Chapter 9
Cortically-Coupled Computer Vision

**Paul Sajda, Eric Pohlmeyer, Jun Wang,
Barbara Hanna, Lucas C. Parra,
and Shih-Fu Chang**

Abstract We have developed EEG-based BCI systems which couple human vision and computer vision for speeding the search of large images and image/video databases. We term these types of BCI systems "cortically-coupled computer vision" (C3Vision). C3Vision exploits (1) the ability of the human visual system to get the "gist" of a scene with brief (10's–100's of ms) and rapid serial (10 Hz) image presentations and (2) our ability to decode from the EEG whether, based on the gist, the scene is relevant, informative and/or grabs the user's attention. In this chapter we describe two system architectures for C3Vision that we have developed. The systems are designed to leverage the relative advantages, in both speed and recognition capabilities, of human and computer, with brain signals serving as the medium of communication of the user's intentions and cognitive state.

P. Sajda (✉) · E. Pohlmeyer
Department of Biomedical Engineering, Columbia University, New York, NY, USA
e-mail: psajda@columbia.edu

E. Pohlmeyer
e-mail: ep2473@columbia.edu

J. Wang · S.-F. Chang
Department of Electrical Engineering, Columbia University, New York, NY, USA

J. Wang
e-mail: jwang@ee.columbia.edu

S.-F. Chang
e-mail: sfchang@ee.columbia.edu

B. Hanna
Neuromatters, LLC, New York, NY, USA
e-mail: bhanna@neuromatters.com

L.C. Parra
City College of New York, New York, NY, USA
e-mail: parra@ccny.cuny.edu

D.S. Tan, A. Nijholt (eds.), *Brain-Computer Interfaces,*
Human-Computer Interaction Series,
DOI 10.1007/978-1-84996-272-8_9,

9.1 Introduction

Today we are faced with more information on a daily basis than ever before. Constantly evolving digital recording devices that can capture large amounts of spatial and/or temporal data, ever increasing digital storage capacities and multitudes of multimedia applications are just a few factors that create this "information tsunami". Searching for something of interest, making rapid decisions and being attentive to relevant information are becoming increasingly complex tasks.

Various technologies, driven by diverse fields of research, have been developed to assist us in consuming information. Yet the fact is that the human capacity to analyze information and make inferences about our surrounding environment remains unsurpassed. For example, our ability to recognize objects is extraordinarily robust, and with trillions of neuronal connections, our brain can react extremely fast to an external stimulus: we respond to the information we receive in the "blink of an eye" (Gladwell 2005), before we are even aware of it.

Recently we, as well as others, have been investigating the application of brain computer interfaces (BCI) for dealing with issues in image search, retrieval and triage (Gerson et al. 2006; Parra et al. 2008; Kapoor et al. 2008; Bigdely-Shamlo et al. 2008). Our group has developed an approach which we term *cortically coupled computer vision* (C3Vision) where the goal is to synergistically couple computer vision with human vision, via on-line real-time decoding of EEG while users' view images as a rapid serial visual presentation (RSVP) (Gerson et al. 2006). As well as being a method for maximizing throughput, the use of RSVP is motivated by our ability to make very rapid and accurate decisions. The ability of the human visual system to do this has sometimes been characterized as getting the "gist" of a scene (Oliva 2005) in a few hundred milliseconds. The C3Vision approach exploits our ability to decode EEG signals that are related to detection and recognition in rapidly shown images (Thorpe et al. 1996; Keysers et al. 2001; Gerson et al. 2006). One of the key signals we exploit in our system is the P300. The P300 is an evoked response in the EEG which reflects perceptual "orienting" or shift of attention which can be driven by the content of the sensory input stream (Linden 2005).

In this chapter we review our work in C3Vision, focusing on two architectures we have developed. The first architecture is tailored to a visual search problem, where a user must find targets in an extremely large image (on the order of 30 K $\times$ 30 K pixels). For this case computer vision serves as a pre-processor to select potential areas of interest, creating chips (or regions of interest—ROIs) of these areas which are subsequently presented to the user via RSVP while the user's EEG is decoded to generate an "interest score" used to rank or prioritize the ROIs (see Fig. 9.1A). Given this first step "triage", the user can proceed to search the large image with the added capability of jumping to locations in the scene which grabbed his/her's attention during the RSVP EEG decoding phase. In Section 9.3 we describe this system and demonstrate results for remote sensing.

The second architecture, presented in Section 9.4, addresses an image retrieval application, using EEG decoded during RSVP presentation to generate an interest

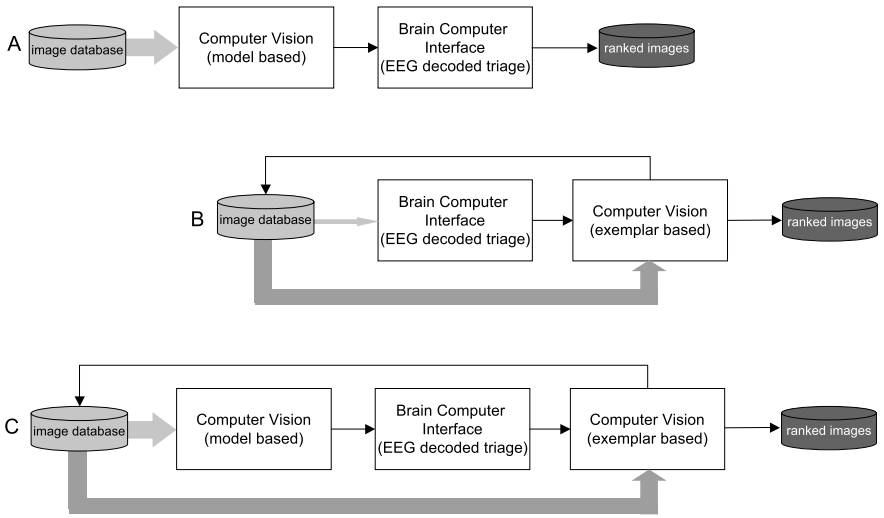


Fig. 9.1 Strategies for integrating computer vision with EEG-based image triage. The goal is to re-rank images in a database so that the result is a dense ranking, with images of interest at the front of the database, given an initial ranking that is sparse. In system A a very large database is processed by model based computer vision to generate a candidate set of images that might be of interest to the user. The computer vision model is tuned to have a high sensitivity at the cost of specificity. The user is presented with rapid sequences of images labeled as potentially interesting by computer vision while high-spatial density EEG (≈64 channels) is recorded. An EEG decoder is trained on data collected from subjects while they look at imagery from an unrelated database and pay attention for target specific or "interesting" imagery in the rapid sequences. The trained EEG decoder is used to process EEG signals while the user observes the barrage of images in the sequence, with the result being an interest score used to re-rank the database. This leads to a dense ranking of the database (note dark gray vs light gray, indicating the database has become more dense in term of the "concentration" of interesting images). System B starts by randomly sampling the database and passing on the samples as rapid sequences to a human user. Note that in this case, the volume of imagery assessed by the human is small, compared with the computer vision in System A, due to speed and fatigue limitations. However an advantage of System B is that the human is able to look for images which are specifically of interest to him/her and which might be difficult to define in terms of a prior model for computer vision. The EEG interest score is used to re-rank images and pass labels to an exemplar based computer vision system which then propagates predicted labels into the database and returns a re-ranking based on the propagated labels. System C combines systems A and B so that computer vision operates both at the front end and back end of the system

score usable for training a feature based computer vision system (see Fig. 9.1B). The computer vision system derives training labels from the EEG interest score and propagates them to re-rank the image database and retrieve for the user those images which match what grabbed his/her attention. Below we begin by describing how we decode the EEG and map it to an "interest score". For additional technical details readers are referred to Gerson et al. (2006), Parra et al. (2008), Wang et al. (2009b), Sajda et al. (2010).

9.2 The EEG Interest Score

Given an RSVP paradigm for presenting a rapid sequence of images to the subject, we simultaneously record EEG, using 64 scalp electrodes, 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 neuro-science perspective it can be seen as the single-trial correlate of the P300-related orienting response, though as can be seen in Fig. 9.2 we allow for flexibility in this definition. The algorithm we use for decoding the EEG, and ultimately mapping it to an interest score, has been described previously (Gerson et al. 2006; Parra et al. 2008). Briefly, our approach begins with the linear model,

$$y_t = \sum_i w_i x_{it} \tag{9.1}$$

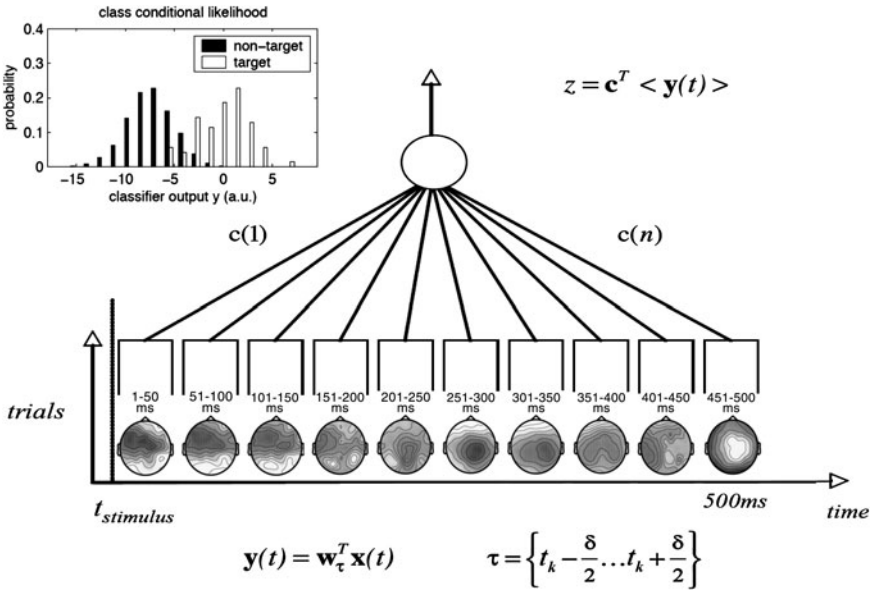


Fig. 9.2 Using hierarchical discriminant component analysis to construct EEG interest scores. Shown is the forward model for the discriminating component at each time window, which can also be seen as the normalized correlation between the component activity in that window and the data (Parra et al. 2005). The series of 10 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 (Gerson et al. 2005; Makeig et al. 1999) 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 signals 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. The 10 components characterized by the scalp maps are linearly integrated to form a single classification score, which can be represented via the class-conditional histograms. This classification score is used as the "interest score" in our C3Vision systems

where 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 will be learned based on a set of training data. The goal is to combine voltages in the electrodes linearly such that the sum y is maximally different between two conditions—e.g. "target of interest" vs "distractor". 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—i.e. we assume a stationarity time T of approximately 100 ms. Thus 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):

$$y_{kt} = \sum_i w_{ki} x_{it}, \quad t = T, 2T, \ldots, (k-1)T, kT. \tag{9.2}$$

These different y_{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", y_{IS} for the image

$$y_{IS} = \sum_t \sum_k v_k y_{tk}. \tag{9.3}$$

For on-line implementation purposes we use the method of Fisher Linear Discriminants to train coefficients w_{ik} within each time window of time. 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, we have termed this algorithm "Hierarchical Discriminant Component Analysis". Figure 9.2 plots the spatial filters that are learned for each time window and shows the subsequent hierarchical integration which enables us to construct interest scores, based on the classifier output. Note in the figure that the scores distribute as a function of whether the image was a target of interest or not.

9.3 C3Vision for Remote Sensing

We first consider architecture A in Fig. 9.1 with computer vision followed by RSVP and EEG decoding. The application we will consider is for remote sensing. In remote sensing, analysts must search images that are several hundreds of giga-pixels in size. In particular, intelligence analysts routinely search very large satellite images in order to find and monitor objects and regions of interest (Fig. 9.3A). As part of their work-flow, analysts use specialized software (e.g. Remoteview, by Overwatch Systems) that lets them rapidly load, display and zoom/pan such images. They conduct their searches using various strategies depending on their level of experience and expertise. Figure 9.3B shows for example the raster scanning search pattern followed by an image analyst during a test. Given the size of the images, this typical search process can be lengthy and inefficient. For example a trained analyst may need 60 minutes to complete the review of a 30 K × 30 K image, and may only identify the targets of interest in the last few minutes of the review. However, searches

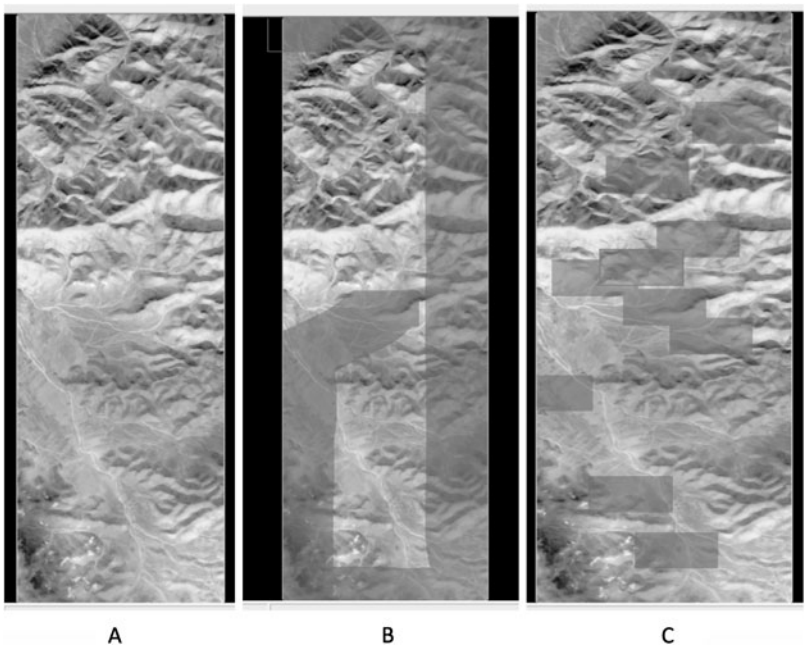


Fig. 9.3 **A**. Satellite image to be searched. **B**. Traditional search approach shows a smooth and continuous path. **C**. Search in which areas are prioritized by EEG triage. *Shaded areas* in **B** & **C** represent regions analyzed by the analyst

could be significantly enhanced and accelerated with means to prioritize the search, and help analysts focus their attention on regions with high target probability.

Leveraging the high sensitivity of computer vision with the high specificity of human visual recognition, we have developed a C3Vision Remote Sensing System, based on the architecture of Fig. 9.1A. In this system potential target regions are automatically identified by computer vision and image chips centered on potential targets are generated and then presented as RSVP to the user. Centering the image chips on potential targets improves the detection performance during the triage, as targets are better foveated when presented to the analysts at a rapid pace. The EEG scores computed during the RSVP are used to prioritize which regions of the image should be searched first, leading to search patterns like those shown in Fig. 9.3C.

Using C3Vision in this way improves on the analysts' typical work-flow by offering a first pass in which they can very rapidly review thousands of image chips extracted from the large satellite image and identify those that will be of most interest to them, as shown in Fig. 9.4. They can then move to a more in-depth second pass during which they can review high priority areas first, thus accelerating and managing their search more efficiently.

This architecture combines three major components: 1. computer vision based automated region selection and target detection; 2. real-time recording and decoding of EEG signals and 3. the interface used to exploit the prioritized image analysts.

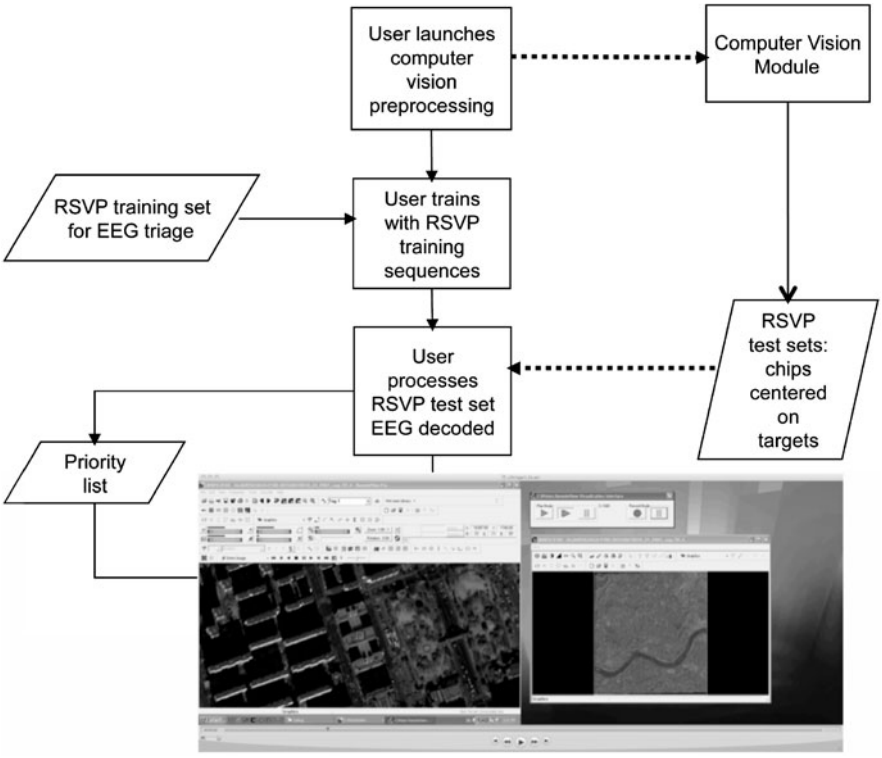


Fig. 9.4 Analyst work-flow with the C3Vision architecture. Potential target regions are automatically identified, and image chips constructed with potential targets centered on those regions. Image chips are triaged using rapid image presentation and classification of neural activity. The results are then reviewed by order of priority in existing specialized software with the help of a dedicated visualization interface

While there is a vast body of computer vision research on object/region detection, the C3Vision architecture itself is agnostic to the choice of a particular method. Such a choice is best guided by the task for which the system is used. The scenario presented here involves targets classes that are known a priori, enabling the use of a model based approach. In particular, we have implemented and tested a framework that extracts low-level features specific to aerial object classes. The framework then infers object classes with a grammar-based reasoning engine that uses domain knowledge and the relationship between object features (see Sajda et al. 2010 for more details). As the image size is typically large and the target detection needs only to be within a few pixels, target detection is only performed on a subsample of image pixels, for example a uniform grid with user specified density. The detection framework associates a confidence score with each pixel in the subsample. Image chips are generated based on those detections with a score exceeding a predefined, task-based threshold.

The image chips are then presented to the analyst as RSVP, using a pace of 5 to 10 Hz (i.e. 200–100 ms per image). While they are presented, EEG signals are recorded using a 64 electrode EEG recording system (ActiveTwo, Biosemi, Germany) in a standard 10–20 montage and at a 2048 Hz sampling rate. Image chips are presented in blocks, typically 100 image chips long. Since detection performance can degrade when target occurrences are too seldom or too frequent, each block is constructed to satisfy a certain target prevalence. In particular, for each block, a number of image chips are randomly drawn from the computer vision list, based on expected true and false positive rates, and additional chips are drawn from a pool of "distractors" in order to achieve the desired block size and prevalence.

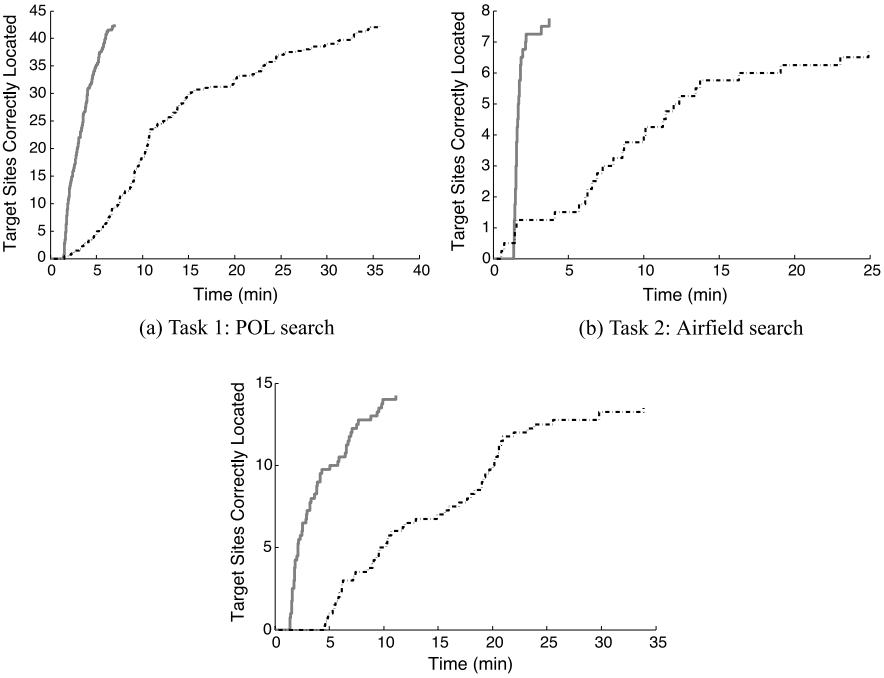
As the EEG stream is acquired, a classifier based on the hierarchical discriminant component analysis, described above, assigns an EEG interest score to each image chip in real time. The EEG classifier is trained at the beginning of a presentation session, with 20 to 30 blocks each containing two known but randomly placed (in the sequence) targets. The content of the training sequences can be related to the content of the test sequences to better familiarize the user to the paradigm and targets. However, from a neuro-physiological perspective, training is not dependent on the choice of imagery, since the classifier is in fact tuned to P300 events. To further help users modulate their responses and obtain better training, feedback can be given at the end of each block, for example by providing a visual indication of how the classifier can re-order the target images within a block based on the EEG interest scores.

The list of prioritized chips is reviewed for validation via a dedicated visualization interface that interacts directly with the analysts' dedicated software. Analysts validate target locations by clicking on corresponding x, y coordinates, which can then be saved in analyst specific products such as shape files. The visualization interface controls the software's viewport, forcing it to show areas centered on the x, y coordinates of the original large image corresponding to the centers of the chips by descending order of EEG interest. Those "jumps" can be triggered by user inputs (e.g. pressing a next button) or be automatically paced. Analysts experimenting with the system have provided positive feedback on both approaches, reporting that the latter helps them rapidly focus their decisions, while the former gives them greater control over the review process.

The architecture has been tested in the course of four semi-operational tests, involving a minimum of 4 image analysts each and imagery from a variety of sensors: EO-gray-scale cameras, EO-color cameras and SAR. Here we show the results of tests where each analyst had to perform three search tasks: 1. look for POLs (Petroleum Oil Lubricant storage); 2. look for airfields in a SAR image; 3. look for buildings in a large EO gray-scale image. For each search task, the times at which image analysts had clicked on a target pixel location was recorded for both baseline and image assisted searches. As a result, several metrics were computed to compare baseline and assisted searches: area throughput at matched sensitivity, i.e. the number of pixels searched per unit time while keeping the baseline and assisted number of targets found the same, detection rate, i.e. the number of targets found over time,

Table 9.1 Throughput comparison between baseline search and C3Vision for the remote sensing application

	Task 1 (POLs—MSI)	Task 2 (Airfields—SAR)	Task 3 (Buildings—EO)
Avg throughput improvement	3.21	11.01	3.16
Standard deviation	0.42	3.48	0.52



(a) Task 1: POL search

(b) Task 2: Airfield search

(c) Task 3: Building search

Fig. 9.5 Average number of target detections as a function of time across subjects and for each task. *Dashed lines* are for baseline and *solid lines* are using C3Vision

and sensitivity. For each task, the system was shown to improve on the baseline area throughput by at least 300% on average (see Table 9.1), as well as on the baseline detection rates (see Fig. 9.5). At the same time, the overall sensitivity and number of false positives were kept the same or moderately improved upon, highlighting the capacity of the system to drastically accelerate search without degrading detection performance.

9.4 C3Vision for Image Retrieval

Due to explosive growth of visual content on the Web, such as personal photographs and video, there is an emerging need for efficient and accurate systems to rapidly analyze visual information. 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. 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 the 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 required manual input.

We consider a C3Vision system for image retrieval using the architecture shown in Fig. 9.1B (and more specifically in Fig. 9.6A). In this architecture, neural signals measured via EEG are used to detect generic objects of interest (OOI) presented in a series of images, while computer vision exploits the EEG labels within the context of a graph-based visual pattern mining algorithm. For the EEG-based OOI detection, only a relatively small subset of images (on the order of few hundred) is first randomly sampled from a larger image database and presented as visual stimuli to the subject. From this window into the larger image collection, the EEG interest detector can identify a small set of highly ranked images to be used as a group of 'pseudo positive' labels for the pattern discovery module. This module then refines and propagates the labels throughout the entire database of images, returning a larger set of images related to those to which the subject showed the greatest interest. In this way, subject participation is minimized, yielding just sufficient information for the neural state decoder and the pattern mining module to effectively infer objects that have attracted a users attention and generate labels for all the images in the collection. Thus, while subjects are only required to review a small subset of the database (avoiding long EEG recording sessions and fatigue), they can still obtain access to a large number of images that interest them.

The imagery used to test the image retrieval architecture was taken from the Caltech 101 database. This database is a well known set of images that are commonly used to evaluate object detection and recognition algorithms (Fei-Fei et al. 2004). It is composed of 101 different image categories, with all the images having been taken from the web. As the categories have large intra class variation and represent a diverse set of image types, while still only consisting of images that have been well-defined, it provides a good testbed for the image retrieval architecture. The Caltech images do vary considerably in both resolution in scale, however. To control for any such fluctuations in image size impacting the subjects' visual responses and fixation capabilities during the RSVP, we selected a subset of categories from the Caltech 101 database to serve as the experimental database. These categories all contained images of similar scale and resolutions, and the images could easily be re-scaled to a consistent size (with negligible distortion) to provide the desired uniformity in the

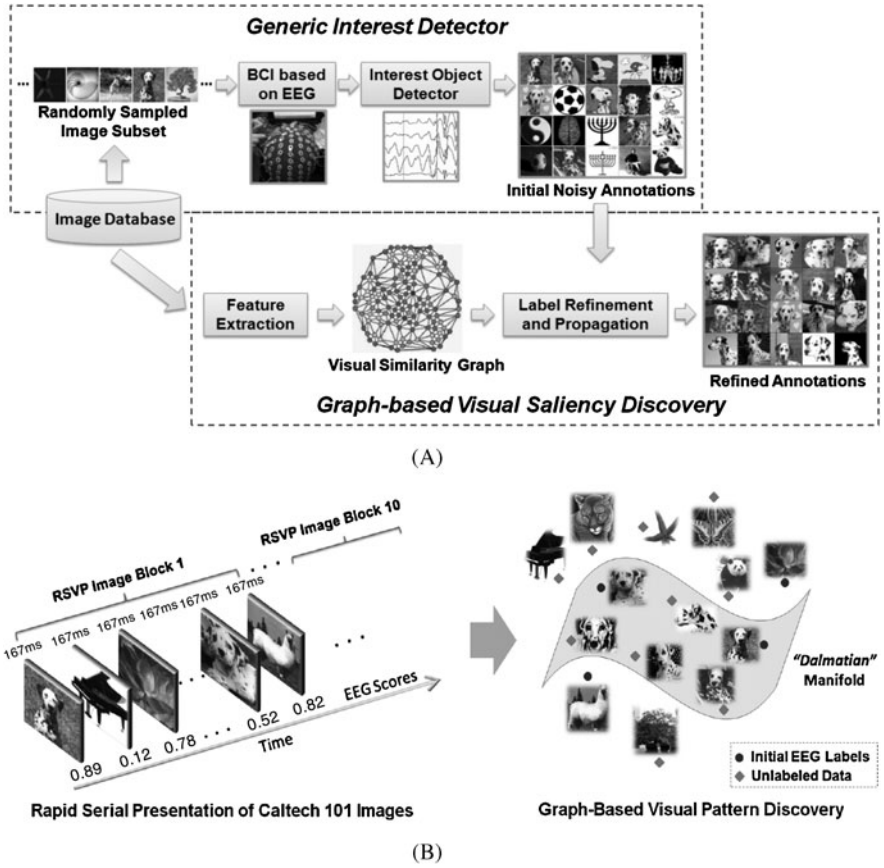


Fig. 9.6 System design for image retrieval. (**A**) Images are sampled from the image database and presented as RSVP to the user while EEG is simultaneously recorded. EEG interest scores are calculated and used to rank the samples based on their score. Concurrently, the entire image database is mapped to a graph structure based on image-based feature similarity. The graph structure is than provided labels derived from the EEG interest scores and these labels are propagated in the graph to re-rank the entire image database. Since the labels derived from EEG interest scores are considered noisy, there is a label refinement/sensitivity analysis step which is used to maximize the value of the labels. (**B**) From the perspective of the graph based model, the interest scores derived from EEG during the RSVP presentation of images can be seen as improving the discovery of manifolds in the feature space of the images. These manifolds represent object categories which are of interest to the user and which are similar in terms of the feature space in which they reside

visual input during the RSVP. The experimental database thus consisted of 62 of the Caltech 101 database categories for a total of 3798 images (42% of the total Caltech 101 images).

Each test sequence involved the users being presented with 1000 images randomly selected from the 62 categories. The images were shown in 10 blocks of 100 images each, with the images within each block being shown at 6 Hz for the RSVP.

During each test sequence, the users were instructed to look for images from one of three target categories: Starfish, Chandeliers, and Dalmatians. The RSVP sequence of 1000 images were then repeated (in a random order) with the participant being instructed to attend to images from the next target category. The ordering of the target categories was varied between subjects. All EEG data from four subjects (who were familiar with EEG work, but who had not been exposed to this experiment), were again recorded during these tests using a 64 channel system (ActiveTwo, Biosemi, Germany) in a standard 10–20 montage, with a sampling rate of 2048 Hz.

The hierarchical discriminate component analysis algorithm was used to create the interest detector component of the image retrieval architecture, see Fig. 9.6A. The format of the training data used to create the detector matched the test data (blocks of 100 images, shown at 6 Hz), with the training images being taken from the Caltech 256 database to differentiate the training and testing data. Similarly to the testing data though, only a subset of relatively similarly sized Caltech 256 images were used for the training database, with several 101 categories that are typically part of the 256 database also having been removed. Typically 25–30 blocks of images (with exactly 2 target images randomly positioned within each block) were presented during the training session, with the subjects being instructed to attend to either soccer balls or baseball gloves as the training target category. The 20 images ranked most highly by the interest detector were then given to the pattern discovery module so that other images similar to those could be identified from the full image database.

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 Fig. 9.6B). 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 (Wang et al. 2008) 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.

Assume that the generic interest detector outputs the EEG score $\mathbf{e} = \{e_1, e_2, \ldots, e_n\}$ from a RSVP sequence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ shown to the subject.[1] Previous work has shown that the existing semi-supervised methods cannot handle cases with extremely noisy labels (Wang et al. 2009a). In order to refine the noisy EEG scores, our method first extracts the salient image pattern and recovers 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]For an RSVP image sequence, the decoded EEG score vector $\mathbf{e} = \{e_1, e_2, \ldots, e_n\}$ is usually normalized as $e_i \in [0, 1], i = 1, \ldots, n$.

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 $\mathbf{e} = \{e_1, e_2, \ldots, e_n\} \rightarrow \mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. 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 ϵ is called interest level for discretizing the EEG scores.[2]
3. Apply the bivariate regularization framework to define the following risk function

$$E_\gamma(\mathbf{f}, \mathbf{y}) = \mathcal{Q}(\mathbf{f}, \mathbf{y}) + \gamma \mathcal{V}_\mathcal{G}(\mathbf{f}) \tag{9.4}$$

   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 $\mathbf{f}^*$

$$\mathbf{f}^* = \arg\min_{\mathbf{f},\mathbf{y}} E_\gamma(\mathbf{f}, \mathbf{y}). \tag{9.5}$$

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 Fig. 9.7B, are expected to be more accurate (in terms of both precision and recall) than the baseline of using EEG based detection alone.

The Caltech 101 image search experiments clearly demonstrated how the C3Vision architecture was able to improve on image identification over chance or even just using EEG detection alone (Wang et al. 2009b). The results were quantified in terms of their average precision (AP), a metric commonly used in information retrieval, and which approximates the area under the precision recall curve (Wang et al. 2009b). For example, the full system achieved 69.1% AP for one subject searching for Dalmatians, as compared to 33.73% when using EEG interest detection alone, and 1.76% for chance. The precision recall curves for this particular case are shown in Fig. 9.7A, with Fig. 9.7B illustrating how the density of target images was increased using the full architecture (bottom panel) versus simply using the EEG scoring (top panel). Overall, the combined EEG-pattern discovery module showed significant improvement in eight of the twelve trials (4 subjects searching for 3 target categories), with AP's in those cases ranging between 25–69% (mean: 42.5%). By comparison, chance levels were 1.76% (Dalmatian), 2.26% (Starfish), 5.11% (Chandelier/Menorah), and the average APs for the EEG detection alone was 15.7%. Furthermore, even in cases where the EEG detection was below 10% AP, the label refinement process was still able to significantly improve the image annotation accuracy.

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

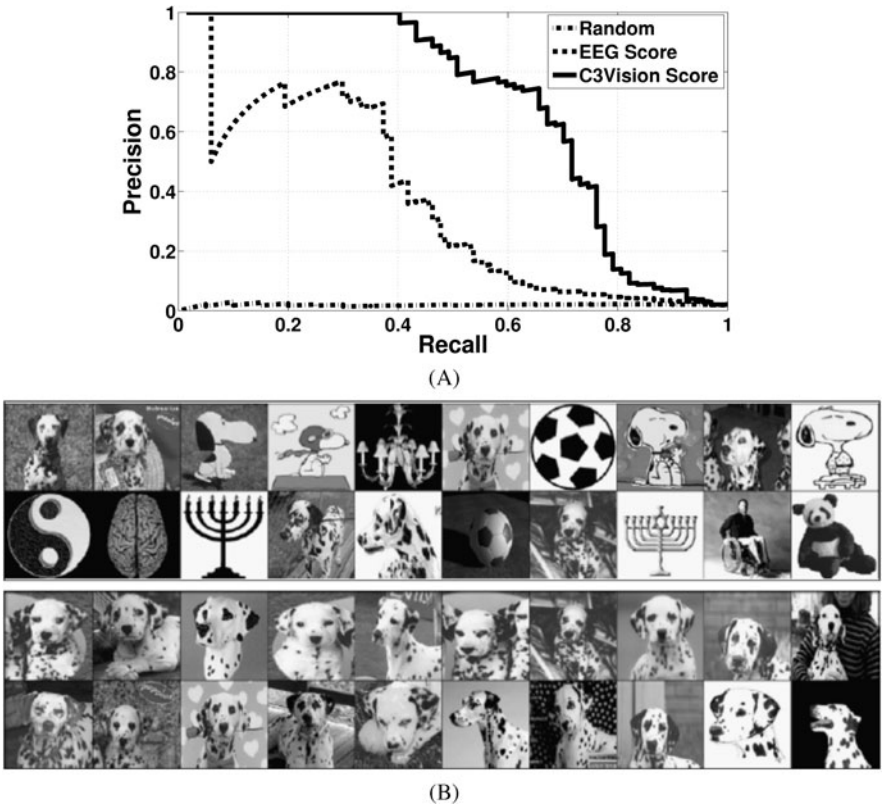


Fig. 9.7 Results for image retrieval for the object class "Dalmatian" in the Caltech 101 database. (**A**) Precision-recall curves for random sampling, retrieval using the EEG interest score alone and the results using EEG + the computer vision based transductive graph (i.e. C3Vision). Note that the C3Vision case results in a $>5\times$ increase in recall while maintaining a 100% precision, over the EEG score ranking alone. (**B**) Top 20 images for one subject, showing (a) ranking by interest scores from EEG detector; (b) ranking by scores after label refinement in transductive graph. Adapted from Wang et al. (2009b)

9.5 Conclusions

The C3Vision framework we describe has potentially many applications in multimedia search and image retrieval. However there are several technical challenges that remain. The results we have described have investigated essentially feedforward one-pass processing, namely there is no feedback between the computer vision system and human (or vice versa). However more recent work by our group has shown that feedback can be used to improve the precision of retrieval, though this comes at the cost of also changing the prevalence of objects of interest in the sample and thus the potential magnitude of the neural target related signal (e.g. P300). More generally, the issue of feedback brings up the interesting problem of co-learning. The human subject, the computer vision system and the EEG decoder can all potentially

adapt in a feedback loop and we are currently investigating co-learning strategies which will improve convergence to high precision recall.

Our approach in developing C3Vision has been to leverage the complementary strengths of rapid, general-purpose scene analysis by humans and the ability of computers to process vast amounts of information. Unique to our approach is that we create an interface between the two vision systems via real-time EEG-based communication channel. A current challenge in BCI system design is that state-of-the-art decoding enables relatively low bit rates—40–60 bits per minute—far below what other communication mediums might offer. For BCI's which focus on assisting those with neurological disease and disability, particularly those that look to assist people that are "locked-in", such a low bandwidth channel is better than no channel at all and thus can substantially improve quality of life. However if BCI systems are going to make an impact in applications in which users are essentially neurologically healthy individuals, then the low bit rate channel of EEG must be exploited with some ingenuity. For example, in our BCI applications, we are looking at ways in which the bits that we can obtain via the EEG channel are very difficult to measure from other channels, for example by monitoring behavior via eye-tracking and/or button/keyboard responses. Future work will continue to investigate approaches that exploit this low bandwidth channel in ways that give us access to information about otherwise latent cognitive states of the user.

Acknowledgements This research was funded by DARPA (contract NBCHC080029). The views, opinions, and/or findings contained in this document are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Aerial images are provided by DigiGlobe.

References

Bigdely-Shamlo N, Vankov A, Ramirez RR, Makeig S (2008) Brain activity-based image classification from rapid serial visual presentation. IEEE Trans Neural Syst Rehabil Eng 16(5):432–441. DOI 10.1109/TNSRE.2008.2003381

Fei-Fei L, Fergus R, Perona R (2004) Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: IEEE CVPR 2004, Workshop on Generative-Model Based Vision

Gerson A, Parra L, Sajda P (2005) Cortical origins of response time variability during rapid discrimination of visual objects. Neuroimage 28(2):342–353

Gerson AD, Parra LC, Sajda P (2006) Cortically-coupled computer vision for rapid image search. IEEE Trans Neural Syst Rehabil Eng 14:174–179

Gladwell M (2005) Blink: The Power of Thinking Without Thinking. Little, Brown and Company: Time Warner Book Group, New York

Kapoor A, Shenoy P, Tan D (2008) Combining brain computer interfaces with vision for object categorization. In: Proc IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp 1–8. DOI 10.1109/CVPR.2008.4587618

Keysers C, Xiao DK, Foldiak P, Perrett D (2001) The speed of sight. J Cogn Neurosci 13(1):90–101

Linden D (2005) The P300: Where in the brain is it produced and what does it tell us? Neuroscientist 11(6):563–576

Makeig S, Westerfield M, Jung TP, Covington J, Townsend J, Sejnowski T, Courchesne E (1999) Independent components of the late positive response complex in a visual spatial attention task. J Neurosci 19:2665–2680

Oliva A (2005) Gist of the scene. In: Encyclopedia of Neurobiology of Attention. Elsevier, San Diego, CA, pp 251–256

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

Parra LC, Spence CD, Gerson AD, Sajda P (2005) Recipes for the linear analysis of EEG. Neuroimage 28(2):326–341

Sajda P, Parra L, Christoforou C, Hanna B, Bahlmann C, Wang J, Pohlmeyer E, Dmochowski J, Chang SF (2010) In a blink of an eye and a switch of a transistor: Cortically-coupled computer vision. Proceedings of the IEEE 98(3):462–478

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

Wang J, Jebara T, Chang SF (2008) Graph transduction via alternating minimization. In: International Conference on Machine Learning (ICML)

Wang J, Jaing YG, Chang SF (2009a) Label diagnosis through self tuning for web image search. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami Beach, Florida, USA

Wang J, Pohlmeyer E, Hanna B, Jiang YG, Sajda P, Chang SF (2009b) Brain state decoding for rapid image retrieval. In: ACM MultiMedia, Beijing, China, pp 945–954