Combining Brain Computer Interfaces with Vision for Object Categorization

Ashish Kapoor
Microsoft Research, Redmond
akapoor@microsoft.com

Pradeep Shenoy
Univ. of Washington, Seattle
pshenoy@cs.washington.edu

Desney Tan
Microsoft Research, Redmond
desney@microsoft.com

Abstract

Human-aided computing proposes using information measured directly from the human brain in order to perform useful tasks. In this paper, we extend this idea by fusing computer vision-based processing and processing done by the human brain in order to build more effective object categorization systems. Specifically, we use an electroencephalograph (EEG) device to measure the subconscious cognitive processing that occurs in the brain as users see images, even when they are not trying to explicitly classify them. We present a novel framework that combines a discriminative visual category recognition system based on the Pyramid Match Kernel (PMK) with information derived from EEG measurements as users view images. We propose a fast convex kernel alignment algorithm to effectively combine the two sources of information. Our approach is validated with experiments using real-world data, where we show significant gains in classification accuracy. We analyze the properties of this information fusion method by examining the relative contributions of the two modalities, the errors arising from each source, and the stability of the combination in repeated experiments.

1. Introduction

Visual category recognition is a challenging problem and techniques based on computer vision often require human involvement to learn good object category models. The most basic level of human involvement is providing labeled data that the system can use to learn visual categories. Since this labeling process is often very expensive, much recent work has focused on ways to reduce the number of labeled examples required to learn accurate models [5, 12, 23]. These systems aim to maximally utilize the human effort involved in labeling examples. Other ingenious solutions for the labeling problem include embedding the labeling task in popular games [39, 40], and asking users to provide finer-grained information by selecting and labeling specific objects within images [1].

This paper explores a new form of human involvement by directly measuring a user's brain signals so as to provide information to the machine with little conscious effort. This approach is built on the realization that the human brain subconsciously processes different images in different ways measurable by certain brain-sensing technologies, even when the user is not trying to categorize images. The advantages of fusing this information with traditional computer vision-based techniques are several-fold. First, by observing how human brain processes help boost traditional vision-based methods we hope to gain insight into aspects of images and categories that are currently unmodeled by computer vision algorithms. This can help us build systems that match the robustness and flexibility of the human visual system. Second, even with explicit human involvement, gathering labels for building visual categorization systems is an expensive process. It is well known that informative brain responses are observed even when images are displayed for only 40ms [15]. By exploiting the implicit processing in the human brain with rapid presentation of images, we can significantly speed up the labeling process and reduce the amount of labeled training data we need to collect. Finally, since computers process images very differently from our brains, the two modalities provide complementary information and should lead to more effective classifiers. Techniques based on computer vision

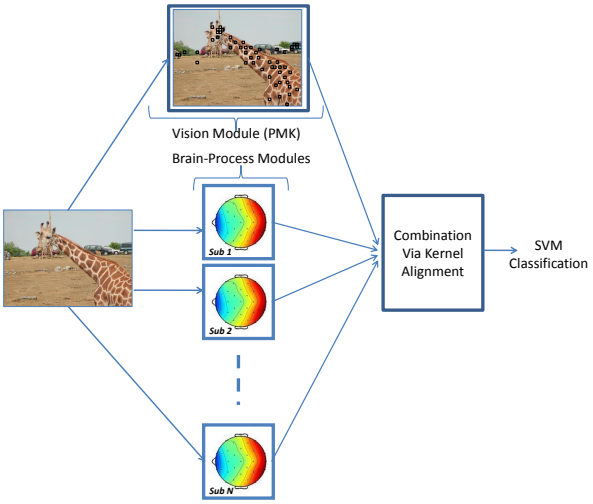

Figure 1. The proposed framework to combine vision computations with human-brain processing for visual category recognition.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

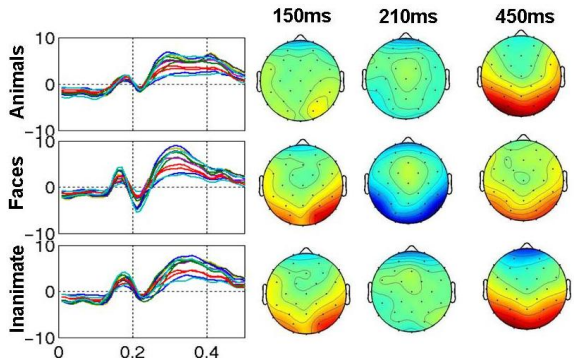

Figure 2. The figure shows the across-subject average EEG response to images containing animals, faces and inanimate objects respectively. The x-axis is time in seconds where each image is presented at time 0. Also shown are the spatial distribution of the signal at three specific time instants for each of the three classes (red corresponds to positive values, and blue corresponds to negative values). The difference in responses is sufficient to build an object categorization system based on EEG.

focus on various imaging transformations and intra-class variations and are often motivated by the specific vision-centric tasks. However, human brain processes tend to be fairly task-invariant and show characteristic responses that are more likely due to contextual or semantic associations.

In this work, we focus on the advantages of effectively combining the two different processing modes to build better visual categorization models. Specifically, the two main contributions of this paper are 1) a system that learns visual categories by combining information from visual image features with the information measured from a human brain processing images and 2) a kernel alignment based fusion scheme that combines the two modalities in a principled and efficient manner. We show, using data from human users, that such a combined system can significantly improve visual category classification.

Figure 1 depicts our overall framework for image categorization. The computer vision component of our system is based on the Pyramid Match Kernel (PMK) [16], a local feature correspondence kernel for object categorization. Object recognition based on local features has been shown to have several important advantages, including invariance to various translational, rotational, affine and photometric transformations and robustness to partial occlusions [27, 28]. The second component of our system is the brain-process module that measures EEG data from single or multiple users. This module complements the visual features with activations in a human brain as images are presented to multiple subjects. Our system combines these two modalities using a fast convex kernel alignment criterion and learns visual category models that are superior to the ones trained on only one of the modules. In the following sections, we present background on object categorization and EEG, describe the technical details of the framework, and then present our validation experiment and results.

2. Background

2.1. Object Categorization with Pyramid Match Kernel

Object category recognition has long been a topic of active interest in computer vision research. Many popular methods are based on local feature descriptors (c.f. [27, 28]) and have been shown to offer invariance across a range of geometric and photometric conditions. Early models captured appearance and shape variation in a generative probabilistic framework [13], but more recent techniques have typically exploited methods based on SVMs or Nearest Neighbor methods [10, 30, 36, 42]. In our work, we adopt Grauman and Darrell's Pyramid Match Kernel [16] and express vision-based similarity between images in terms of partial match correspondences. We chose this method largely for its efficient linear-time approximation of the optimal partial-match correspondence.

Sets of local features provide a useful image representation for object categorization, as they often show tolerance to partial occlusions, object pose variation, and illumination changes. Generally an image is decomposed into local regions or patches, possibly according to an interest operator, and then a local descriptor is extracted to describe the shape or appearance of these patches. The matching or correspondence between two such sets can often reveal their overall similarity and localize highly distinctive object features. Recent research has produced several specialized set-correspondence kernels to exploit this property for object recognition [16, 26, 41, 42].

The Pyramid Match Kernel approximates the partial match similarity between sets of unordered feature vectors. Given a set of feature vectors, $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_{|\mathbf{S}|}\}$ where all $\mathbf{s}_i \in \Re^d$, an L-level multi-resolution histogram $\Psi(\mathbf{S}) = [H_0(\mathbf{S}), \ldots, H_{L-1}(\mathbf{S})]$ is computed. This pyramid bins the features in such a way that an implicit hierarchical matching between $\mathbf{S_1}$ and another set $\mathbf{S_2}$ can be read off in time linear in $\max(|\mathbf{S}_1|, |\mathbf{S}_2|)$. The pyramid match kernel (PMK) value between two input sets $\mathbf{S}_1$ and $\mathbf{S}_2$ is defined as the weighted sum of the number of feature matches found at each level of their pyramids [16]:

$$\mathbf{K}_\Delta\left(\Psi(\mathbf{S}_1), \Psi(\mathbf{S}_2)\right) =$$

$$\sum_{i=0}^{L-1} w_i \Big(\mathcal{I}\left(H_i(\mathbf{S}_1), H_i(\mathbf{S}_2)\right) - \mathcal{I}(H_{i-1}(\mathbf{S}_1), H_{i-1}(\mathbf{S}_2)) \Big)$$

where $\mathcal{I}$ denotes histogram intersection, and the difference in intersections across levels serves to count the number of new matches formed at level i, which were not already counted at any finer resolution level. The weights are set to be inversely proportional to the size of the bins, in order to reflect the maximal distance two matched points could be from one another. As long as $w_i \geq w_{i+1}$, the kernel is Mercer.

The matching is partial in that some features may not have good matches but are not penalized in the matching

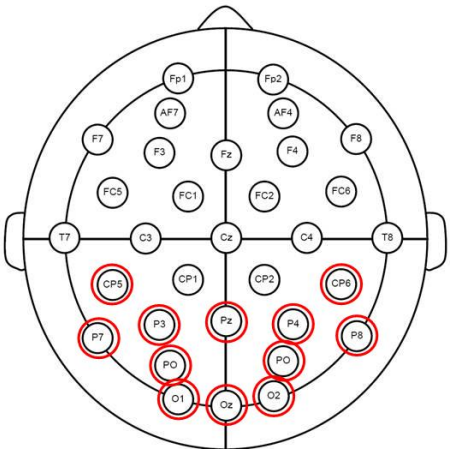

Figure 3. The figure shows a standardized layout for electrode placement in a 32-electrode EEG measurement system [21], pictured from the top, with nose and ears shown for orientation. Also marked in red are the electrodes used for analysis in this paper.

score, and thus some clutter and background features is tolerated. The linear-time PMK offers a computationally appealing alternative to the cubic-time optimal matching. This is useful in our application since densely sampled local features are known to often yield better accuracy on category-level recognition problems [20, 26]. In addition, since PMK is a Mercer kernel, we can train an SVM based on a pool of labeled images using $\mathbf{K}_\Delta$ [16], thus using the unordered sets of interest points in each image to determine visual similarity between images.

2.2. Brain Computer Interface

We use an electroencephalograph (EEG) device to observe cognitive activity as the images are being presented to a human subject. Electroencephalography or EEG, is a neurophysiological measurement of brain activity using electrodes placed on the surface of the scalp (see e.g. [14]). Researchers often examine behavioral correlates in EEG signals by measuring the event-related potential (ERP), which represents the spatiotemporal shape of brain measurements in response to a discrete stimulus. By averaging this response across multiple presentations of stimuli and multiple subjects, researchers can learn about aggregate differences in response between different classes of stimuli [18, 22]. As an example, the presentation of a human face is commonly connected with a pronounced negative drop in signal amplitude in certain channels approximately 170ms following stimulus presentation [33].

For example, Figure 2 depicts brain responses as images with three different labels are shown to human subjects from the data we use in our work. In this figure, responses are averaged across multiple image presentations and each line represents the measurement of one of the channels from the EEG device in the first 0.5s following stimulus presentation. For visualization purposes, scalp maps of the electrode readings for three time points are shown, highlighting the spatiotemporal difference in the EEG response to each category of images. Note that the responses are significantly different for the three categories and there is enough discriminatory signal to train a classifier, indicating the discriminative power that may exist in this signal.

Related to this research is the study of brain-computer interfaces (BCI), which aim to allow users to communicate with the external world using brain signals alone [4]. Many BCIs are based on a "recognition response" called a P300[1] that is evoked by stimuli of interest to the user. By detecting which of a series of stimuli (e.g., images, menu options, letters) generate this response, such systems can decode the user's intent or attention, and establish a communication channel such as a spelling device [11].

Gerson and colleagues [15] exploit this P300 response in their system for "cortically coupled computer vision", in which the user intentionally performs visual search on a sequence of rapidly presented images, looking for a designated target image. The system can detect target images using the brain response alone, in certain cases faster than possible by manual identification using button presses. This system requires the user's explicit intent in searching for a single target or category of targets, and is a "target detector" system, rather than a detector for a specific category of objects. As a result, the study did not use computer vision algorithms to enhance the EEG-based results.

We base our work on that done by Shenoy and Tan [35], who propose a complementary system for "human-aided computing", in which the user is passively viewing images while performing a distracter task that does not consist of explicitly labeling or recognizing the images. The distracter task serves only to capture visual attention and cognitive processing. Their results showed that passive EEG responses can be used to label images with one of 3 category labels, namely human faces, animals, and inanimate objects, with average accuracy of 55.3% using only a single presentation of an image. They further showed that the accuracy could be boosted by using multiple presentations to one or multiple users. With up to 10 presentations, they raised the average labeling accuracy to 73.5%. This system demonstrated that EEG signals could in principle be used as a new modality for extracting features from images for use in an object recognition system. Our work extends this [35] and explores a method for combining the information from EEG responses with the state-of-the-art vision algorithms for object recognition. Our work is significantly different as we focus on the vision algorithms based on correspondence kernels with local features and show that there is a significant gain obtained by incorporating EEG information. This suggests that there exists a set of complementary features in EEG that are not yet captured by vision-based methods.

[1] named for the positive amplitude change seen in certain EEG channels roughly 300ms after stimulus presentation

3. Combining BCI with Visual Features

Much recent research has focused on the general problem of combining information from multiple sources. Many feature fusion methods, including Boosting [34] and Bagging [6], concatenate features extracted from all the modalities to form a single representation, and train a classifier using this joint feature representation. Since the visual category algorithm based on the Pyramid Match Kernel operates at the kernel level where instead of features the Pyramid Match criterion provides us with a similarity ($\mathbf{K}_\Delta$) between any two given images, it is nontrivial to use such feature-fusion methods in our framework.

An alternative is to use decision-level fusion [24], with many possibilities for combining decisions from multiple modalities, including majority vote, sum, product, maximum, and minimum. However, it is difficult to predict which of these fixed rules would perform best. There are also methods that adaptively weigh and fuse the decisions in an expert-critic framework [19, 29, 31, 37]. Unfortunately, these methods require a large amount of training data.

Our solution is to fuse modalities at the kernel level, allowing us to seamlessly combine the visual category recognition algorithms based on local feature correspondence kernels. Specifically, assuming that we have similarities (kernels) from vision features and the EEG responses, our aim is to additively combine the kernel matrices such that the resulting kernel is "ideal" for classification. Our formulation of the kernel combination is a convex program and can naturally handle multiple classes. Kernel alignment techniques have been explored in the past [9, 25] and more recently kernel learning techniques have been used in vision [38]. Our method of kernel combination is most similar to [25] and [38], however, their formulation is a semidefinite program and a second order cone program respectively and it is non-trivial to extend them to the multi-class case besides formulating the classification as either 1-vs-all or a series of pairwise classification formulations.

3.1. Kernel Alignment for Fusion

Given a set of training images and corresponding EEG responses from k different users, we start with kernels that determine the similarity of the images in the visual as well as the EEG signal space. The kernel $\mathbf{K}_\Delta$ that describes the visual similarity between example images is computed via the Pyramid Match as described in section 2.1. Further, given EEG responses from a user i we assume that we can compute the kernel $\mathbf{K}_{\xi_i}$ that depicts similarity in the ERP space (we defer the details of EEG kernels to section 4.1.1).

Given the kernels $\mathbf{K}_\Delta, \mathbf{K}_{\xi_1}, \cdots, \mathbf{K}_{\xi_k}$ we seek a linear combination of these base kernels such that the resulting kernel $\mathbf{K}$ is well-aligned with an ideal kernel $\mathbf{A}$. We define an ideal kernel $\mathbf{A}$ such that the entry $A_{ij} = 1$, if and only if the i^{th} and the j^{th} image have the same visual category label, otherwise $A_{ij} = 0$. This definition is different from the target kernel used for alignment in earlier approaches [9, 25]. However, those approaches focus on binary classification problems and it is non-trivial to optimize kernels simultaneously when the number of classes are more than two. Since the proposed target kernel $\mathbf{A}$ assigns a value of 0 when the examples belong to different classes, it assumes no similarity between them irrespective of their true labels; thus, allowing the measure to be invariant to the number of classes. Formally, we have

$$\mathbf{K} = \alpha_0 \mathbf{K}_\Delta + \sum_{i=1}^{k} \alpha_i \mathbf{K}_{\xi_i} \tag{1}$$

Here, $\boldsymbol{\alpha} = \{\alpha_0, .., \alpha_k\}$ are the parameters that we wish to optimize for. The objective $\mathcal{L}(\boldsymbol{\alpha})$ that we minimize is the squared Frobenius norm of the difference between $\mathbf{K}$ and $\mathbf{A}$:

$$\begin{aligned} &\arg\min_{\boldsymbol{\alpha}} ||\mathbf{K} - \mathbf{A}||_{\mathcal{F}}^2 \\ \text{subject to:} \quad &\alpha_i \geq 0 \text{ for } i \in \{0, .., k\} \end{aligned}$$

The non-negativity constraints on $\boldsymbol{\alpha}$ ensure that the resulting $\mathbf{K}$ is positive-semidefinite and can be used in an SVM formulation (or other kernel-based methods). The proposed objective is a convex function, which can be easily seen by considering $\mathbf{K}$ as a linear combination of vectors constructed by unfolding the basis matrices. With the linear non-negativity constraints, the resulting optimization problem is a convex program and has a unique minimum. Similar criteria has been proposed in the context of Gaussian Processes [32] and Geostatistics [8]. In a manner similar to the alignment measure used by [9, 25], it can be shown the measure defined by the Frobenius norm is also consistent [32].

The proposed convex program can be solved using any gradient-descent based procedure and in our implementation we use a gradient descent procedure based on projected BFGS method that uses a simple line search. The gradients of the objective are simple to compute and can be written as: $\frac{\delta \mathcal{L}(\boldsymbol{\alpha})}{\delta \alpha_i} = 2 \cdot sum(\mathbf{K}_{\xi_i} \circ (\mathbf{K} - \mathbf{A}))$, where $sum(\cdot)$ denotes summation over all the elements of the matrix and the '$\circ$' operator denotes the Hadamard product, which is simply the product of corresponding entries in the matrices. Once the parameters $\boldsymbol{\alpha}$ are found, then the resulting linear combination of kernel ($\mathbf{K}$) can be used in any kernel-based learning procedure.

4. Experiments

We performed experiments with real-world data to (1) show the advantage of the combined approach, (2) analyze strengths and weaknesses of the two modalities and (3) examine the stability of the combined visual categorization system.

4.1. Description of Experimental Data

The EEG data for our experiments are taken from [35]. The EEG signals were originally captured using a Biosemi system [3]) at 2 kHz from 32 channels. In the Biosemi system, users wear a cap of electrodes placed in the 10-20 standard electrode layout [21] (see Figure 3). Electrodes measure the electrical activity on the scalp (typically in the microvolt range) and represent a noisy stream of neural activity occurring in the brain.

The images used in the study were taken both from the Caltech-256 dataset and from the web. For the Animals class, random images from multiple categories of the Caltech 256 dataset were chosen. For the Inanimate and Face classes, the authors [35] used keyword search on the web using the keywords "Face" and "Object". They then had independent people rank the collected images according to relevance to the particular category, and used the top ranked images as stimuli for these classes. EEG responses were recorded from 14 users as they viewed the animal, face and inanimate images while performing a "distracter task," which consisted of counting images that contained butterflies in them. Users were not told of the classification task and were not explicitly trying to perform classification.

The data set consisted of two groups of images drawn from the three categories. The first group (*group-1*) consisted of 60 images per class shown to each of the subjects only once, whereas the second group (*group-2*) consisted of 20 images per class presented 10 times each to the subject in a block randomized fashion.

4.1.1 Kernel Computation

Computing Pyramid Match Kernel: For experiments described in this paper, we used the *libpmk* package[2] that used SIFT descriptors extracted at salient points in the image, where each descriptor was concatenated with the normalized image position. For computing PMK values, we used data-dependent partitions [17]. The SIFT features were clustered to create a vocabulary tree of depth 4 and branch factor 10. Using this tree, we built pyramids for each feature set, and computed the match kernel between each pair of images.

EEG Measurement and Processing: We down-sampled the data to 100Hz and filtered it using a butterworth filter in the range 0.5-30Hz. We restricted the data to include only the time window 100-500ms following stimulus presentation. These processing steps are identical to those used in [35] and are typical of EEG studies in the literature. Also, in our analysis we used only data from 12 electrodes of interest (CP5, CP6, P3, Pz, P4, P7, P8, PO3, PO4, O1, O2, Oz), the channels expected to most closely measure human visual processing activity. We concatenated the chosen time window of measurements for the channels of interest to form a single vector representing the "EEG feature" for the image. We converted these responses into a similarity measure (a kernel) by using a gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\beta||\mathbf{x}_i - \mathbf{x}_j||^2)$, where the scaling factor $\beta = 10^{-5}$ was chosen empirically and kept constant across all experiments and subjects.

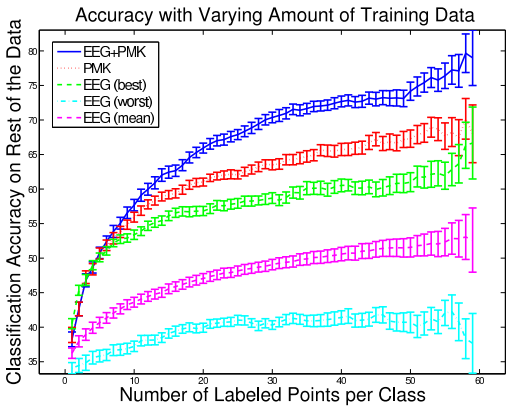

Figure 4. Performance of different modalities on the held-out set as the number of labeled examples are varied. Non-overlapping error bars, which are standard error scaled by 1.64, signify 95% confidence in performance difference. The combined classification significantly outperforms the individual modalities.

4.2. Results

All the experiments in this paper were carried out using the *libsvm* package [7]. Also, we fix $C = 10^6$ which worked well; we experimented with other values but found that classification with SVM was fairly insensitive to the choice of C.

Benefits of the combination: First we examine the gains obtained by combining EEG signals with PMK. For this experiment, we follow the standard testing protocol in object recognition, where a given number (say 25) of training images are taken from each class at random, and the rest of the data is used for testing. The mean recognition rate is used as a metric of performance. We repeated this process 100 times on the group-1 images. Figure 4 shows the mean performance for different modalities and the combination, along with the performance of the best and worst of the EEG users. The errorbars correspond to standard error scaled by 1.64 and non-overlapping errorbars signify 95% confidence in performance difference. We see that significant gains are obtained by combining BCI with vision features. Although the vision features consistently outperform the EEG features, the combination performs better than both, suggesting complementary properties between the modalities.

Contribution of the modalities: Next, we look at the discriminative power of the different modalities. Specifically, we are interested in the relative weight α_0 of the Pyramid Match Kernel in the kernel combination, characterized as:

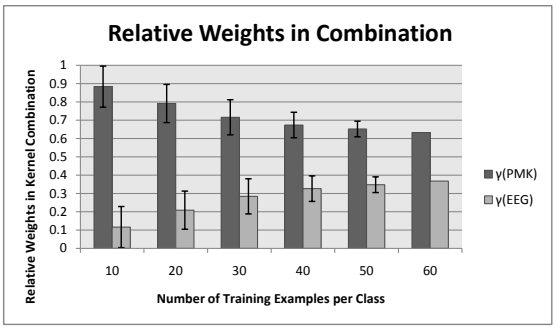

Figure 5. Comparison of relatives weights of the different modalities as we vary the number of labeled examples per class.

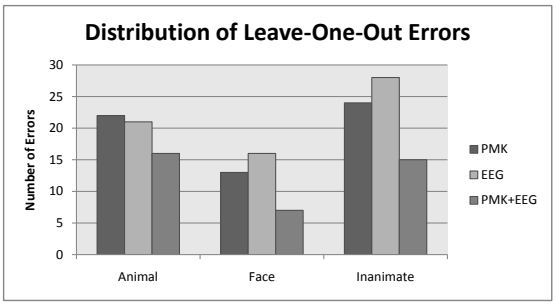

Figure 6. Comparison of relatives weights of the different modalities as we vary the number of labeled examples per class.

$\gamma(\text{PMK}) = \frac{\alpha_0}{\alpha_0 + \sum_{i=1}^{k} \alpha_i}$. Similarly, the relative weight for EEG can be written as: $\gamma(\text{EEG}) = 1 - \gamma(\text{PMK})$. By looking at the statistics of these quantities we can form estimates about the relative contribution of each modality. Figure 5 shows these relatives weights averaged over 100 different runs (error bars are standard deviation) on group-1 for various amounts of training data. We can see that the vision modality has higher discriminative power, however, note that the weight of the EEG modality is highly significant and leads to significant gains in accuracy as shown above. Further, the relative contribution of EEG signal increases with data suggesting that better gains can be expected with more training data.

Analysis of the errors: We also look at the distribution of errors being made by the different channels and their combinations. To this end, we do a leave-one-out analysis on group-1 images where we train the system on all the images except a held-out test image. This process is repeated for all the images in the data and we log the errors that were made by the system when classifying the test image. We do this analysis for the combined system, the PMK-only system and a system that uses only EEG data from all the users. Figure 6 shows the distribution of errors, where we see that the combined system consistently reduces the number of errors being made across all the classes. Interestingly when compared to EEG, the PMK modality performs better on the face and the inanimate categories, while performing slightly worse on the animal one. This might be due to the fact that the face category has enough visual information for a vision-only classifier to work well. The animal images on the other hand usually have inanimate objects as well in the background that might confuse a vision-only system. The EEG modality on the other hand is more influenced by the semantic association of the image; hence, can distinguish classes even when the low-level visual features are confusing. Although this observation is made on a small data set and needs further study, we see that the combination does indeed benefit from individual strengths of both modalities to increase the recognition performance on all the classes.

Stability of the system: Since the human brain concurrently handles multiple tasks and may show significant "background activity", we expect significant variations in measured EEG responses, and thus variations in recognition accuracy. We explored the stability of the combined system in the face of these variations. Here we look at classification results on group-2 images that were presented to the subjects 10 times each. We trained the classification system on the 180 images from group-1 and tested the classification performance with each round of presentation to all the users. We found that in terms of behavior the classification performance was similar for all the runs, with the combined system outperforming the individual modalities 9 out of the 10 times. Figure 7 shows the performance of different modalities for the very first presentation, where we see that the performance is similar to the one obtained with group-1 images. We obtained similar curves for the rest of the presentations as well, which we do not reproduce due to the space constraints. However, note that we can further boost the classification accuracy by voting (EEG+PMK voting) among the presentations and choosing the class label for a test image based on classification agreement across different presentations. This agrees with the phenomenon observed in earlier studies based only on EEG [35], where it is reported that multiple presentations improve accuracy.

Sensitivity to the number of subjects: We also examined how many human-brains are required to get a significant boost and how the performance scales as we increase the number of subjects. Again, we used group-2 as the test set for the classifiers trained with the images in group-1. Figure 8 shows the performance as we vary the number of users, and compares it with the baseline performance of the PMK classifier. Each point was generated by averaging over 15 runs, with subjects randomly selected from the group. As before, the error bars signify 95% confidence intervals. We can see that a significant boost is achieved with as few as 2 users and the performance increases almost linearly as EEG

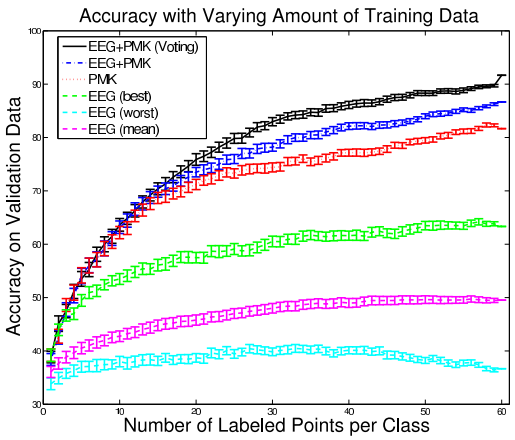

Figure 7. Performance of different modalities on the validation set as the number of labeled examples are varied for a single presentation to the subject. The combined classification based on EEG and PMK significantly outperform the individual modalities. Presenting the same image multiple times to the subject and voting among those classification outcomes further improves the accuracy.

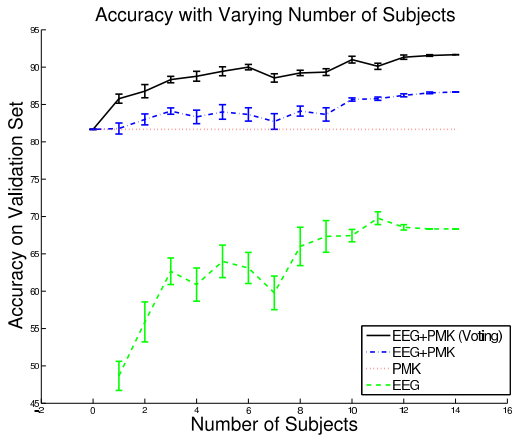

Figure 8. Performance of different modalities on the validation set as the number of subjects are varied. The EEG response helps boost the classification accuracy even with a low number of users. Presenting images multiple times further improves accuracy.

data from more subjects is incorporated. Note that the EEG and EEG+PMK curves are shown only for a single presentation to each user. If we consider all the 10 presentations and vote among the classifications as we did before, then the performance further improves (EEG+PMK voting). We can see that there is a significant gain for EEG+PMK voting even with a single user.

Table 1 summarizes the accuracies obtained on the group-2 images (test set) obtained by classifiers trained on group-1 images. The combination with single presentation outperforms each individual channel with an accuracy of 86.67%. Further improvement is achieved by presenting images to the subjects 10 times and then voting among

Method	Accuracy
PMK	81.67%
EEG	68.33%
PMK + EEG	86.67%
PMK + EEG (Voting)	91.67%

Table 1. Classification accuracies on the validation when trained with 60 labeled examples per class.

	Recognized Class		
True Class	**Animals**	**Faces**	**Inanimate**
Animals	19	0	1
Faces	1	19	0
Inanimate	1	2	17

Table 2. Confusion matrix obtained on the validation data with the combination of EEG and PMK. (60 labeled images per class, mean accuracy over all the classes = 91.67%)

the outcomes of 10 combined classifiers. Figure 9 analyzes different kinds of errors that were made by the different modalities. All of the errors, except the ones highlighted by the thick red double-lined border were corrected in the combined system. Further, there were not any additional errors made in the fused system. Interestingly, the chimps were misclassified as faces by the EEG modality. This is not surprising as objects that look similar to faces are known to elicit "face-like" responses in specific areas of brain. Table 2 shows the confusion matrix obtained by the EEG+PMK voting strategy on the group-2 images, which shows a tremendous boost over the vision only classifier with an accuracy of 91.67%.

5. Conclusion and Future Work

We have presented a framework for combining computer-vision algorithms with brain-computer interfaces for the purpose of visual category recognition. Our SVM based discriminative framework combines correspondence-based notion of similarity between sets of local image features with EEG data using a fast convex kernel alignment criterion. The EEG data we used was collected from users who were not even explicitly trying to perform image classification. Our empirical results demonstrate that such a combination between vision and the human-brain processing can yield significant gains in accuracy for the task of object categorization. This suggests a set of complementary properties between local feature based vision algorithms and the way a human-brain processes the image.

As future work, we plan to extend the framework to incorporate other local feature based kernels and explore alternate kernel combination techniques (such as [38]). By incorporating other vision algorithms we should be able to further improve classification performance. We also plan to explore the possibility of using the EEG signal as a weak label and incorporating those weak labels in active and semi-supervised learning formulations. Finally, we also aim to extend and test this system on larger, more varied data sets.

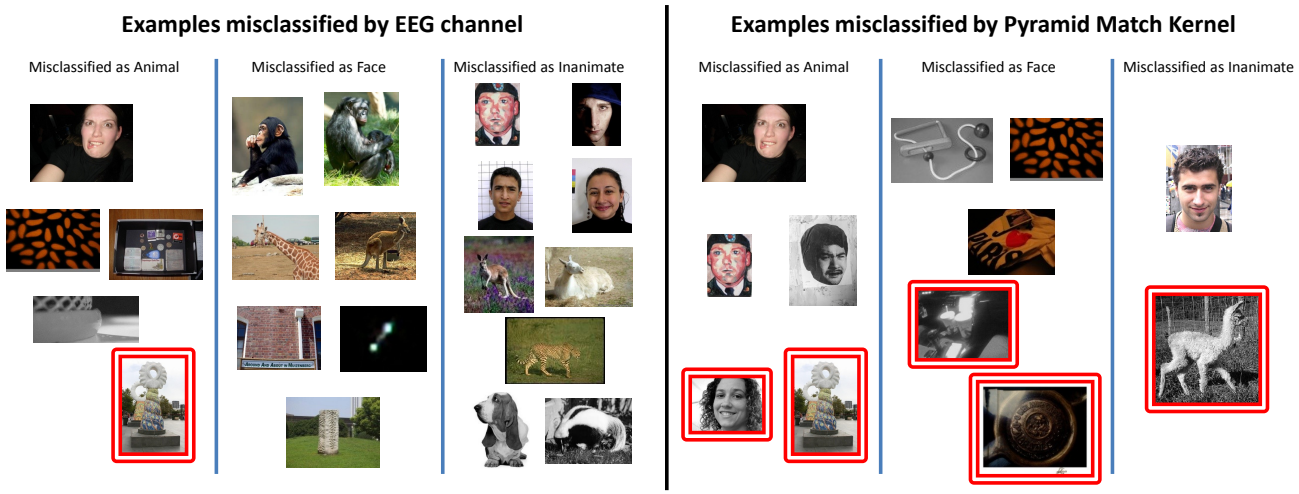

Figure 9. The classification errors made by the classifier based on EEG and the vision features (PMK). All of these errors were corrected by the combination (EEG+PMK) except for the ones that have the red double-lined bounding box around them.

References

[1] *http://labelme.csail.mit.edu/.*

[2] *http//people.csail.mit.edu/jjl/libpmk/.*

[3] *http://www.biosemi.com/.*

[4] Special issue on brain-computer interface technology: The third international meeting. *IEEE Transactions on Neural Systen Rehabilitation Engineering*, 2006.

[5] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 26(2), 1996.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[8] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[9] N. Cristianini, J. Shawe-Taylor, and A. Elisseeff. On kernel-target alignment. In *NIPS*, 2001.

[10] B. T. F. Moosmann and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, 2007.

[11] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography & Clinical Neurophysiology*, 70(6):510–23, 1988.

[12] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian Approach Tested on 101 Object Cateories. In *Workshop on Generative Model Based Vision*, 2004.

[13] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[14] B. Fisch. *Fisch & Spehlmann's EEG primer: Basic principles of digital and analog EEG*. Elsevier: Amsterdam, 2005.

[15] A. Gerson, L. Parra, and P. Sajda. Cortically-coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):174–179, 2006.

[16] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[17] K. Grauman and T. Darrell. Approximate correspondences in high dimensions. In *NIPS*, 2007.

[18] K. Grill-Spector. The neural basis of object perception. *Current opinion in neurobiology*, 13:1–8, 2003.

[19] Y. Ivanov, T. Serre, and J. Bouvrie. Confidence weighted classifier combination for multi-modal human identification. Technical Report AI Memo 2005-035, MIT Computer Science and Artificial Intelligence Laboratory, 2005.

[20] Z. J., M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: A comprehensive study. *IJCV*, 2006.

[21] H. Jasper. The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–375, 1958.

[22] J. S. Johnson and B. A. Olshausen. The earliest EEG signatures of object recognition in a cued-target task are postsensory. *Journal of Vision*, 5(4):299–312, 2005.

[23] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with Gaussian Processes for object categorization. In *ICCV*, 2007.

[24] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[25] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 2004.

[26] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[27] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[28] K. Mikolajczyk and C. Schmid. Indexing Based on Scale Invariant Interest Points. In *ICCV*, 2001.

[29] D. J. Miller and L. Yan. Critic-driven ensemble classification. *Signal Processing*, 47(10), 1999.

[30] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *CVPR*, 2006.

[31] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. In *ICMI*, 2002.

[32] J. C. Platt, C. J. C. Burges, S. Swenson, C.Weare, and A. Zheng. Learning a Gaussian Process prior for automatically generating music playlists. In *NIPS*, 2002.

[33] B. Roisson, I. Gauthier, J.-F. Delvenne, M. Tarr, R. Bruyer, and M. Crommelinck. Does the N170 occipito-temporal component reflect a face-specific structural encoding stage? In *Object Perception and Memory 1999*, 1999.

[34] R. Schapire. A brief introduction to boosting. In *Proceedings of International Conference on Algorithmic Learning Theory*, 1999.

[35] P. Shenoy and D. Tan. Human-aided computing: Utilizing implicit human processing to classify images. In *ACM CHI*, 2008.

[36] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, 2003.

[37] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *ACCV*, 2000.

[38] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[39] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004.

[40] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *ACM CHI*, 2006.

[41] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, 2003.

[42] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.