. ***Computer vision results for the surface-to-air missile (SAM) site detection problem. Detections are sorted by confidence (row-first). SAM site chips are marked in green.***

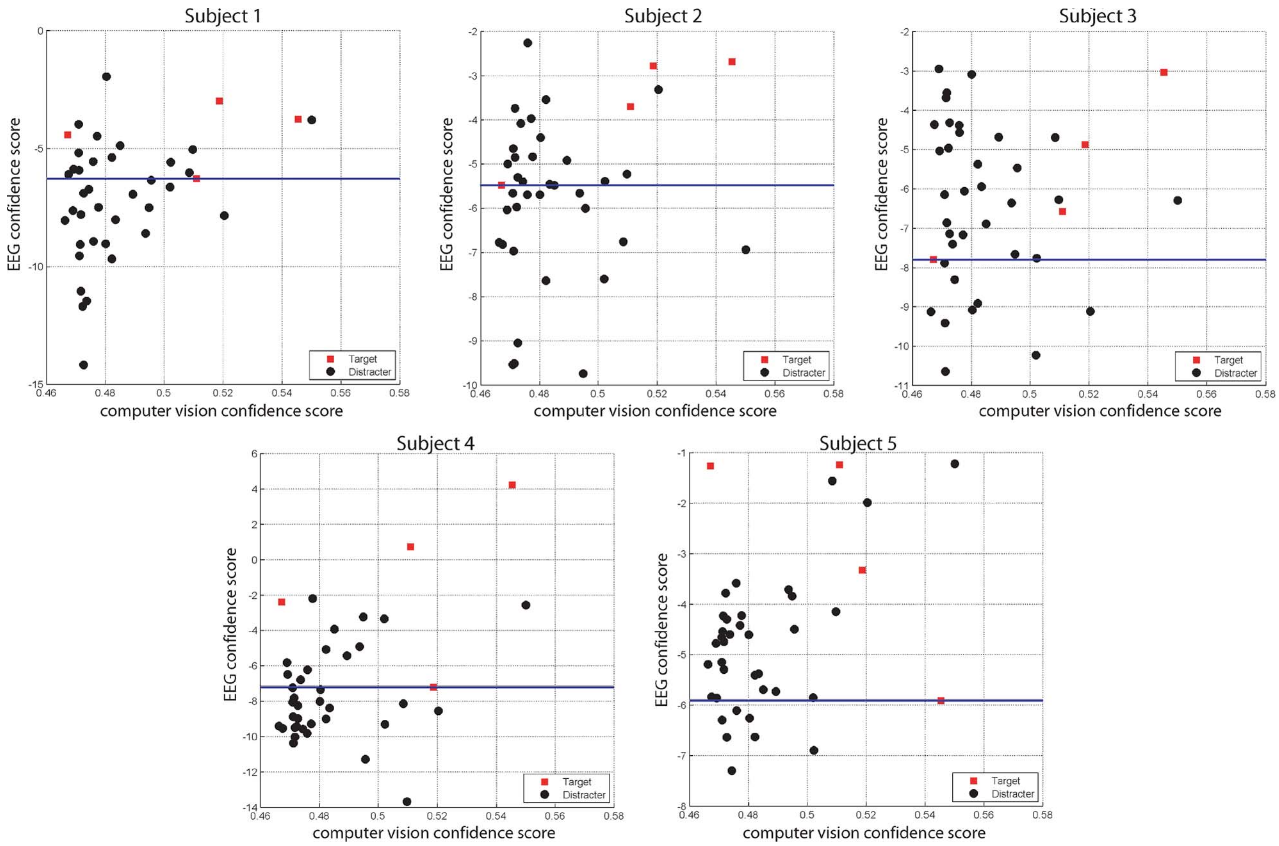


Fig. 4. ***Comparison of the detection performance of the integrated CV-RSVP triage system with the CV module alone: the integrated system's priority scores for the top 40 images are plotted against the CV detection confidences. The threshold for the priority scores needed to capture all true positives is plotted in blue, thus showing the reduced number of false positives with the integrated system.***

5 Hz. This frame rate was chosen based on preparatory experiments. Fig. 4 shows scatter plots combining the priority scores generated by the CV system alone with those of the integrated CV-RSVP triage system for these 40 chips. The plots also show the minimum priority score that should be chosen in order to capture all true positives with the integrated system, thus distinguishing the false positives that would be obtained with the integrated system for that threshold and those obtained with the CV alone. The figure shows that the integrated system leads to an overall reduction in the number of false positives of approximately 50% across all subjects. To place this back in the context of an image analyst's everyday tasks, an analyst would only need to review 40 locations in order to detect all targets using the integrated CV-RSVP triage system compared to potentially all 27 552 × 16 324 pixels without assistance. The productivity savings are dramatic, as demonstrated by Fig. 5 and Table 1. The CV-RSVP assisted condition results in a detection hit rate which on average is a factor of four improvement over baseline, with no increase in false alarms—in fact a decrease in false alarms. The time cost of the triage is small ($<$ 5 min) while the performance improvement substantial.

B. EEG Followed by Computer Vision: Bootstrapping a Computer Vision System

1) System Description: The system described above assumes some prior knowledge about the object of interest in order to construct a computer vision model. However, what if the system does not yet know what exactly the subject is looking for? Also, how could one use computer vision as a postprocessor to the EEG-based triage?

We have developed a second system to address these scenarios. The system allows a subject to browse through a limited number of images; it uses the EEG triage to take an initial "guess" as to what attracted the observer attention, and uses a computer vision module to pull additional positive examples from a database and correct potentially erroneous labels given by EEG classification. The system (see Fig. 6; additional detail can be found in [24]) is similar to the first system described, in that it uses the same type of components: computer vision module, EEG triage, visualization/review module. However, here the EEG triage module precedes the computer vision module. Additionally, the number of examples provided by the EEG triage may be insufficient to train conventional supervised

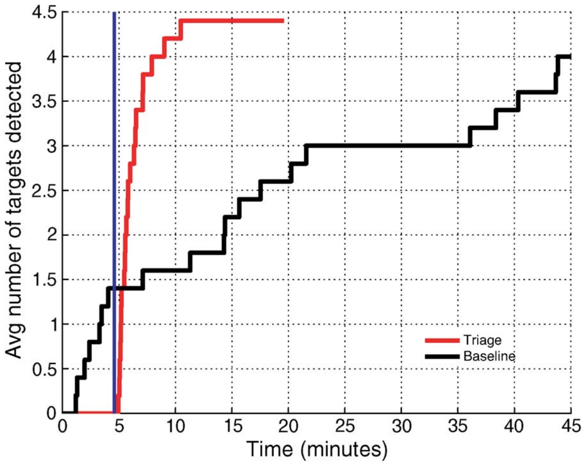


Fig. 5. ***Results showing average targets detected across subject (*$N = 5$*) as a function of time. Red curve is CV-RSVP assisted condition, black curve is baseline search. The relatively short time at the beginning of the search (*< 5 *min) in which the red curve is at zero represents the time it takes to do the RSVP triage. After that time, denoted by vertical blue bar, the triage list is loaded into the viewing software and the analysts use the triage results to jump to areas that resulted in a high EEG score and thus caught his/her attention during the RSVP. Clear is the substantial improvement in target detection rate for the assisted condition relative to baseline once the CV-RSVP generated triage listed is used.***

learning algorithms; and there may be inaccuracies in the EEG outputs due to typically lower sensitivity of the triage approach. So we use a computer vision module underpinned by a graph-based semisupervised learning algorithm. With this approach, the outputs of the EEG triage is a set of positive and negative examples (as determined by a suitable EEG confidence threshold), that serve as labeled inputs to a graph-based classifier to predict the labels of remaining unlabeled examples in a database and refine the initial labels produced by EEG-based classification.

Most semisupervised techniques focus on separating labeled samples into different classes while taking into account the distribution of the unlabeled data. The performance of such methods often suffers from the scarcity of the labeled data, invalid assumptions about classification models or distributions, and sensitivity to unreliable label conditions. Instead, our graph-based classification scheme, referred to as [26], incorporates novel graph-based label propagation methods and in real or near real-time receives refined labels for all remaining unlabeled data in the collection. By contrast to other semisupervised approaches, this graph-based label propagation paradigm makes few assumptions about the data and the classifier. One central principle of this paradigm is that data share and propagate their labels with other data in their proximity, defined in the context of a graph. Data are represented as nodes in a graph and the structure and edges in the graph define the relation among data. Propagation of labels among data in a graph is intuitive, flexible, and effective, without requiring complex models for the classifier and data distributions. This graph inference method has also been shown to improve over existing graph learning approach in terms of the sensitivity to weak labels, graph structure, and noisy data distributions [26].

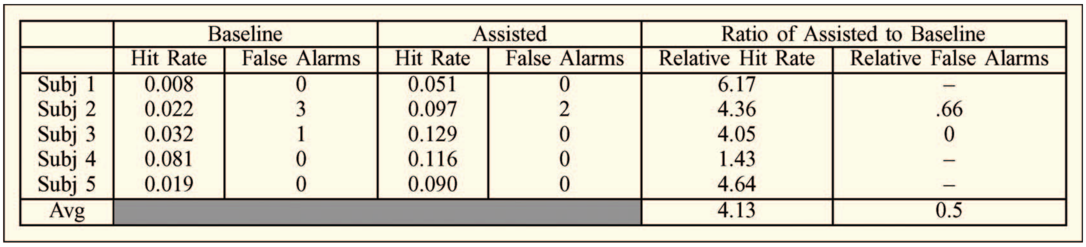
Based on GTAM algorithm, we developed a prototype system called transductive annotation by graph (TAG) for image search and retrieval. The processing pipeline of the TAG module contains the following components:

- input of labeled example provided by the EEG triage;
- image preprocessing components, such as denoising, enhancement, and filtering;
- image feature extraction to quantize the visual context;
- affinity graph construction;
- automatic label prediction via graph based inferencing and label correction.

One important aspect of this integration of EEG with computer vision is the ability for the computer vision TAG module to deal with uncertainty in the EEG labeling. Specifically, we have developed a "self-tuning" approach [24] that is able to identify the most reliable EEG inputs and reverses the labels of the most unreliable samples in order to optimize an objective function that captures labeling consistency and smoothness properties.

2) Experiments and Results: To illustrate the combination of EEG triage and automated labeling, we present

Table 1 Comparison of Hit Rate and False Alarm for SAM Site Search in 27 552 × 16 324 Satellite Image Without (Baseline) and With (Assisted) the Integrated CV-RSVP Triage System. Hit Rate Is Expressed in Fraction of Total Targets Detected per Minute



	Baseline		Assisted		Ratio of Assisted to Baseline	
	Hit Rate	False Alarms	Hit Rate	False Alarms	Relative Hit Rate	Relative False Alarms
Subj 1	0.008	0	0.051	0	6.17	–
Subj 2	0.022	3	0.097	2	4.36	.66
Subj 3	0.032	1	0.129	0	4.05	0
Subj 4	0.081	0	0.116	0	1.43	–
Subj 5	0.019	0	0.090	0	4.64	–
Avg					4.13	0.5

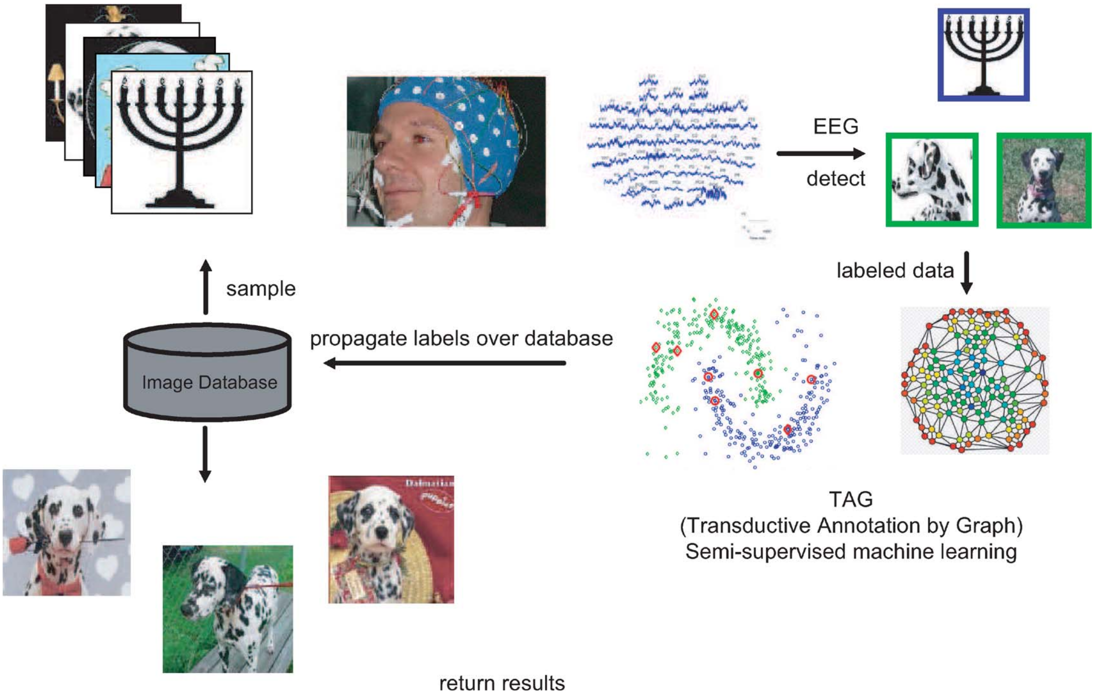

EEG
detect
labeled data
sample
propagate labels over database
Image Database
TAG
(Transductive Annotation by Graph)
Semi-supervised machine learning
return results

Fig. 6. ***System architecture for using EEG to bootstrap computer vision model construction. A sample set of images is taken from a database and the subject processes these images in RSVP mode while EEG is simultaneously recorded. The EEG is decoded and used to tag images in terms of how strong they grabbed the user's attention. The images can be seen as being a small set of labeled images, some of which might be the images of interest (e.g., images of soldiers) and some of which just grabbed the users attention because of novelty (e.g., the fellow with the interesting hairstyle). These small sets of labeled images are used as training data in a transductive graphic model which operates in the features space of the image. The transductive model uses the limited training data and manifold structures in the image feature space to propagate the initial labels to the rest of the images in the database The system includes a self-tuning mechanism which enables removal of tagged by the EEG as being interesting, but that deviate from the manifold structures. For example, the image with the blue border can be interpreted as a false positive and removed based on self-tuning. The computer vision model is then used to predict the relevance (priority) scores of the rest of images in the database. Images taken from Caltech image database.***

experiments with the detection of helipads in EO satellite imagery. To perform the EEG triage, a 30 K × 30 K satellite image (DigiGlobe, Longmont, CO) was chipped into (500 × 500 pixels) tiles; tiles containing helipads were centered on those helipads and were presented using the RSVP paradigm described above to human observers at a rate of 10 Hz. The output priority list was provided to the automated labeling module and the EEG scores were used as positive and negative labels according to a predefined threshold. Fig. 7 shows the top 20 images ranked by the

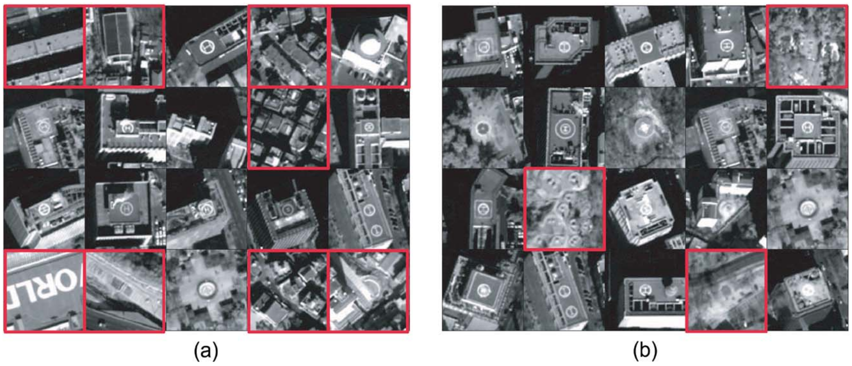


Fig. 7. ***Top 20 chips returned by: (a) the EEG triage alone and (b) the automated labeling module using the EEG triage priority scores as inputs. Chips with red squares are false positives. 45% of the top 20 chips returned by EEG are false positives and after passing into TAG system with self tuning this is reduced to 15%.***

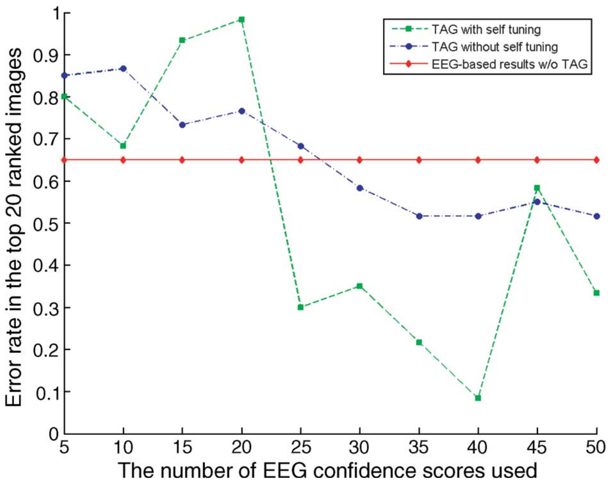


Fig. 8. ***Comparison of the error rate in the top 20 returned chips using the EEG triage alone, the automated labeling without self-tuning and the automated labeling with self-tuning.***

EEG priority list alone, and the top twenty images ranked by the combination of EEG triage and automated labeling. Fig. 8 shows the error rate in the top 20 returned images as a function of the number of EEG scores used as labels by the automated labeling algorithms. This figure highlights that without refining EEG-based classification results, the EEG triage error rate is significant, at approximately 65%. The unreliability of the EEG scores is reflected in the performance of the labeling algorithm without self-tuning, which is only able to marginally improve upon the EEG triage. However, with self-tuning, the automated labeling is able to infer the most unreliable EEG scores and significantly lower the error rate. The number of EEG labels used needs to be optimally chosen to obtain the best detection improvement, and Fig. 8 points to the existence of a single optimal number. Being able to automatically determine this number is one of the efforts we are currently focusing on.

IV. CONCLUSION

In this paper we have described two systems for cortically coupled computer vision which use computer vision (CV) and EEG to construct triage transforms for prioritizing imagery. Our results for $T(\cdot)_{\mathrm{CV}}$ followed by $T(\cdot)_{\mathrm{EEG}}$ show that EEG-based prioritization is effective for increasing the specificity of the CV system classification, without loss in sensitivity, for a realistic aerial image search task. Comparison of the system to baseline ultimately results in a factor of four improvement in the rate of target detection, without an increase in the rate of false alarms. Our preliminary results for $T(\cdot)_{\mathrm{EEG}}$ followed by $T(\cdot)_{\mathrm{CV}}$ show that computer vision can effectively use the EEG-based priority scores as noisy labels for building a visual similarity based model which can improve the specificity of the EEG triage. Note that this result was for a predefined and well-localized target (e.g., helipads). Our current efforts are investigating using $T(\cdot)_{\mathrm{EEG}}$ followed by $T(\cdot)_{\mathrm{CV}}$ system for identifying "targets" in which the user is not cued to identify a particular target class and/or the target class is less well-defined *a priori*. In addition, we are considering these triage systems for the case in which multiple classes in a database might attract a given user's interest.

We have described our first attempts to synergistically couple state-of-the-art computer vision with brain–computer interface technology to improve the searching of imagery and video content. The basic framework, namely coupling the speed of computer processing and analysis with the general purpose recognition and detection of sensory information by the human brain, is not limited to imagery. In general, systems for information triage, regardless of the signal type could be constructed using a similar framework, with the caveat being that the system design should maximize the complementarity of the two systems (human and computer). For example, if one were to use the same system design for triaging audio recordings, a rapid playback of the audio to the user would likely result in substantial distortion and result in low detection rates via EEG decoding. It is clear that careful attention must be paid toward how the particular sensory system (visual, auditory, etc.) is best presented with the information to be triaged, specifically so that it is presented to be maximally selective to patterns of interest in the data when the data is presented rapidly.

The systems we are developing can potentially shed light on some basic questions underlying the neuroscientific basis of rapid decision making. For example, can subjects be trained to improve their sensitivity for target and/or "interesting" objects and is such improvement accompanied by characteristic changes in the neural activity? Previous work by our group [27]–[29] has shown that there is a cascade of processes, detectable via single-trial analysis of EEG, which represent the constituent processes of decision making and that some of these processes are modulated by task difficulty. As subjects perform better for identical stimuli, we expect that the EEG correlates of these processes could change. By seeing which signatures are affected by training and which are not, we might get better insight into whether changes in the neural signatures precede behavioral changes and perhaps develop better theories on how training affects rapid decision making.

There is some debate on whether the system can be driven so that we can detect subliminal (or subconscious) events. In our work we have assumed that all detections are consciously processed by the subject—i.e., we have no evidence that the signals represent unconscious or "subliminal" processing. However it is interesting to consider how signals based on conscious events might be used to lower a behavioral threshold for initiating a decision. For example, in some cases, subjects might be instructed to look for a given type of target, find it (resulting in a neural signature of the detection event), and continue to analyze the image in spite of the detection. This type of "overanalyzing" of the imagery can be reduced by using the decoded neural signature of the recognition event to disengage the user from the current search, forcing him/her to effectively lower their decision threshold and move to the next image.

Finally, the potential for applications of brain–computer interfaces, outside the area of neurorehabilitation and neuroprosthetics, is tremendous. In addition to imagery, or more generally information triage, two potentially significant areas are video gaming and neuromarketing, which are already receiving substantial attention and interest. As BCI systems are developed and deployed, two fundamental issues must be considered. The first is ethical, and pertains to issues of privacy and the ramifications of being able to read someone's "thoughts" or intent, even if they do not act on them. The field of neuroethics [30] has emerged in response to the complicated and important questions related to such a new form of brain monitoring technology. A second fundamental issue is one related to human factors, and addresses the important question of how we will interact with a system that can "read our minds." When we interact with our personal computer we see the response of the computer following our own actions, such as moving a mouse or typing a key. How will we perceive the interaction when the human–computer interface does not require us to behaviorally interact? These and other neuroethical and human factors questions ultimately will play an important role in how BCI systems are integrated into our society. ■

APPENDIX
OTHER LINEAR METHODS FOR EEG DECODING

In Section II-B we described the hierarchical discriminant component analysis (HDCA) algorithm for decoding EEG for the image triage application. This algorithm can be easily implemented in real time and thus it is the algorithm of choice for our current C3Vision system. Recently we developed a set of new algorithms to improve EEG detection performance. Currently, their increased computational complexity limit their use to non-realtime applications. However, it is worth reporting them here given their significant improvement in classification accuracy.

A. Bilinear Discriminant Component Analysis

The HDCA algorithm described in Section II-B combines activity linearly. This is motivated by the notion that a linear combination of voltages corresponds to a current source, presumably of neuronal origin within the skull [17]. Thus, this type of linear analysis is sometimes called source-space analysis.[6] The most general form of combining voltages linearly in space and time would be

$$y = \sum_t \sum_i w_{it} x_{it}. \tag{5}$$

However, the number of free parameters w_{it} in this most general form is the full set of dimensions—6400 for the examples we consider—with only a handful of positive exemplars to choose their values. To limit the degrees of freedom one can restrict the matrix w_{it} to be of lower rank, say K. The linear summation can then be written as

$$y = \sum_{k=1}^{K} \sum_t \sum_i v_{tk} u_{ik} x_{it} \tag{6}$$

where $w_{it} = \sum_{k=1}^{K} v_{tk} u_{ik}$ is a low-rank bilinear representation of the full parameter space.

This bilinear model assumes that discriminant current sources are static in space with their magnitude (and possibly polarity) changing in time. The model allows for K such components with their spatial distribution captured by u_{ik} and their temporal trajectory integrated with weights v_{tk}. Again, the goal is to find coefficient u_{ik}, v_{tk} such that the bilinear projection is larger for positive examples than for negative examples, i.e., $y_+ > y_-$.

In addition, it is beneficial to assume that these coefficients are smooth, i.e., they do not differ much from their neighbors, thus implicitly assuming that the discriminant activity is correlated across neighboring electrodes and neighboring time samples (i.e., the discriminant activity is low frequency). In [31] we present an algorithm to find these coefficients simultaneously for all K.[7]

B. Bilinear Feature Based Discriminants

The algorithms presented so far will only capture a type of activity called event-related potentials (ERP). This term refers to activity that is evoked in a fixed temporal relationship to an external event; that is, positive and negative deflections occur at always the same time relative to the event—in our case, the time of image presentation. In addition to this type of evoked response activity the EEG shows variations in the strength of oscillatory activity. Observable events may change the magnitude of ongoing oscillatory activity or may induce oscillations in the EEG. A linear summation will not be able to capture oscillatory activity, since for oscillations the phase and therefore the polarity of the signal may change from trial to trial. To capture the strength of an oscillation, irrespective of polarity, it is common to measure the "power," or the square of the signal, typically after it has been filtered in a specific frequency band. Instead of a linear combination to capture power, one has to allow for a quadratic combination of the electrical potentials.

[6] "Beam-forming" is a common misnomer for the same.

[7] More recently we have found it beneficial to estimate the parameters for one component, then subtract the activity spanned by the bilinear subspace of that component, and then estimate the activity for an additional component on the remaining subspace—a process that may be repeated several times to estimate additional components (executable code implementing this idea can be found at [32]).

A difficulty that arises in this context is the choice of frequency band. In addition to spatial and temporal coefficients one has to now choose coefficients for different frequencies, which in practice may increase the degrees of freedom by one order of magnitude or more. Thus, it may be difficult to find an optimal combination of space-time-frequency features without *a priori* knowledge as to which frequency band contains discriminative information. Here we present a novel algorithm that can identify an appropriate combination of first and second-order features.[8] The main idea is to identify feature invariances by analyzing data collected from multiple subjects on the same experimental paradigm. While the specific coefficient combining different first and second order features may vary from subject to subject, we assume that the relevance of evoked potentials (first order) and induced oscillations in different frequency bands (second order) does not change significantly across subjects. The features we consider here are the instantaneous power in different frequency bands so as to capture temporal changes in oscillatory power in addition to frequency and spatial information. We denote here with $f_{kt}(x_1,\ldots,x_T)$ the kth feature of the time sequence $\mathbf{x}=x_1,\ldots,x_T$ evaluated around time t. A bilinear discriminant model can be formulated for each feature as follows:

$$y_k=\sum_t\sum_i v_{tk}u_{ik}f_{kt}(\mathbf{x}_i) \tag{7}$$

where the spectro-temporal features are evaluated separately for each electrode i providing the time sequence $\mathbf{x}_i$. Note that in this formulation one of the features could simply be the original evoked response signal, $f_{1t}=x_t$, i.e., the linear features as before. The total model combines different features

$$y=\sum_k w_k y_k \tag{8}$$

with the goal of including only a small subset of nonzero values for w_k. While $v_{tk}u_{ik}$ will be chosen differently for different subjects, the goal is to pick w_k with the same set of nonzero values for different subjects, i.e., the same features are selected for different subjects. For the sake of computational efficiency we assume that the information provided by each feature is independent from another feature. Thus, the bilinear coefficient v_{kt},u_{ki} may be selected separately for each subject and each feature. Once selected, all v_{kt},u_{ki} remain constant, and only w_k has to be found with a subset of nonzero coefficient such that consistently good performance is found across all subjects. This is a potentially large combinatorial search problem which can be solved in limited time only using greedy methods. Various heuristic strategies for a greedy feature search can be envisioned. Here we begin by selecting a single feature that performs for the largest number of subjects among the top M features. Then we test this feature as a pair ($K=2$) in combination with each one of the other features and select again the one that (together with the first) performed most often among the top M features. This process can be repeated several times to increase the the number of selected features. We restrict ourselves here to a set of $K=3$ features. There is nothing particular about this specific version of greedy search; any other features selection strategy that is based on invariance across subjects is expected to perform equally well.

C. Comparing Results for the Three Algorithms

Discrimination results for the three algorithms we have discussed are shown in Fig. 9. The data are from 14 datasets obtained during a set of RSVP experiments on 5 subjects (3 datasets per subject with one dataset excluded as it was used to optimize regularization parameters). Due to memory limitations we only used 250 example images out of a total of 2500 (50 positives and 200 negatives). For this result the data was down-sampled to 256 Hz from 2048 Hz. The BDCA used a single component here ($K=1$). Test-set performance reported here is the result of fivefold cross validation. The feature selection procedure had no access to the test data. The algorithm was given quadratic features that capture power in various frequency bands in addition to the linear features that were used also by the other two algorithms. Specifically, we used a time-resolved estimate of power obtained with a sliding multitapered windows of 150 ms duration. The two most important features extracted by the algorithm are the

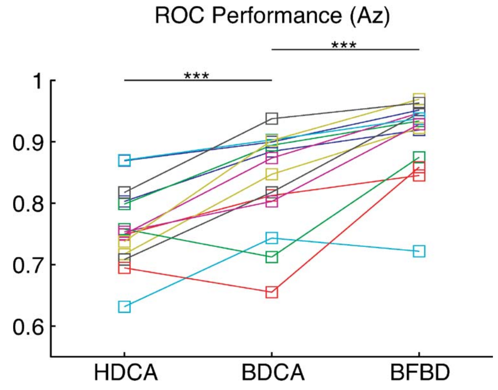


Fig. 9. ***Performance comparison for 14 datasets collected from 5 subjects on various types of images on the RSVP task. The Az performance for the three algorithms is (mean ± std.): 0.76 ± 0.07, 0.83 ± 0.08 and 0.91 ± 0.07. Statistical significance $***$ indicates here $p<0.001$ and was computed using a Wilcoxon signed rank test.***

[8] For full detail see [33]. An earlier algorithm that combines linear and quadratic features in source-space was presented in [34].

conventional linear features and an estimate of power in higher frequencies (20–40 Hz). The resulting bilinear coefficients, v_{kt}, u_{ki}, indicate that images of interest elicited increased power in this frequency band following image presentation with a nonuniform spatial distribution.

Acknowledgment

Satellite imagery provided by DigitalGlobe. C. Bahlmann and M. Singh would like to thank Dr. Visvanathan Ramesh and Dr. Vinay Shet from Siemens Corporate Research for their discussions and technical support.

REFERENCES

[1] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[2] A. Oliva, "Gist of the scene," in *Encyclopedia of Neurobiology of Attention*. San Diego, CA: Elsevier, 2005, pp. 251–256.

[3] M. C. Potter and E. I. Levy, "Recognition memory for a rapid sequence of pictures," *J. Exp. Psychol.*, vol. 81, no. 1, pp. 10–15, 1969.

[4] M. M. Chun and M. C. Potter, "A two-stage model for multiple target detection in rapid serial visual presentation," *J. Exp. Psychol., Hum. Percept. Perform.*, vol. 21, no. 1, pp. 109–127, 1995.

[5] C. Keysers, D.-K. Xiao, P. Foldiak, and D. I. Perrett, "The speed of sight," *J. Cogn. Neurosci.*, vol. 13, no. 1, pp. 90–101, 2001.

[6] R. VanRullen and S. J. Thorpe, "The time course of visual processing: From early perception to decision-making," *J. Cogn. Neurosci.*, vol. 13, no. 4, pp. 454–461, 2001.

[7] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1996.

[8] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortically-coupled computer vision for rapid image search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 174–179, Jun. 2006.

[9] P. Sajda, A. Gerson, and L. Parra, "High-throughput image search via single-trial event detection in a rapid serial visual presentation task," in *Proc. 1st Int. IEEE EMBS Conf. Neural Eng.*, Mar. 20–22, 2003, pp. 7–10.

[10] L. C. Parra, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. G. Philiastides, and P. Sajda, "Spatiotemporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 95–115, Jan. 2008.

[11] S. Mathan, S. Whitlow, D. Erdogmus, M. Pavel, P. Ververs, and M. Dorneich, "Neurophysiologically driven image triage: A pilot study," in *CHI '06 Extended Abstracts Hum. Factors in Comput. Syst.*, New York, USA, 2006, pp. 1085–1090,1-59593-298-4, ACM.

[12] A. Kapoor, P. Shenoy, and D. Tan, "Combining brain computer interfaces with vision for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR 2008)*, Jun. 23–28, 2008, pp. 1–8.

[13] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Oct. 2008.

[14] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, pp. 352–356, 2008.

[15] 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*, vol. 60, no. 5, pp. 915–929, 2008.

[16] D. E. J. Linden, "The P300: Where in the brain is it produced and what does it tell us?" *Neuroscientist*, vol. 11, no. 6, pp. 563–576, Oct. 2005.

[17] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Recipes for the linear analysis of eeg," *Neuroimage*, vol. 28, no. 2, pp. 326–341, Nov. 2005.

[18] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortical origins of response time variability during rapid discrimination of visual objects," *Neuroimage*, vol. 28, no. 2, pp. 342–353, 2005.

[19] S. Makeig, M. Westerfield, T.-P. Jung, J. Covington, J. Townsend, T. Sejnowski, and E. Courchesne, "Independent components of the late positive response complex in a visual spatial attention task," *J. Neurosci.*, vol. 19, pp. 2665–2680, 1999.

[20] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis. (Special Issue on Texture Analysis and Synthesis)*, vol. 62, no. 1–2, pp. 61–81, Apr. 2005.

[21] V. Shet, D. Harwood, and L. Davis, "Multivalued default logic for identity maintenance in visual surveillance," in *Computer Vision—CCV*. Graz, Austria: Springer-Verlag, 2006, pp. IV: 119–IV: 132.

[22] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis, "Bilattice-based logical reasoning for human detection," presented at the IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), Minneapolis, MN, 2007.

[23] M. L. Ginsberg, "Multivalued logics: A uniform approach to inference in artificial intelligence," *COMIN*, vol. 4, no. 3, pp. 256–316, 1988.

[24] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang, "Brain state decoding for rapid image retrieval," in *ACM MultiMedia*, Beijing, China, 2009, pp. 945–954.

[25] J. Wang and S. F. Chang, "Columbia Tag System—Transductive Annotation by Graph Version 1.0," Columbia University, Tech. Rep. ADVENT Tech. Rep. 225-2008-3, 2008.

[26] J. Wang, T. Jebara, and S. F. Chang, "Graph transduction via alternating minimization," presented at the Int. Conf. Mach. Learn., Helsinki, Finland, Jul. 2008.

[27] M. G. Philiastides and P. Sajda, "Temporal characterization of the neural correlates of perceptual decision making in the human brain," *Cereb. Cortex*, vol. 16, no. 4, pp. 509–518, 2006.

[28] M. G. Philiastides, R. Ratcliff, and P. Sajda, "Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram," *J. Neurosci.*, vol. 26, no. 35, pp. 8965–8975, 2006.

[29] M. G. Philiastides and P. Sajda, "EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making," *J. Neurosci*, vol. 27, no. 48, pp. 13082–13091, Nov. 2007.

[30] M. J. Farah, "Emerging ethical issues in neuroscience," *Nature Neurosci.*, vol. 5, no. 11, pp. 1102–1123, 2004.

[31] M. Dyrholm, C. Christoforos, and L. C. Parra, "Bilinear discriminant component analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1097–1111, 2007.

[32] M. Dyrholm, Source Code and Improvements of Dyrholmd *et al.*, 2007, bdca.googlecode.com.

[33] C. Christoforou, "The Bilinear Brain: Bilinear Methods for EEG Analysis and Brain Computer Interfaces," Ph.D. thesis, Graduate Center, City Univ., New York, Jan. 2009.

[34] C. Christoforou, P. Sajda, and L. C. Parra, "Second order bilinear discriminant analysis for single trial eeg analysis," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 313–320.

ABOUT THE AUTHORS

Paul Sajda (Senior Member, IEEE) received the B.S. degree in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1989 and the M.S. and Ph.D. degrees in bioengineering from the University of Pennsylvania, Philadelphia, in 1992 and 1994, respectively.

In 1994 he joined the David Sarnoff Research Center where he went on to become the Head of the Adaptive Image and Signal Processing Group. He is currently an Associate Professor of Biomedical Engineering and Radiology at Columbia University, New York, where he is Director of the Laboratory for Intelligent Imaging and Neural Computing (LIINC). His research focuses on neural engineering, neuroimaging, computational neural modeling, and machine learning applied to image understanding.

Prof. Sajda has received several awards for his research including an NSF CAREER Award and the Sarnoff Technical Achievement Award. He is an elected Fellow of the American Institute of Medical and Biological Engineering (AIMBE). He serves as an Associate Editor for IEEE Transactions on Biomedical Engineering, and a member of the IEEE Technical Committee on Neuroengineering.

Eric Pohlmeyer received the B.S. degree in mechanical engineering from the University of Cincinnati, Cincinnati, in 2001, and the M.S. and Ph.D. degrees in biomedical engineering from Northwestern University, Evanston, IL, in 2004 and 2008, respectively.

He is a Postdoctoral Fellow in the Biomedical Engineering Department in the Laboratory for Intelligent Imaging and Neural Computing (LIINC) at Columbia University, New York. He works in brain–computer interfacing and has constructed a system capable of translating desired hand movements from the brain into electrical stimulation of paralyzed muscles in order to restore wrist function in nonhuman primates. He has also worked with EEG-based neural interfaces, in particular with cortically coupled computer vision (C3Vision) systems that incorporate EEG recordings with computer vision systems in order to help individuals sort through large image databases to find specific images.

Jun Wang received the B.S. degree from Shanghai JiaoTong University in 1998, and the M.S. degree from Tsinghua University in 2003. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering Department, Columbia University, New York.

He also worked as Research Assistant at Harvard Medical School in 2006, and as Research Intern at Google New York in 2009. His research interests include image retrieval, machine learning, and hybrid neural–computer vision systems.

Lucas C. Parra (Senior Member, IEEE) received the Ph.D. degree in physics from the Ludwig-Maximilian University, Germany, in 1996.

He is Professor of Biomedical Engineering at the City College of the City University of New York. Previously he was head of the adaptive image and signal processing group at Sarnoff Corporation (1997–2003) and member of the machine learning and the imaging departments at Siemens Corporate Research (1995–1997). His areas of expertise include machine learning, acoustic array processing, emission tomography, and electroencephalography. His current research in biomedical signal processing focuses on functional brain imaging and computational models of the central nervous system.

Christoforos Christoforou received the Ph.D. degree in computer science from the Graduate Center of the City University of New York in 2009.

He is the Chief Research Scientist at R.K.I Leaders Limited, Aradippou, Cyprus. He holds the rank of Special Scientist of Electrical Engineering at the Cyprus University of Technology. His research focuses on machine learning, pattern recognition applied in the areas of single-trial EEG analysis, computational biology, and natural language processing.

Jacek Dmochowski received the B.Eng. degree (with High Distinction in Communications Engineering) and the M.A.Sc. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2003 and 2005, respectively, and the Ph.D. degree in Telecommunications (granted "exceptionnelle") from the University of Quebec-INRS-EMT, Canada, in 2008.

He is currently a Postdoctoral Fellow at the Department of Biomedical Engineering of the City College of New York, City University of New York, and is the recipient of the National Sciences and Engineering Research Council (NSERC) of Canada Post Doctoral Fellowship (2008–2010). His research interests lie in the area of multichannel statistical signal processing and include machine learning of neural signals, decoding of brain states, and neuronal modeling.

Barbara Hanna (Member, IEEE) received the B.A. and M. Eng. degrees from the University of Cambridge, U.K., in 1997 and the Ph.D. degree in computer vision from the University of Surrey, U.K., in 2001.

In 2001, she joined the David Sarnoff Research Center where she designed and developed real-time video processing systems, and went on to become Technical Lead for medical vision initiatives. In 2007, she became the Program Manager for Research and Development led by the LIINC lab at Columbia University under the DARPA NIA Phase 2 Program. She is currently the CEO of Neuromatters, New York, and focuses on the design and development of novel brain machine interfaces to deal with information overload. Her areas of expertise include computer vision and real-time image and video processing.

Claus Bahlmann received the Ph.D. degree in computer science with the highest of honors from the University of Freiburg, Germany, in 2005.

Since 2004, he has been Postdoctoral Staff Member, Research Scientist, and Project Manager in the Real-time Vision and Modeling Department at Siemens Corporate Research (SCR), Princeton, NJ. His research interests include pattern recognition, computer vision, and machine learning. He has applied these techniques in various fields, including handwriting recognition, automotive, and medical.

Dr. Bahlmann was awarded Best Paper at the IWFHR 2002 conference for his work "On-line Handwriting Recognition with Support Vector Machines—A Kernel Approach." In 2005, his Ph.D. thesis, "Advanced Sequence Classification Techniques Applied to Online Handwriting Recognition," earned the Wolfgang-Gentner Nachwuchsförderpreis award from the University of Freiburg.

Maneesh Kumar Singh (Member, IEEE) received the B.Tech. degree in electrical engineering and the M.Tech. degree in communication and radar engineering from the Indian Institute of Technology, Delhi, in 1993 and 1996 respectively, and the Ph.D. degree in electrical and computer engineering from the University of Illinois, Urbana, in 2003.

He was a Postdoctoral Research Associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign in 2003–2004. Since 2004, he has been a Research Scientist at Siemens Corporate Research, Princeton, NJ, in the Real-time Vision and Modeling Department. His current research interests include nonparametric statistics, density estimation, statistical computer vision, and applications of computer vision for medical diagnostics, industrial inspection, security, and surveillance.

Shih-Fu Chang (Fellow, IEEE) is Professor and Chairman of Electrical Engineering and Director of Digital Video and Multimedia Lab at Columbia University, New York. He has made significant contributions to multimedia search, media forensics, video adaptation, and international standards for multimedia indexing.

Prof. Chang has been recognized with several awards, including the IEEE Kiyo Tomiyasu Award, Navy ONR Young Investigator Award, IBM Faculty Award, ACM Recognition of Service Award, and NSF CAREER Award. He and his students have received several Best Paper and Best Student Paper Awards from IEEE, ACM, and SPIE. He has worked in different advising/consulting capacities for IBM, Microsoft, Kodak, PictureTel, and several other institutions. He was was Editor-in-Chief for IEEE Signal Processing Magazine during 2006–8.