# Cognitive Vision: Examining Attention, Engagement and Cognitive load via Gaze and EEG

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in*
*Computer Science and Engineering*
*by Research*

*by*

*Viral Parekh*
*201507535*
parekh.viral@research.iiit.ac.in

*International Institute of Information Technology*
*Hyderabad - 500 032, INDIA*
*July 2018*

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled "Cognitive Vision: Examining Attention, Engagement and Cognitive load via Gaze and EEG" by Viral Parekh, has been carried out under our supervision and is not submitted elsewhere for a degree.

_____ _____

Date

Adviser: Prof. C. V. Jawahar
Center for Visual Information Technology,
Kohli Center on Intelligent Systems,
IIIT Hyderabad

_____ _____

Date

Adviser: Dr. Ramanathan Subramanian
University of Glasgow Singapore,
Singapore

**To**

*My Parents and My Sister*

*for their unconditional love and support*

# Acknowledgments

As I submit my thesis, I would like to thank all the people who have supported me and motivated me throughout my journey at IIIT.

I would like to express my sincere gratitude to my supervisors Prof. C. V. Jawahar and Dr. Ramanathan Subramanian for their continuous guidance, and encouragement at every phase of the projects. Working with both of them not only helped with my thesis but I also learned the importance of dedication, consistency, and patience required in the field of research.

I am grateful to Dr. Avinash Sharma who motivated me to join MS by research program at IIIT Hyderabad. I would like to thank Dr. Dipanjan for his guidance in my very first research project. I thank Dr. Vishal Garg for allowing us to use the space in building science department to conduct our experiments with EEG device.

I would also like to thank my colleagues Sourabh, Praveen, Pritish, Aditya, Harish, Govinda and Maneesh for their constant support and for providing a positive learning environment. It was a good learning opportunity to work as a sysadmin with Sourabh. Working with Maneesh on a couple of projects was a great experience. I thank him for his help and suggestions in my research work. Special thanks to Siva, Silar, Varun, and Rajan for all of their help on various occasions. I profusely thank my friends Tejas, Utsav, Abhishek, Shashank, and Soham who made my journey at IIIT memorable.

Last but not the least, I express my deepest gratitude toward my parents, Maheshkumar Parekh and Meenaben Parekh and my elder sister, Khyati Nandha, for their unconditional support in all my endeavors.

# Abstract

Gaze and visual attention are very related. Analysis of gaze and attention can be used for behavior analysis, anticipation or to predict the engagement level of a person. Collectively all these problems fall into the space of cognitive vision systems. As the name suggests it is the intersection of two areas: computer vision and cognition. The goal of the cognitive vision system is to understand the principles of human vision and use them as inspiration to improve machine vision systems.

In this thesis, we have focused on Eye gaze and Electroencephalogram (EEG) data to understand and analyze the attention, cognitive workload and demonstrated a few applications like engagement analysis and image annotation.

With the presence of ubiquitous devices in our daily lives, effectively capturing and managing user attention becomes a critical device requirement. Gaze-lock detection to sense eye-contact with a device is a useful technique to track user's interaction with the device. We propose an eye contact detection using a convolutional neural network (CNN) architecture, which achieves superior eye-contact detection performance as compared to state of the art methods with minimal data pre-processing; our algorithm is furthermore validated on multiple datasets, Gaze-lock detection is improved by combining head pose and eye-gaze information consistent with social attention literature.

Further, we extend our work to analyze the engagement level in the person with dementia via visual attention. Engagement in dementia is typically measured using behavior observational scales (BOS) that are tedious and involve intensive manual labor to annotate, and are therefore not easily scalable. We propose AVEID, a low-cost and easy to use video-based engagement measurement tool to determine the level of engagement of a person with dementia (PwD) when interacting with a target object. We show that the objective behavioral measures computed via AVEID correlate well with subjective expert impressions for the popular MPES and OME BOS, confirming its viability and effectiveness. Moreover, AVEID measures can be obtained for a variety of engagement designs, thereby facilitating large-scale studies with PwD populations.

Analysis of Cognitive load for a given user interface is an important measure of effectiveness or usability. We examine whether EEG-based cognitive load estimation is generalizable across the character, spatial pattern, bar graph and pie chart-based visualizations for the n-back task. Cognitive load is estimated via two recent approaches: (1) Deep convolutional neural network and Proximal support vector machines. Experiments reveal that cognitive load estimation suffers across visualizations suggesting

that (a) they may inherently induce varied cognitive processes in users, and (b) effective adaptation techniques are needed to benchmark visual interfaces for usability given pre-defined tasks.

Finally, the success of deep learning in computer vision has greatly increased the need for annotated image datasets. We propose an EEG (Electroencephalogram)-based image annotation system. While humans can recognize objects in 20-200 milliseconds, the need to manually label images results in a low annotation throughput. Our system employs brain signals captured via a consumer EEG device to achieve an annotation rate of up to 10 images per second. We exploit the P300 event-related potential (ERP) signature to identify target images during a rapid serial visual presentation (RSVP) task. We further perform unsupervised outlier removal to achieve an F1-score of 0.88 on the test set. The proposed system does not depend on category-specific EEG signatures enabling the annotation of any new image category without any model pre-training.

# Contents

# List of Figures

# List of Tables

*Chapter 1*

# Introduction

The cognitive computer vision (CCV) is the understanding of the principles of human vision and use them as inspiration to improve machine vision systems. More formal definition of CCV is given by ECVision [4] as follows,
"Cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events & structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment."

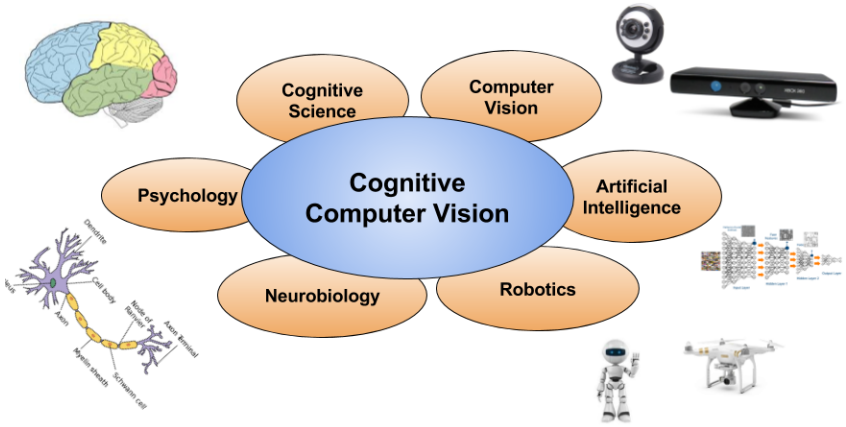


**Figure 1.1** Cognitive Computer Vision is the intersection of several domains like Computer Vision, Cognitive Science, Psychology, Artificial Intelligence etc.

## 1.1 Problems of interest

As depicted in figure 1.1, CCV stimulates a wide spectrum of associations like Psychology, Cognitive Science, Robotics, Computer Vision etc. There are vast number of open problems associated with each

of the domains. Here in this thesis we have selected a few challenges and tried to solved the problem and proposed some real world applications emerged from the field of CCV.

We have restricted our work in the area of attention and engagement analysis. However the applications of these areas could be useful to solve problems in other domains as well. In the following sections we have provided details about the problems of our interest.

## 1.2 Gaze tracking and Gaze locking

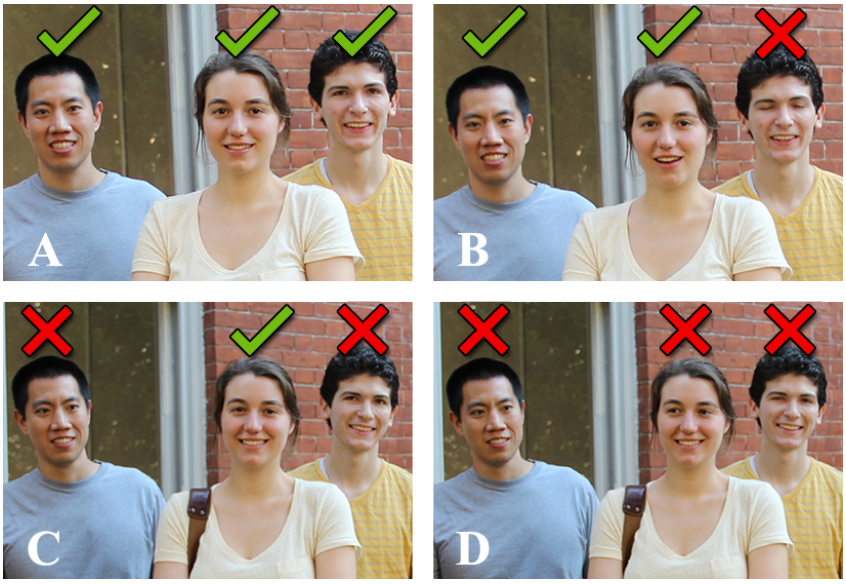


**Figure 1.2** Representation of a Gaze locking (binary classification) problem. In (A) everyone is looking at the camera. In (B) 3rd person is not looking at the camera. (All of these figures are taken from [85])

Gaze tracking has application in many areas like human-computer interaction, medical diagnosis, and psychological studies. Gaze is an indicator of visual attention. There are many works related to gaze estimation and gaze tracking. This survey [32] conveys most of the attempts and contributions in the field for the past 30 years. Gaze estimation methods are either model-based or appearance-based. Some of the model-based methods uses external light sources to detect eye features [103, 107, 108] and other methods uses pupil centers and iris edges to infer the gaze direction [15, 33, 39, 96]. These model based methods do not perform well with low image quality and variable lightning conditions. Appearance-based approaches uses eye images as input [58, 94, 95]. Thus such methods perform well even with low resolution images and produces good results with large datasets [51, 106]. Most of the available solutions require active infrared illumination [63, 66] or intrusive equipment [9, 56] and work only for short distance [10]. Commercial eye tracking devices like Tobii X2-60 are expensive. In [85] the method for gaze locking problem produces good results but that involves feature extraction,

geometric rectification and dimensionality reduction step like PCA. Krafka *et al.* [51], have proposed convolutional neural network architecture for eye tracking in mobile devices. Zhang *et al.* [106] have proposed a deep learning model for gaze estimation. They have used images of one eye and head pose vector.

Gaze locking is a sub-problem of *gaze-tracking*, where the objective is to determine where the user is looking. Gaze-tracking techniques (with the exception of few such as [51]) have inferred the point-of-gaze using eye-based cues even though social attention literature has identified that other cues such as head orientation contribute significantly to this end [53]. Gaze tracking is a regression problem while gaze locking is binary classification problem.

## 1.3 Image annotation

Image annotation is a critical task in computer vision, intended to bridge the ***semantic gap*** between automated and human understanding via the use of tags and labels. Image annotation is useful for building large-scale retrieval systems, organizing and managing multimedia databases, and for training deep learning models for scene understanding. A trivial way to annotate images is to tag them manually with the relevant labels, but this approach is slow and tedious for huge databases. Therefore, many efforts have been undertaken to address/circumvent this problem. Some methods are completely automatic [23, 28, 98, 102, 105], while some others are interactive [6, 43, 87, 93, 100]– these approaches have considerably reduced the human effort required for annotation.

## 1.4 Cognitive Workload

Working memory(WM) and cognitive load are the linked concepts. Working memory, the part of our brain that consciously processes information, is the mental sticky note we use to keep track of information until we need to use it. Unlike long-term memory(LTM), working memory has highly limited capacity to hold information for long duration, and quite sensitive to overload. Cognitive load is correlated with the number of statements in working memory. As human short-term memory is severely limited, any problem that requires a large number of items to be stored in short-term memory may contribute to an excessive cognitive load [92].

Direct and indirect methods have been employed to measure workload, and these can be classified under four main groups. First the subjective assessment methods, that use self-reported rating scores. The second is behavior and performance measures, of the participant, recorded during the task, to identify any workload effects within them. Accuracy and speed of response are widely in use. Next is physiological measures, such as eye movements [2], eye blinks [97], pupil dilation [41], skin temperature , galvanic skin response , heart rate , blood pressure, and respiration rate. The final approach is to adopt a more direct perspective and monitor cognitive activity-related signals directly from the brain. Physiological measures can provide a continuous record of workload over time and their measurement

does not interfere with primary task performance, Hence, the physiological measures may be more useful to asses human mental workload.

## 1.5 Major Contributions

The main contribution of our thesis are

- **Problem:** Eye contact detection
  For this problem we have proposed a deep neural network architecture that requires minimal prepossessing of the input images. Further we have used transfer leaning method to overcome the limitation of annotated data.

- **Problem:** Engagement analysis for PwD
  We have proposed an automatic video system that analyze the engagement of the dementia patients using gaze and emotion analysis.

- **Problem:** Fast image annotation using EEG
  In this work we exploited P300 event related potential in EEG data to provide BCI based fast image annotation solution.

- **Problem:** Cognitive workload estimation using EEG
  This work investigates the suitability of a single EEG framework to assess (extraneous) memory workload across multiple visual interfaces under similar task difficulty. If cognitive load estimation (CLE) is generalizable, it would naturally facilitate usability evaluation of Visualization User Interfaces (Viz UIs). We designed the experiment protocol and collected EEG data from 20 participants to evaluate the results.

## 1.6 Thesis Outline

The chapters in this thesis are organized as follows. First we have discussed our contribution in the domain of gaze analysis and its applications in engagement studies. Then we have discussed the cognitive workload and attention analysis using EEG data.

**Chapter 2:** Eye Contact Detection

**Chapter 3:** AVEID: Automatic Video System for Measuring Engagement In Dementia

**Chapter 4:** An EEG Based Image Annotation System

**Chapter 5:** EEG-based Cognitive Load Estimation Across Visualizations

*Chapter 2*

# Eye Contact Detection

The importance of ***eye-contact*** in non-verbal human communication cannot be understated. Right from infanthood, humans use eye-contact as a means for attracting and acknowledging attention, and can effortlessly sense others' eye-gaze direction [31]. In today's ubiquitous computing environment, it becomes critical for devices to effectively attract and manage users' attention for proactive communication and information rendering. Therefore, HCI would greatly benefit from devices that can sense user attention via eye-contact– a phenomenon termed ***gaze locking*** in [85].

This paper proposes gaze-locking detection using deep convolutional neural networks (CNNs), which have recently become popular for solving visual recognition problems as they obviate the need for hand-crafted features (*e.g.*, expressly modeling head pose). Specifically, our work makes the following research contributions:

(1) Even though the gaze-locking methodology outlined in [85] detects eye-contact from distant faces, it requires an elaborate processing pipeline which includes: eye region rectification for head pose compensation, eye mask extraction, compression of a high-dimensional eye appearance feature vector via dimensionality reduction and a classifier for gaze-lock detection. Differently, we leverage the learning power of CNNs for gaze-locking with minimal data pre-processing. We validate our model on three datasets, and obtain over 90% detection accuracy on the Columbia Gaze (CG) [85] test set. In comparison, [85] reports 92% accuracy on the CG *training set*.

(2) Different from [85] and most gaze-tracking methods, we use facial appearance, which implicitly conveys face pose, in addition to eye appearance. As seen in Fig. 2.1, face orientation crucially determines if the user is gaze-locked with a (reference) camera or not. The eyes in the left and right images have very similar appearance; however, eye-contact is clearly made only in the right instance when one infers gazing direction as the eye orientation *relative* to head pose. Combining face and eye cues achieves superior gaze locking than either of the two as demonstrated in prior works [88].

(3) CNNs are usually implemented on CPU/GPU clusters given their huge computation and memory requirements; their implementation on mobile platforms is precluded by the limited computation and energy resources in these environments. We demonstrate gaze-locking on an Android mobile platform via CNN compression using ideas from the *dark knowledge* concept [36].

**Figure 2.1** Left image is ***non-gaze-locked***, while right image is ***gaze-locked***. Their eye crops however look very similar.

## 2.1 Methodology

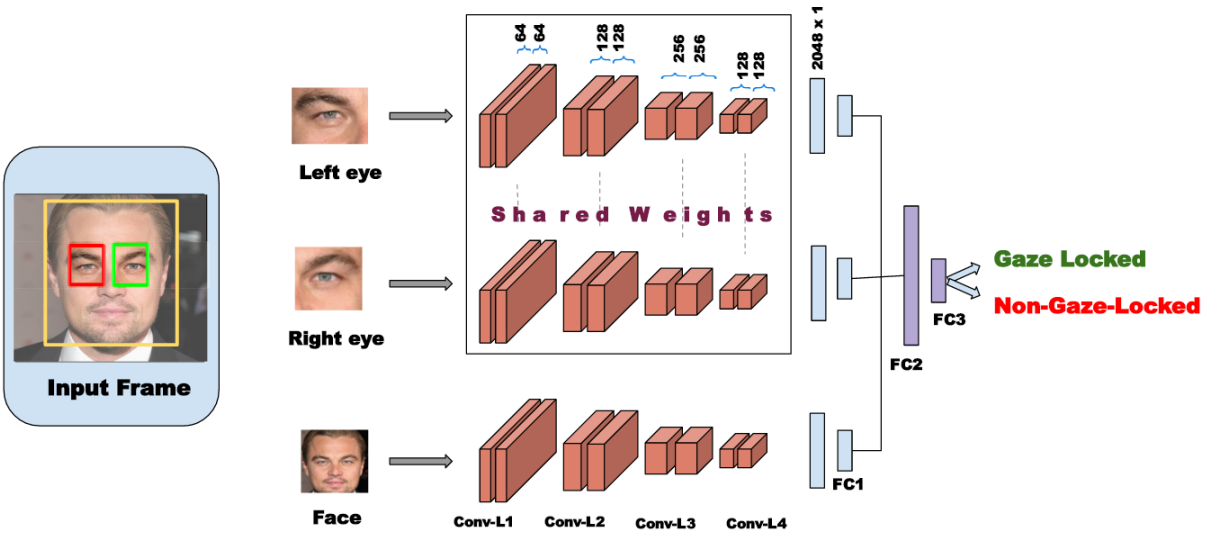


**Figure 2.2** Overview of our ***gaze-lock detector***. Inputs include 64 × 64 *left eye*, *right eye* and *face images*, and the detector outputs a binary label assigned as either *gaze-locked* or *non-gaze-locked*. CNN architecture has ***three*** parallel networks each comprising ***four*** convolutional layer blocks (denoted as filter size/number of filters): CONV-L1: 3 × 3/64, CONV-L2: 3 × 3/128, CONV-L3: 3 × 3/256, and CONV-L4: 3 × 3/128, and ***three*** fully-connected layers denoted as FC1 (of size 2048 inputs × 128 outputs), FC2: 384 × 128 and FC3: 128 × 2.

Fig. 2.2 presents our proposed system and the convolutional neural network (CNN) architecture. CNNs *automatically learn* problem-specific features, obviating the need for devising hand-crafted descriptors like HoG [21]. Furthermore, replacing the largely independent *feature extraction* and *feature learning* modules by an end-to-end framework allows for efficient handling of classification errors. System components are described in following subsections.

### 2.1.1 Image pre-processing

We essentially use the face and eye appearance to detect eye-contact, and pre-processing is limited to extraction of these regions. A state-of-the art facial landmark detector [7] is used to obtain $64 \times 64$ left and right eye patches. Since face pose serves as an additional cue, a $64 \times 64$ face patch obtained using the Viola-Jones detector [99] is also fed to the CNN. The red, green and blue channels for each patch are $z$-normalized prior to input.

### 2.1.2 CNN architecture

Our system comprises three parallel networks (one each for face, left eye and right eye) with a VGGnet [84]-like configuration. CNNs are stacked with *convolutional* (Conv) layers composed of groups of *neurons* (or filters), which automatically compute locally salient features (or activations) from input data. Conv layers are interleaved with *max pooling* layers, which isolate the main activations on small data blocks and allow later layers to work on a 'zoomed out' version of previous outputs facilitating parameter reduction. Convolutions are also usually followed by a non-linear operation (called *rectified linear unit* or ReLU [65]) to make the CNN more expressive and powerful. Finally, in a *fully-connected* (FC) layer, neurons have access to *all* activations from a previous layer as against a Conv layer whose neurons only access local activations.

Each of our three networks have four blocks, with each block including two Conv layers, a ReLU and a max-pooling layer (only Conv layers are shown in Fig. 2.2). Similar activations are enforced for the left and right eye networks by constraining their neurons to learn identical/shared weights. The *filter size* or spatial extent of activations input to a Conv layer neuron is $3 \times 3$ for all blocks, and there are 64, 128, 256 and 128 neurons respectively in the four blocks. A stride length of 1 is used while convolving (computing dot product of) the filter with the input patches. The Conv-L4 outputs are vectorized to a 2048 dimensional vector, which is input to the FC1 layer with 128 outputs. FC1 outputs from the three networks are combined and fed to FC2 followed by FC3, which assigns the input label as either *gaze-locked* or *non-gaze-locked*. The CNN model was implemented on *Torch* [20], and trained over 250 epochs with a batch size of 100. An initial learning rate of 0.001 was reduced by 5.0% after every epoch. To avoid overfitting, a dropout technique was used to randomly remove 40% of the FC layer neurons during training. Interested readers may refer to [84] for further details.

## 2.2 Experiments and Results

### 2.2.1 Datasets

To expressly address eye-contact detection, authors of [85] compiled the **Columbia Gaze** (CG) dataset which comprises 5880 images of 56 persons viewing over 21 different gaze directions and 5 different head poses. Of these, 280 are *gaze-locked*, while 5600 are *non-gaze-locked*– sample CG images

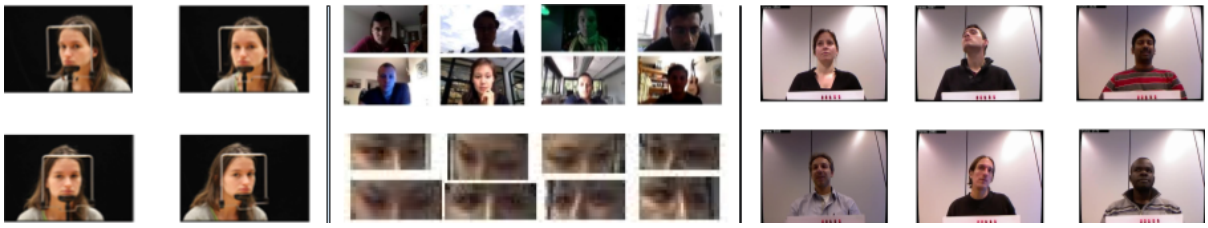

**Figure 2.3** (left) Sample images from the **CG** dataset. (center-top) Original exemplars and (center-bottom) publicly available eye-only images from **MPIIGaze**. (right) Sample images from **Eyediap**.

are shown in Fig. 2.3(left). The CG dataset is compiled in a controlled environment, and contains little variation in terms of illumination and background. The limited size of the CG dataset makes it unsuitable for training CNNs, and we therefore used two large datasets to train our CNN, namely, (1) **MPIIGaze** [106] comprising 213,659 images compiled from 15 subjects during everyday laptop use. As shown in Fig. 2.3(center-top), MPIIGaze images vary with respect to illumination, face size and background. However, only cropped eye images (center-bottom) are publicly available for MPIIGaze; (2) The **Eyediap** dataset [29] (Fig. 2.3(right)) contains 19 HD videos with more than 3000 images each captured from 16 participants. We ignore the depth information available for this dataset, and only use the raw video frames for our purpose.

### 2.2.2 Data Synthesis & Labeling

As only 280 *gaze-locked* images exist in the CG dataset, we generated 2280 *gaze-locked* and 5900 *non-gaze-locked* samples by scaling and randomly perturbing original images as described in [85]. We have downsampled the number of images for the MPIIGaze and Eyediap datasets. MPIIGaze comprises images with continuous gaze direction from 0° to-20° *pitch* (vertical head rotation) and -20° to 20° *yaw* (horizontal rotation). The 3D gaze direction $(x,y,z)$ is converted to 2D angles $(\theta,\phi)$ as $\theta = \arcsin(-y), \phi = \arctan(-x,-z)$. Then, *gaze-locking* implies $(\theta,\phi) = (0,0)$. This way, we obtained 6892 *gaze-locked* and 12000 *non gaze-locked* images from MPIIGaze. Likewise, Eyediap images show users making eye-contact with various screen regions on a 24" PC monitor. We labeled images with the target looking straight ahead (around screen center) as *gaze-locked*, and others as *non gaze-locked*. Table 2.1 presents the training and test sets statistics for the three datasets. We now discuss gaze-locking results with different train and test sets.

**Experiment 1 (Ex1)**

To begin with, we used only the CG dataset for model training[1]. Specifically, we trained our detector with (a) images of only *one eye*; (b) images from *both eyes*; (c) only *face* images, and (d) *face-plus-eye* images as in Fig. 2.2.

[1]The CNN was trained and validated with a 80:20 split of the training set in all experiments.

| Attribute | CG | MPIIGaze | EYEDIAP |
|---|---|---|---|
| **Total Images** | 5880 | 214076 | 125000 |
| **Synthesized** | 8180 | 18892 | 24575 |
| **Training Set** | 7000 | 15000 | 19660 |
| **Test Set** | 1180 | 3892 | 4915 |

**Table 2.1** Training and test set details for the various datasets.

**Experiment 2 (Ex2)**

Here, we repeated Ex1(a) and (b)[2], but first pre-trained the CNN with MPIIGaze and *fine-tuned* the same using CG. Fine-tuning involved modifying only the FC layer weights by re-training with CG images, assuming that the learned Conv-L4 activations were relevant for both MPIIGaze and CG.

**Experiment 3 (Ex3)**

We repeated Ex1(a–d), but pre-trained the CNN with Eyediap followed by fine-tuning on CG.

**Experiment 4 (Ex4)**

To examine the effect of our framework on datasets other than CG, we repeated Ex1(a–d) with a CNN trained on CG and fine-tuned with Eyediap.

| | **1(a)** | **1(b)** | **1(c)** | **1(d)** | **2(a)** | **2(b)** | **3(a)** | **3(b)** | **3(c)** | **3(d)** | **Smith *et al.* [85]** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Acc (%)** | 70.8 | 70.6 | 68.4 | 64.4 | 86.1 | 90.8 | 85.5 | 90.2 | 88.4 | **92.7** | 92.00 |
| **MCC** | 0.69 | 0.72 | 0.67 | 0.36 | 0.74 | 0.81 | 0.74 | 0.80 | 0.78 | **0.83** | 0.83 |

**Table 2.2** Detection performance for Ex1(a)–3(d) and comparison with [85]. Model tested on **CG** in all cases. [85] reports results only on the training set.

### 2.2.3 Results & Discussion

Gaze-locking results are tabulated in Tables 2.2 and 2.3. Detection performance is evaluated in terms of accuracy, and the Mathews correlation coefficent (MCC). MCC is useful while evaluating binary classifier performance on unbalanced datsets, as with our case where the number of *gaze-locked* instances are far less than *non-gaze-locked* ones. In Ex1, accuracy and MCC decrease as more information is input to the CNN (*e.g.*, face=plus-eyes vs eyes/face only), contrary to our expectation. This reduction is attributable to *overfitting* due to the small CG dataset size in comparison to the number of CNN parameters.

However, the benefit of using additional information for gaze-lock detection is evident from Ex2, Ex3 and Ex4 (Ex2 and Ex3 involve pre-training of the CNN model with larger and visually richer datasets).

[2]Since MPIIGaze does not contain face images, we could not repeat Ex1(c) and (d).

| Input | One eye | Both eyes | Face only | Face & eyes |
|---|---|---|---|---|
| **Accuracy** | 62.9 | 65.6 | 64.5 | **66.9** |
| **MCC** | 0.57 | 0.58 | 0.57 | **0.61** |

**Table 2.3** Detection results for Ex4. Model trained on **CG** and fine-tuned/tested on **Eyediap**.

Using *two-eye* information as against *one-eye* in Ex2 improves accuracy and MCC by 4.7 and 7% respectively. Ex3 and Ex4 results are consistent with social attention literature. They confirm that while gaze direction is more critical than head pose for inferring eye contact, combining head and eye orientation cues is optimal for gaze-locking. Our system achieves a best accuracy of 93% and MCC of 0.83 on the CG dataset. Table 2.2 also compares our results with the state-of-the-art [85]. [85] reports detection results on the *training set*, while our results are achieved on an independent test set. With minimal data pre-processing, our model performs similar to [85] using only eye appearance, and outperforms [85] with face-plus-eye information. Finally, while the results for Ex4 again confirm the insufficiency of the CG dataset for training the CNN, the gaze-locking performance significantly improves on incorporating facial and binocular information.

### 2.2.4 Visualizing CNN activations

Fig. 2.4 illustrates four neuronal activations learned in the Conv-L1 layer of our CNN model for the input eye and face images. Conv-L1 activations are informative as ReLU network activations are dense in the early layers, and progressively become sparse and localized. As eye gaze direction is given by the pupil orientation, the eye activations capture edges and textures relating to the pupil. Similarly, the face network activations encode face shape and structural details for pose inference.

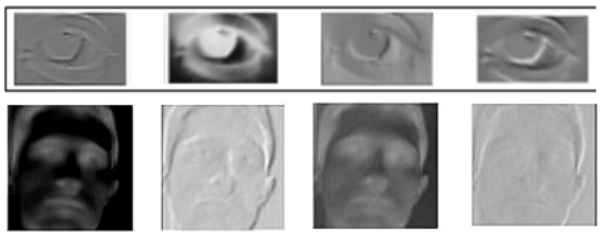

**Figure 2.4** Exemplar Conv-L1 neuron outputs for input eye (top) and face (bottom) images.

## 2.3 CNN Implementation on Android

While our CNN based gaze-lock detector requires minimal pre-processing, the end-to-end framework obviates need for heuristics as with the eye mask extraction phase in [85]. Our system achieves 15 fps throughput on an Intel Core I7 2.6 GHz, 16 GB RAM PC with GeForce GTX 960M GPU. How-

ever, CNNs require large computational and memory resources which precludes their implementation on mobile devices with limited computation and energy capacity.

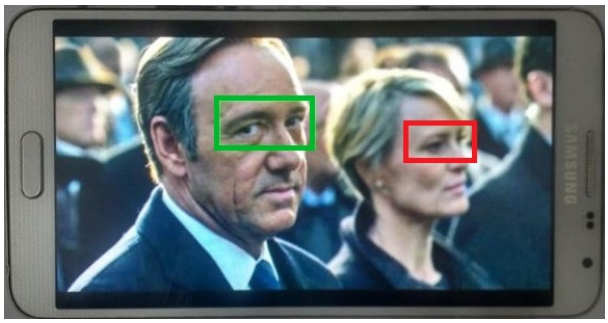

**Figure 2.5** Compressed version of our model working on an Android (Quad-core, 2.3 GHz, 3GB RAM) phone. Green rectangle denotes gaze-locking, while red denotes non-gaze-locking.

This problem can be circumvented by compressing knowledge in a large, complex model to train a simpler model with minimal accuracy loss using the *dark-knowledge* concept [36]. Fig. 2.5 shows our gaze-lock detector on an Android platform, which has a throughput of 1 fps. A more efficient implementation described in [51] can achieve upto 15 fps throughput.

## 2.4 Summary

This work exploits the power of deep CNNs to perform passive eye-contact detection with minimal data pre-processing. Combining facial appearance with eye information improves gaze-locking performance. Our end-to-end system with minimal heuristics can be leveraged by today's smart devices for capturing and managing user attention (*e.g.*, a *smart selfie* application), as well as in image/video retrieval (detecting shots where a certain character is facing the camera).

*Chapter 3*

# AVEID: Automatic Video System for Measuring Engagement In Dementia

Engagement activities for people with dementia (PwD) are an important, non-pharmacological method to promote quality of life and reduce undesirable outcomes such as apathy, depression and aggressive behaviors [19]. Hence, HCI researchers have been developing various systems to supply more interactive and interesting engagement activities for PwD. Examples of these are conversation support systems [5], art therapy [61] and music therapy [24]. Equally important are engagement measurement tools as they provide feedback to facilitators on the effectiveness of engagement systems, while also providing a basis for forming and adjusting interventions. As memory impairments in PwD preclude the use of self-reports as a measurement tool, researchers primarily use some form of observation to code outcome behavior. Behavioral coding requires the use of behavioral observational scales (BOS), and the training of coders who can accurately encode observed behaviors for robust inference. Due to these requirements, behavioral coding as a measurement tool presents the following challenges: 1) It is human-effort intensive; 2) Training coders is time-consuming; 3) Large-scale data annotation becomes tedious, and 4) It supports only coarse-grained behavior analytics due to limitations in human annotation capability.

Hence, automated measures of engagement might be useful for researchers. Despite the availability of such tools for neurotypical target groups, they are not appropriate for use with PwD. For example, PwD tend to resist any type of on-body physical instrumentation [60] so the use of wearable devices or bio-signal systems for measuring engagement is typically not viable. Patel et. al. [71] suggest that PwD require monitoring systems that are "unobtrusive, and preferably collected in a transparent way without patient intervention due to their cognitive impairment."

In this regard, we present AVEID, a low-cost and easy to use video-based system for measuring engagement in PwD. AVEID employs deep learning-based computer vision algorithms to continuously capture a dementia patient's engagement behavior during an interaction session, thereby enabling fine-grained behavior analytics. Consistent with BOS that quantify the patient's attention and attitude towards an engagement system, AVEID estimates the patient's attentional behavior based on gazing direction, and attitude based on facial emotions (Figs. 3.1,3.6).Gazing mannerisms have been extensively studied as cues indicative of attention/engagement during interactions [78, 89], while facial emotions

are inherently reflective of a user's attitude towards the environment. Also, since deep learning systems are 'end-to-end' requiring no manual intervention for model synthesis, AVEID only requires manual annotation of bounding boxes to denote positions of the patient, facilitator (if present) and engagement device at the beginning of the examined video. Unlike gaze-tracking or wearable systems that involve specialized hardware, AVEID only requires a video recording as input. These features facilitate practical, day-to-day usage of AVEID in care homes by therapists or researchers from other domains.

We validated AVEID against human (expert behavioral coder) impressions corresponding to two well-known BOS, namely, the Menorah Park Engagement Scale (MPES) [40] and the Observational Measure of Engagement (OME) [19]. Experiments confirm that measures derived from AVEID agree well with human opinion. Furthermore, AVEID can save the time and effort expended by the behavioral coder, allow for personalized treatment and enable timely analytics on large sample sizes. AVEID measures would also be applicable across small-space engagement activities, facilitating replicability and ecological validity of engagement evaluation studies; we ultimately envision AVEID to provide a strong basis for effective non-pharmacological intervention in dementia care environments.

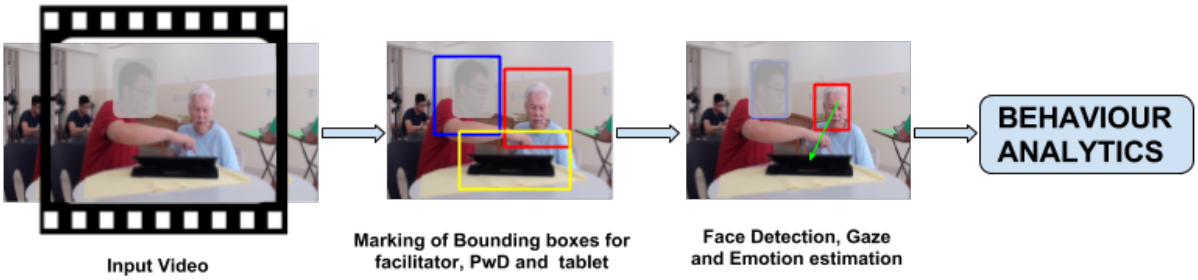


**Figure 3.1** AVEID overview - The AVEID system enables non-intrusive and automated behavioral analytics of persons with dementia (PwD) by capturing their attention (gazing behavior) and attitude (emotion) characteristics.

## 3.1 AVEID Implementation

Consistent with popular BOS used for measuring engagement with PwD, AVEID is designed to quantify the attention and attitude of the patient towards the engagement system over the observed period (Table 1). AVEID employs the patient's gaze focus as a cue towards inferring attention, while utilizing facial affect to infer attitude. We use the following terms to describe the system:

**User:** who operates AVEID to measure engagement

**Target subject:** PwD undertaking an engagement activity.

**Target activity space:** the 2D area where we expect the PwD's gaze to be directed in order to engage with the designed activity. In the AVEID context, the engagement activity involves interaction with a tablet as in [27]. The applications used in tablets are categorized by interests and abilities

of the PwD. For interest areas, these five categories are selected reminiscence, household-linked activities, games, arts and crafts, and chatting.

**Facilitator:** a second person in the frame whose role is to support and promote engagement of the PwD.

## 3.2 AVEID Modules

The AVEID system comprises four modules, namely, *User input, Face detection, Gaze and emotion detection and Behavior analytics* (Fig. 3.1).

The **User Input** module allows users to select the video for analysis, and enables them to mark bounding boxes corresponding to the target subject, target activity space and facilitator. This initialization needs to be performed at the beginning of each video for accurate face detection (under varied video acquisition conditions), and estimation of attentional measures.

The **Face Detection** module implements the *Tiny face* [37] state-of-the-art face detection method. Tiny face performs robust face detection across a wide range of illuminations, face sizes, head poses and facial occlusions. The face detection module detects patient's and facilitator's faces (within the input bounding boxes) in each video frame.

**Gaze Detection** forms the core of AVEID, as its output is used to compute *attention* measures, which are of prime importance in engagement measurement. This module implements the *GazeFollow* deep network architecture of Recasens et al. [77], and utilizes head orientation as a cue to determine where a target is gazing at [69, 85].


**Figure 3.2** The three inputs utilized for gaze detection.

The target's gaze focus is determined based on three inputs (Fig. 3.2): 1) An image (video frame) capturing the scene of interest, 2) Cropped head of the target (output of *Tiny face*), and 3) Location of the head in the scene (denoted by the highlighted grid square). The model comprises two computational pathways. The *gaze pathway* uses the target head appearance and location to produce a *gaze map* that estimates the general direction of the target's gaze. The *saliency pathway* examines scene content to output the *saliency map* that detects interesting objects capable of capturing the target's attention. The two maps are then combined to infer the target's gaze focus. As the target activity space and facilitator

are the two entities of interest in AVEID, the gaze detection module outputs for each video frame, a label signifying whether the target is gazing at the *target activity space, facilitator or elsewhere*.

The **Emotion Detection** module implements the deep network for emotion recognition described in [67]. Given that recognizing emotions of elderly people is challenging even for trained human experts, the deep network described in [67] is fine-tuned with 950 elderly face examples from the FACES dataset [22]. The emotion detection module outputs per video frame a label corresponding to one of the six Ekman emotions plus neutral as illustrated in Fig. 3.6.

The **Behavior Analytics** module processes the outputs of the gaze and emotion detection modules to compute measures reflecting the patient's *attention* and attitude. For characterizing attention, the per-frame gaze labels are combined to compute three raw (or basic) statistics, namely, gaze proportion on tablet, facilitator and elsewhere over the period of observation. In addition to these coarse-grained features, we also derived 18 fine-grained statistics from the gaze labels for analysis, described as follows.

Upon determining *episodes of focus* on the tablet, facilitator and other *entities* within the observation period, we computed the means and standard deviations (std) of these episode durations (**6 features** in total); likelihood of transitioning from one entity to another- e.g., transition from focusing on tablet to focusing elsewhere; this gives rise to **six transition probability features** corresponding to 3 permute 2 entity transitions. An additional *six gaze flux features* denoting gaze flux into and out of the three entities were obtained from marginal likelihoods– e.g., P(gaze flux into tab) = P(fac $\rightarrow$ tab) + P(others $\rightarrow$ tab), where $\rightarrow$ denotes a gaze transition.

To quantify attitude, we computed the proportions of positive (neutral or happy emotion) and negative (angry, sad or disgusted) affect over the observation period from the per-frame facial emotion labels.

In terms of computational hardware, AVEID requires a Graphics Processing Unit (GPU) for video processing. The current system is implemented on a Xeon processor with 64GB RAM, and 12 GB NVIDIA GeForce GTX 1080 Ti GPU memory.

## 3.3 BOS for Validation

We validated AVEID by comparing the obtained attention and attitude measures against expert annotations acquired for the OME and MPES scales, whose descriptions follow.

The **Observational Measure of Engagement** (OME) [19] is an observational scale to directly assess engagement in PwD. For this BOS, observers are first required to detect *time periods denoting PwD engagement*, and score the attention and attitude levels of the patient within these engagement periods. OME represents a very coarse-grained assessment of PwD engagement, and can best facilitate examination of engagement periods, as no codings are made when the dementia patient is disengaged from the target activity. The **Menorah Park Engagement Scale** (MPES) [40] is a more fine-grained BOS, as PwD engagement is assessed over 5-minute time periods. Three types of engagement, namely, active engagement with the target device/activity, passive engagement and engagement with others, are measured in this BOS (Table 3.1).

**Table 3.1** Table 1. Measuring attention and attitude via the MPES and OME scales, and the matching measures used with AVEID

| Bos | MPES BOS | OME BOS | AVEID |
|---|---|---|---|
| **Unit of Assessment** | 5-minute observed periods, coded with 0,1 or 2 | Identified period of engagement, rated on a 7-point scale | User-specified observed periods of time (flexible granularity) |
| **Attention** | Active engagement (Did target activity), Passive engagement (Watched target activity), Other engagement. | Attention intensity (1 denoting no attention) | 3 raw + 18 derived gaze-based statistics over observed period. |
| **Attitude** | Pleasure and anxiety as proportion over observed period. | Attitude valence with (1 denoting strongly negative, 3 denoting neutral, and 7 denoting strongly positive affect). | Proportion of negative and neutral-or-positive affect over observed period. |

**Table 3.2** Scores were obtained from experienced therapists or trained researchers.

| Bos | No. Videos | No. of Video Segments | Annotated by |
|---|---|---|---|
| **OME** | 7 | - | Therapist |
| **MPES** | 20 | 130 | Trained researchers |

## 3.4 Expert Score acquisition

All annotated videos were as shown in Fig.3.1, where a PwD engages with an interactive tablet aided by a facilitator [27]. For OME scoring, a dementia care therapist with 10 years experience indicated periods of patient engagement in seven 15-minute video segments according to the following OME definition: "amount of attention the person was visibly paying to the stimulus (tablet) via eye movements; manipulating/holding and talking about it." [19]. MPES scores were provided by researchers trained to attain 0.8 (Kappa) inter-rater reliability. They scored 5-minute segments from 20 videos (30 minutes each), for active engagement (*Did target activity*), passive engagement (*Watched target activity*) and engagement with other stimuli (*Attention on activity other than target*) on an ordinal scale (Table 3.1). Table 3.2 summarizes the annotation statistics.

## 3.5 Validation Study Result and Discussion

### 3.5.1 Measures of Attention

In the OME BOS, engagement periods are identified and then annotated for attention level and attitude. So, we computed the proportion of patient's gaze focus on the target activity space for a) those segments where the therapist indicated engagement, and b) the remaining video segments where the therapist inferred disengagement. Fig. 3.3 presents the computed gaze proportions for seven videos. Higher distribution of gaze focus on tablet was clearly noted during engagement periods, as confirmed by a two-sample Kolmogorov-Smirnov test at $p<0.001$. Therefore, gaze focus on the target was sufficient to convey the notion of attention as with the OME.

The MPES BOS quantifies attention over 5-minute intervals. Fig. 3.4 presents Pearson correlations computed between active, passive and other engagement MPES scores, and the 21 AVEID attention features in the form of a 3×21 grayscale image. Negative and positive correlations are respectively denoted by darker and lighter shades. Red and cyan symbols respectively denote significant ($p<0.05$) and marginally significant ($p<0.1$) correlations.

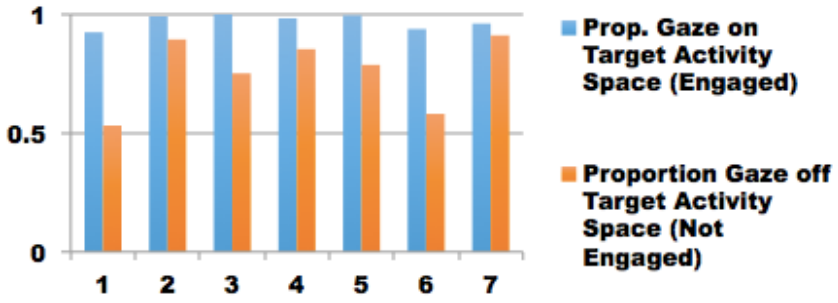


**Figure 3.3** Gaze proportions on tablet (engagement device) during engaged and not-engaged periods as per OME BOS.

Active engagement is significantly and positively correlated with the extent of gaze focused on tablet, mean duration of tablet gazing episodes and standard deviation of these episodic durations, while being negatively correlated with the extent of gaze focus on the caregiver and other areas. Active engagement is also marginally and positively correlated with the gaze flux in and out of the target activity area, suggesting that focus on target activity area as well as periodic gaze switching when communicating with the facilitator are linked with higher active engagement scores as assessed by the expert.

On the other hand, passive engagement marginally and negatively correlates with the gazing durations on the facilitator. This pattern is concurrent with the MPES description of passive activity where the PwD behaves with less enthusiasm and is not having social interactions with the facilitator. The final MPES item, engagement with other, positively correlates with gazing on facilitator, and negatively with the mean and standard deviation of gazing episode durations on the target activity area. This suggests

that engagement with other activities, as coded by the experts, is associated with behaviors where the PwD directs attention more toward the facilitator rather than toward the presented activity.

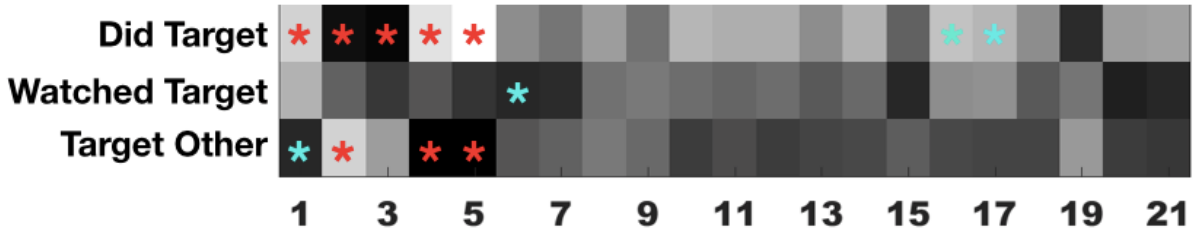


**Figure 3.4** Correlations between gaze-based AVEID features and MPES scores for active (Did Tgt), passive (Watched Tgt) and other engagement (Tgt Other). Red and cyan marks denote correlations significant at ($p<0.05$) and ($p<0.1$).

To summarize, gaze on target as a correlate of engagement is validated by both the OME and MPES coding methods. Furthermore, attention measures computed via AVEID are able to capture a number of aspects concerning these BOS.

### 3.5.2 Measures of Attitude

Inferring facial emotions of PwD is a known challenge due to older adults exhibiting facial emotions in a controlled manner, ageing skin and muscles significantly modulating facial appearance, and indicating a prominently negative affect [26]. Additionally, PwD often display flattened affect [55]. However, greater engagement from the PwD should also elicit a positive reaction from the facilitator whose role is to promote such behavior. As facilitators are neurotypical adults whose facial emotions can be better recognized with available computer vision tools, we therefore examined if higher MPES attitude scores correlate better with the facilitator's facial emotions (Fig. 3.6). This hypothesis turned out to be true, with pleasure scores correlating significantly and positively with the proportion of positive facial affect exhibited by the facilitator ($r$=0.24, $p$<0.01). Therefore, examining facilitator behavior could provide crucial cues for measuring engagement in PwD.

Overall, results reveal that AVEID can effectively capture patient (and facilitator) behavior indicative of attention and attitude. Also, since AVEID measures are based on per-frame gaze and emotion labels, it is possible to go beyond coarse engagement measures that BOS provide. E.g., even though the patient's verbal behavior was not captured in the videos, the frequency with which the patient directs gaze towards the facilitator may serve as an effective cue to this end. Finally, gazing and attitude estimation can be reliably accomplished for small-space activities (where the patient's face is clearly visible), facilitating evaluation of multiple engagement designs.

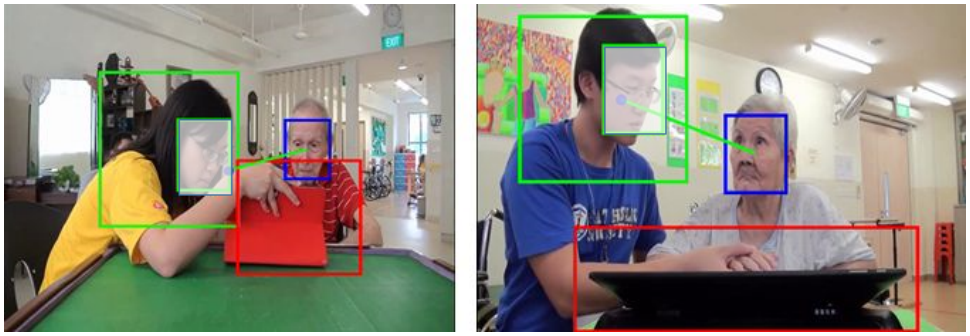

**Figure 3.5** Examples where gaze focus estimation is incorrect (zoom to view).

**Figure 3.6** Exemplar emotion estimation results. Facial emotions of the PwD and facilitator are correctly identified (left). PwD's emotion is incorrectly estimated, but facilitator's emotion is correct (right) (zoom to view).

## 3.6 Challenges and Limitations

Figures 3.5 and 3.6 illustrate the challenges involved in video-based engagement measurement for PwD, where the observation videos are captured under unconstrained settings. Fig.3.5 presents two examples of incorrect gaze focus estimation due to the closeness of the facilitator to the tablet, and due to the 2D video information being insufficient to model the 3D world. Likewise, in Fig. 3.6 (right), the patient's facial appearance is mistaken by the algorithm as an exhibition of sadness. It needs to be acknowledged that gazing behavior can only indicate passive engagement; the use of wearable physiological sensors [91] may be necessary for inferring cognitive engagement.

*Chapter 4*

# An EEG Based Image Annotation System

Human vision is a very powerful system for object recognition and scene understanding. It is also robust to variations in illumination, scale or pose. We are habitually used to recognizing objects even in cluttered scenes. Humans can identify objects in tens of milliseconds [45, 68], but the representation of the perceived information via hand movements or verbal responses for annotation is very slow compared to the processing speed of contemporary digital devices. In this regard, the emerging field of brain-Computer Interfaces (BCI) offers us an innovative way to exploit the power of human brain for data annotation with minimal effort.

Brain-Computer Interfaces rely on various technologies for sensing brain activity such as Electroencephalography (EEG), MEG (Magnetoencephalography), PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), fMRI (functional Magnetic Resonance Imaging) and fNIRS (functional near infrared spectroscopy). Among these, EEG provides a high temporal resolution (sampling rate of up to 1 KHz) and adequate spatial resolution (1-2 cm). In this work, we specifically use the portable and easy-to-use consumer grade *Emotiv* EEG device, which enables a minimally intrusive user experience as users perform cognitive tasks, for sensing and recording brain activity. While having these advantages, consumer EEG devices nevertheless suffer from a high signal-to-noise ratio, which makes subsequent data analytics challenging.

In this work, we focus on the annotation of a *pre-selected* object category over the entire image dataset instead of labeling all categories at once. If the images are presented serially in a sequence for annotation, then the task is equivalent to that of *target detection*. Now whenever an image containing a target class instance is observed by the human annotator, an event-related potential (ERP) signature known as **P300** [57] is observed in the EEG data. By examining the EEG signals generated during image presentation, we can discover the images of interest and annotate them accordingly. In this paper, we provide the pipeline and architecture for image annotation via EEG signals.

## 4.1 Related work

The use of EEG as an additional modality for computer vision and scene understanding tasks has been explored by a number of works. In [62], EEG signals are used to automate grab cut-based image segmentation. In [74], authors exploit ERP signatures such as P300 for image retrieval. In [49], authors use the N400 ERP to validate tags attached to video content. Emotions from movies and ads are inferred via EEG signals in [90] and [82].

Few studies directly use image category-based EEG signatures for recognizing aspects related to multimedia content as well as users. For example, the authors of [42] use EEG signals to classify images into three object categories– animals, faces and inanimate. In a recent work [86], the authors present how EEG features can be employed for multi-class image classification. Another recent work recognizes user gender from EEG responses to emotional faces [11]. Given the state-of-the-art, the key contributions of our work are we how how (i) the P300 ERP signature can be employed for image annotation; (ii) the model trained for one object category can be directly used for a novel category, and (iii) the image presentation time affects annotation system performance for complex images.

## 4.2 System architecture

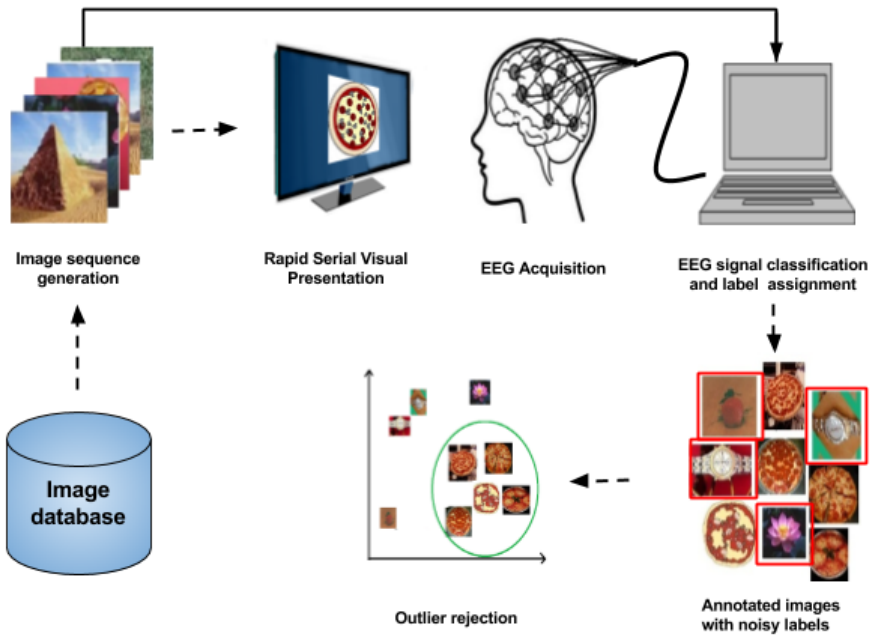


**Figure 4.1 EEG-based annotation pipeline:** An exemplar illustration for the *pizza* object class is presented. Best viewed in color and under zoom.

The proposed image annotation system consists of several components– RSVP generation, EEG data acquisition, EEG pre-processing, (binary) classification and outlier removal. Fig.4.1 presents an overview of the EEG-based annotation pipeline. The RSVP generation unit prepares the set of images for viewing, so that a few among those correspond to the target object category. The image sequence is created via random sampling from the whole dataset. A human annotator is then asked to identify the target category images as the sequence is presented rapidly, and the annotator's brain activity is recorded via an EEG headset during the visual recognition task. The compiled EEG data is first pre-processed for artifact removal. Then, the classification unit categorizes the EEG responses into *target* and *non-target* annotations based on P300 patterns. Images classified as *target* are annotated with the target label class. However, this labeling is noisy due to the presence of false positives and imbalance towards the negative (non-target) class. An outlier removal unit finally performs unsupervised dimensionality reduction and clustering to improve the labeling precision.

### 4.2.1 Rapid Serial Visual Presentation and Oddball paradigm

Rapid Serial Visual Presentation is popularly used in psychophysical studies, and involves a series of images or other stimuli types being presented to viewers with a speed of around 10 items per second. This paradigm is basically used to examine the characteristics pertaining to visual attention. In RSVP studies, the *oddball* phenomenon [73] is widely used. In the oddball paradigm, a deviant (target) stimulus is infrequently infused into a stream of audio/visual stimuli. For EEG-based annotation, we generated an RSVP sequence by combing a few *target* category images with many *non-target* images via random sampling from the original dataset. Each image in the sequence was then shown to the viewer for 100 ms, and a fixation cross was presented for 2 seconds at the beginning of the sequence to minimize memory effects and to record resting state brain activity (see Fig.4.4).

### 4.2.2 EEG data preprocessing and classification

We used the *Emotiv EPOC* headset to record EEG data. This is a 14 channels (plus CMS/DRL references, P3/P4 locations) Au-plated dry electrode system. For ERP analysis, the Emotiv provides signals comparable to superior lab-grade EEG devices with 32, 64 or 128 channels. The headset uses sequential sampling at 2048 Hz internally which is down-sampled to 128 Hz. The incoming signal is automatically notch filtered at 50 and 60 Hz using a $5^{th}$ order sinc notch filter. The resolution of the electrical potential is 1.95 $\mu$V. The locations for the 14 channels are as per International 10-20 locations as shown in Fig.4.2.

The recorded EEG data is contaminated by various noise undesirable signals that originate from outside the brain. For instance, while recording EEG, one often encounters 50/60Hz power-line noise and artifacts caused by muscle or eye movements. We extracted one second long *epochs* corresponding to each 100 ms long *trial* denoting the presentation of an image, with 128Hz sampling rate. Our EEG preprocessing includes (a) baseline power removal using the 0.5 second pre-stimulus samples, (b)

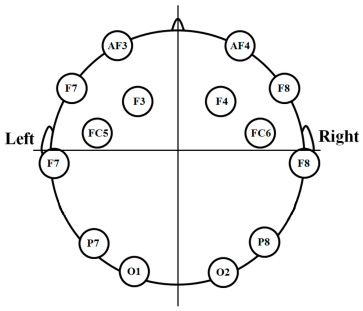


**Figure 4.2 Sensor configuration:** Emotiv electrode locations as per International 10-20 system.

band-pass filtering in 0.1-45Hz frequency range, (c) independent component analysis (ICA) to remove artifacts relating to eye-blinks, and eye and muscle movements. Muscle movement artifacts in EEG are mainly concentrated between 40-100 Hz. While most artifacts are removed upon EEG band-limiting, the remaining are removed manually via inspection of ICA components.

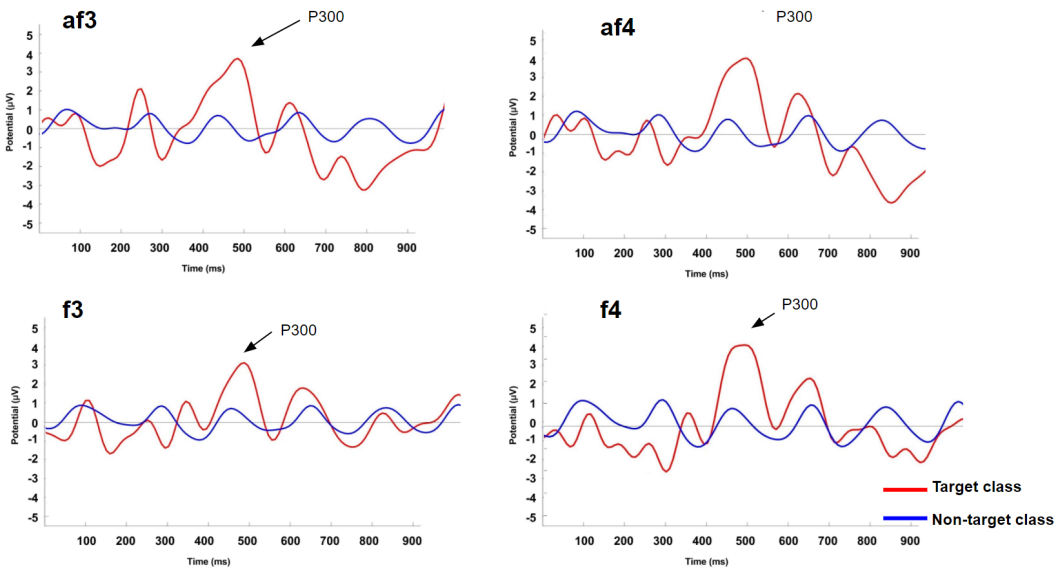


**Figure 4.3 ERP plots:** ERP curves for the Emotiv af3, af4, f3 and f4 channels for *target* (red) and *not-target* (blue) images. P300 signatures are evident for targets but not for non-targets.

The human brain's response to a stimulus can be measured as a voltage fluctuation resulting from the ionic current within the neurons. The event-related potential is one such measure that is directly related to some motor, cognitive or sensory activation. Out of various ERP components, the P300 signature is commonly elicited in the oddball paradigm where very few targets are mixed with a large number of non-targets. In our experimental setup, we employed a 1:12 ratio for target-to-non-target images.

As shown in Fig.4.3, the P300 ERP signature is observed between 250 to 500 ms post *target* stimulus presentation. Also, the ERP response is significantly different for target and non-target images, and therefore can be exploited for EEG-based image annotation.

We used the Convolutional Neural Network (CNN)-based EEGNet architecture [54] to classify our EEG data based on P300 detection in the RSVP task. The EEGnet architecture consists of only three convolutional layers. All layers use the Exponential Linear Unit (ELU) [17] as nonlinear activation function with parameter $\alpha = 1$. We trained the model using the minibatch gradient descent algorithm with categorical cross-entropy criterion and Adam optimizer [46]. The models were trained on a NVIDIA GEFORCE GTX 1080 Ti GPU, with CUDA 8 and cuDNN v6 using the Pytorch [70] based Braindecode [80] library.

### 4.2.3 Outlier removal

We select one category at a time for the annotation task, which results in class imbalance for the RSVP task. The selected object category forms the *target* class, while all other categories collectively form the *non-target* class. Due to this heavy class imbalance and the characteristics of P300 as discussed in Section 4.4, the false positive rate of the predicted labels is high. Therefore we performed unsupervised outlier removal on the predicted *target* images. Deep learning features have proven advantages over hand-crafted features like SIFT and HoG [104]. We used a pre-trained VGG-19 model [84] to obtain the feature descriptors for the targets. These feature descriptors provide compact representation of raw images while preserving the information required to distinguish between image classes. Each target image was fed forwarded within the VGG-19 model to obtain the 4096 dimensional feature vectors. Target images need not belong to the image classes on which the model is pre-trained. Then, we perform dimensionality reduction with t-SNE [59] to generate low-dimensional features. The t-SNE algorithm retains the local structure of the data while also revealing some important global structure, and hence it performs better than principal component analysis (PCA) alone.

In our case, we assume that samples from the target class should be close in feature space as compared to non-target samples. By performing a grid search on hyper-parameters, we found that the algorithm works best with perplexity value 20, 50 PCA components and 3-5 output dimensions. Then, we performed $k$-means clustering for two classes assuming that target class samples will form a cluster distinct from the false positives. Also, since the false positive cluster would contain samples from many categories, the cluster would not be as dense as the target cluster.

## 4.3 Protocol design and Experiments

### 4.3.1 Datasets

To evaluate the performance of our image annotation system, we used the Caltech101 (CT) [25] and Pascal VOC2012 (PV) [23] datasets. The CT dataset consists of 101 object categories with 40 to 800

images per category. The PV dataset contains a total of 11,530 images from 20 categories, and multiple object categories can be present in one image.

### 4.3.2 Experimental setup

We utilized 2500 images for training, and 2500 images for testing. Both these image sets comprised 200 images of a particular target category that we wanted to annotate. All images were resized 512×512 pixels, and images were displayed at 10 Hz frequency in blocks of 100 in order to minimize viewer distraction and fatigue. During the RSVP task, participants were shown a fixation display for 2 seconds at the beginning of each 100 image sequence. Train and test EEG data were captured using an identical experimental setup with the temporal gap of 5 minutes. Target image categories were decided *a priori* before every experiment.

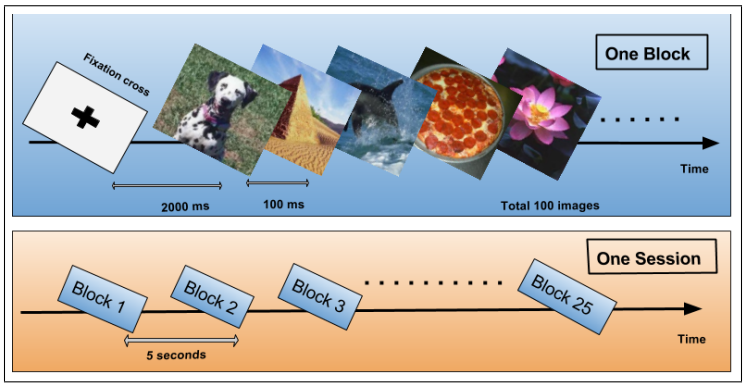


**Figure 4.4 Experimental protocol:** Participants completed two identical sessions (one used for training and the other for test) which were 5 minutes apart. Each session comprised 25 blocks of 100 images, and lasted about six minutes.

Our study was conducted with five graduate students (5 male, age 24.4 ± 2.1) with 10/20 corrected vision, seated at a distance of 60 cm from the display. A total of three sessions (each involving train and test set) were performed with each participant. To facilitate engagement, viewers were instructed to count the number of target images during the experiment. Target image classes were different for each session, and included categories like *bike*, *pizza*, *panda*, *sofa*, *etc*. Each participant performed two sessions on the CT dataset and one session on the PV dataset.

## 4.4 Results and Discussion

Due to a heavy class imbalance between *target* and *non-target* category images, we use the F1-score to evaluate our annotation results. The F1-score is a popular performance metric used in retrieval studies,

and denotes the harmonic mean of the precision and recall scores. All reported results denote the mean F1 achieved with five-fold cross validation.

**Table 4.1 Results synopsis:** Annotation performance obtained for the CT and PV datasets across total 15 sessions (5 viewers).

| **Dataset** | **Caltech101** | **Pascal VOC 2012** |
|---|---|---|
| **Before outliers removal** | | |
| F1 score | 0.71 | 0.68 |
| Precision | 0.66 | 0.63 |
| Recall | 0.81 | 0.72 |
| **After outliers removal** | | |
| F1 score | **0.88** | **0.83** |
| Precision | 0.99 | 0.97 |
| Recall | 0.81 | 0.72 |
| Target image percentage | 8% | 8% |
| Image presentation speed | 10 Hz | 10 Hz |
| Number of images in test set | 2500 | 2500 |

In Table 4.1, we report the averaged F1 and precision-recall values for the CT and PV datasets across all participants. Note that the precision and F1 scores improve significantly upon outlier removal due to a stark reduction in the number of false positives via feature-based clustering. Overall F1 scores for the PV dataset are lower than for the CT dataset. This can be attributed to the fact that the PV dataset is more complex, as it contains multiple object classes in many images, as compared to CT which contains only one object class per image.

As our annotation system is dependent on viewer ability, its performance is sensitive to human factors. One key factor is the image presentation rate. The image display latency (100 ms) is lower than the P300 response latency ($\approx$ 300 ms) [75]. The rapid image display protocol results in (i) viewers confusing between similar object classes, (ii) viewers unable to fully comprehend visual information from complex images, and (iii) EEG data for consecutive images having significant overlap leading to misclassification.

Therefore, we hypothesized that reducing the image display rate would (a) allow the viewer to better comprehend the visual content (especially for complex images), (b) better delineation of EEG responses, and (c) better manifestation of ERP signatures. These in turn, would improve our annotation performance while marginally reducing the annotation throughput. Fig.4.5 presents the observed results. Note that a 3% increase in F1-score is observed when the image presentation rate is reduced from 10 to 4 images/second, validating our hypothesis.

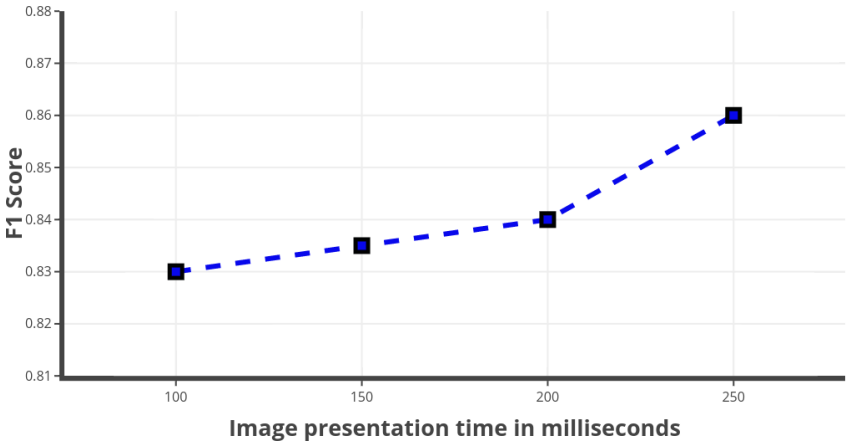


**Figure 4.5 Presentation rate vs annotation performance:** Variation in F1-score with image display rate.

Conversely, since our annotation system is solely based on P300 signatures which are task specific but target class agnostic. Therefore, it is not mandatory to train the EEGNet with object class-specific EEG responses. To validate this aspect, we trained and tested the EEGNet with EEG responses corresponding to different object categories. Table 4.2 presents the F1 scores achieved for the five viewers with class-agnostic train and test EEG data. Note that only a marginal difference in annotation performance is noticeable with class-specific and class-agnostic EEG data across viewers. Since we are using the pre-trained VGG-19 model exclusively to extract feature descriptors, it can be used without further fine tuning for any new target class categories.

**Table 4.2** Annotation performance with class-specific vs class-agnostic EEG data for five viewers.

| **F1 Score** | **P1** | **P2** | **P3** | **P4** | **P5** |
|---|---|---|---|---|---|
| Class-specific train and test | 0.88 | 0.86 | 0.89 | 0.87 | 0.88 |
| Class-agnostic train and test | 0.85 | 0.85 | 0.84 | 0.86 | 0.86 |

## 4.5 Summary

In order to facilitate large-scale image annotation efforts for computer vision and scene understanding applications, we propose an EEG-based fast image annotation system. Our annotation system exclusively relies on the P300 ERP signature, which is elicited upon the viewer detecting a pre-specified object class in the displayed image. A further outlier removal procedure based on binary feature-based clustering significantly improves annotation performance.

Overall, our system achieves a peak F1-score of 0.88 with a 10 Hz annotation throughput. Another advantage of our method is that the P300 signature is specific to the target detection task, but not the

underlying object class.Therefore, any novel image category can be annotated with existing models upon compiling the viewer EEG responses.

*Chapter 5*

# EEG-based Cognitive Load Estimation Across Visualizations

*A picture is worth a thousand words*– this aphorism has spurred the growth of ***visual analytics*** (VA), where the combined strength of big data mining, interactive visualization and human analytical reasoning is exploited to solve complex, real-world problems [1]. Given that the human expert is a key component of the visual analytic process, prior research has established that visual interfaces need to effectively cater to strengths and limitations in human *perception* and *cognition*. Perception relates to the optimal use of human sensory resources, mainly visual and auditory (use of touch, smell and taste has also been explored recently), to provide the user with an informative 'mental image' of the data/problem at hand. Cognition mainly relates to the cognitive or mental workload[1], imposed on the user during interactive problem solving.

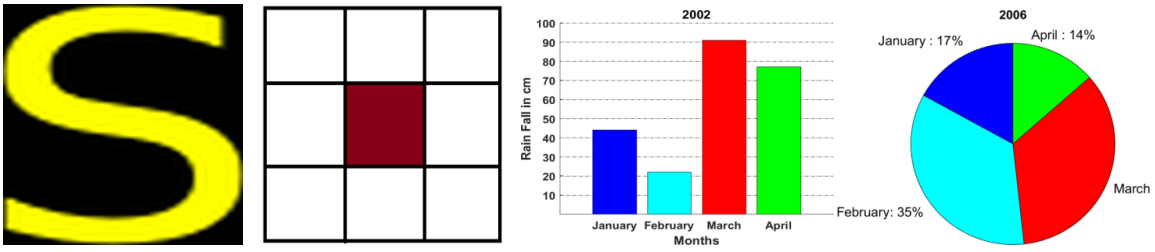


**Figure 5.1 Problem Statement:** Under varying mental workload levels induced by the *n*-back task, we examined if there was any similarity in user cognitive behavior captured via EEG across four visualizations. Figure shows (from left to right) exemplar *character*, *position*, *bar* and *pie* visualizations.

According to Chandler [14], cognitive load can be categorized as either natural or extraneous. In the context of visual interfaces, natural cognitive load is inherently imposed by the task on hand while extraneous cognitive load is dependent on the type of visualization used. As with traditional interfaces, a key usability requirement of visual interfaces is that the presented visualizations be intuitive and induce minimum cognitive load in users. However, since visual analytic systems are also required to support exploratory user behavior and provide insights, the use of traditional usability heuristics or design questionnaires is unsuitable for evaluating interactive visualization interfaces (or Viz UIs) [72, 79].

[1] Mental workload is loosely defined as the amount of mental resources expended during task performance.

Neuroergonomics, which examines human factors by employing neuroscientific methods, represents an interesting and viable alternative for evaluating Viz UIs. Lately, there has been considerable interest in *cognitive sensing* via eye movements [38,76] and neural activity in the form of EEG [3,101] or fNIRS signals [72] for Viz evaluation. These cognitive sensing methods focus on unobtrusively assessing users' mental workload in a fine-grained manner using light-weight, wireless devices that minimally impact task performance [101]. While these methodologies reliably and objectively measure cognitive load, their applicability is highly task and visualization specific and not generalizable [44].

This work investigates the suitability of a single EEG framework to assess (extraneous) memory workload across multiple visual interfaces under similar task difficulty. If cognitive load estimation (CLE) is generalizable, it would naturally facilitate usability evaluation of Viz UIs. We would then be able to automatically assess if Viz UI **B** is *more* or *less* usable than Viz UI **A** for a specific task when user cognitive data are available for **A** and **B**. Alternatively, a smart visual interface could replace/augment the current visualization with a potentially more interpretable one upon detecting high mental workload.

The possibility of realizing a generalizable CLE model has brightened with the success of deep convolutional neural networks (deep CNNs), which robustly learn problem-specific features, and effectively adapt to related tasks with minimal additional training [82]. We examined if user EEG responses obtained for the *character*, *spatial pattern* or *position*-based, *bar graph* and *pie chart* visualizations, under different mental workload levels induced by the *n*-back task [8,44,72,101] had any compatibility so as to facilitate usability benchmarking (Figure 5.1). In lieu of learning a single CLE model, we learned one model for each visualization and evaluated the same on data obtained from the others. We employed two state-of-the-art algorithms for EEG-based CLE: deep CNN [8], and proximal SVM (pSVM) [101]. From experimental results, we observe that (a) both models perform well when the train and test EEG data are from the same (visualization) source, with pSVM generally outperforming deep CNN, and (b) the deep CNN better predicts cognitive load across visualizations, even though both models perform relatively poorly in these conditions.

## 5.1 Related Work

We now overview related work on (a) Viz UI evaluation, and (2) Cognitive sensing for mental workload assessment.

### 5.1.1 Viz UI evaluation

User studies have been the traditional tool for evaluating Viz UIs. While focused studies such as [50] can help answer specific questions regarding visualization techniques, they are grossly inadequate for evaluating visual interfaces as (a) VA tools need to provide the human expert insights by presenting the same data from multiple perspectives and (b) subjective human opinions are more indicative of holistic experience, and cannot effectively capture the UI's effectiveness at supporting data exploration. The

position paper of Riche [79] champions the use of quantitative measures such as the number of insights made by users for Viz evaluation, and espouses how physiological and cognitive sensing can enable passive detection of 'aha moments' denoting insight discovery. Greitzer *et al.* [30] also remark the need for progressing with the science and practices for VA design and evaluation, and call for the application of cognitive theories and empirical results thereof in this regard.

### 5.1.2 Cognitive sensing for assessing mental workload

Following Klimesch's findings that spectral power in the $\alpha$ (8-13 Hz) and $\theta$ (4-8 Hz) bands are reflective of cognitive and memory performance [47], Anderson *et al.* [3] assess mental workload from EEG data passively recorded via the wireless Emotiv headset during plot interpretation. Visualization interpretation requires processing of the visual working memory, and higher stress induced on the visual working memory (VWM) reflects in the form of decreased $\alpha$ and increased $\theta$ power with respect to the baseline. Functional near-infrared spectroscopy (fNIRS) represents another popular cognitive sensing technique for assessing mental workload, as fNIRS devices are portable, light-weight and convenient for passively recording long-term user cognitive activity. Peck *et al.* [72] demonstrate how fNIRS can capture the difference between pie chart and bar graph visualizations. Nevertheless, fNIRS systems have a much lower temporal resolution than EEG counterparts, making them ill-equipped to capture rapidly dynamic brain activity.

Many CLE works including the aforementioned employ the *n*-back test to induce load in the VWM. The *n*-back test is reflective of VWM functionality as it involves storage and retrieval of information over short durations, and requires regular updation of VWM contents– it can be used to elegantly manipulate VWM load without impacting the nature of the input stimulus or motor output [12]. Wang *et al.* [101] estimate mental workload from *n*-back with EEG, employing a host of spectral and statistical features. Upon determining informative features via an information-theoretic feature selection method, a proximal SVM (pSVM) is employed to benchmark workload levels. Deep CNNs, which have become extremely popular in computer vision, are used for CLE from *n*-back in [8]. Unlike most methods that represent EEG activity as a vector, the spatial EEG structure is preserved by synthesizing multi-spectral images which are then input to the CNN for CLE.

### 5.1.3 Analysis of related work

Examination of related work clearly reveals that the VA community favors quantitative evaluation of Viz UIs, and in particular the use of cognitive sensing as a tool to this end. Considerable effort has been devoted to CLE from brain activity analysis. However, (a) many works are restricted to *validating* a cognitive sensing mechanism for Viz evaluation rather than *predicting* mental workload using the same, and (b) the handful that attempt prediction implicitly assume that the influence of different visualizations on the elicited cognitive processes is minimal given task difficulty.

To elaborate, the fNIRS work of Peck *et al.* [72] examines both bar graph and pie chart visualizations and hypothesizes that the two will generate varied cognitive processes. While the authors find that the cognitive demands imposed by the two vary for different users which reflects via deoxygenated hemoglobin levels in the fNIRS signal, CLE based on fNIRS is not performed in this work. Conversely, the recent work of Wang *et al.* [101] explicitly attempts to categorize mental workload levels via a pSVM-based classification strategy. While employing both letter and spatial pattern-based comparisons in the *n*-back task, it nevertheless pools the data acquired for both these visualizations for analysis, disregarding any difference in cognitive behavior that the two Viz types may induce. A recent and closely related work of Ke *et al.* [44] proposes feature selection for improving CLE across the working memory and multi-attribute tasks, but does not focus on comparing visualizations *per se*.

Therefore, despite the existence of a large body of prior work, we are still unable to answer whether ***on fixing the task type and difficulty level, can cognitive sensing inform whether one visualization more intuitive than the other?*** An answer to this question is possible if we gain an understanding of the cognitive processes associated with the Viz processing pathway (*e.g.*, visual or auditory), and the cognitive features needed to benchmark visualizations on a canonical scale. This is not trivial as cognitive sensing has its own pitfalls– considerable intra and inter-subject differences in brain activity patterns occur due to intrinsic and extrinsic factors such as brain sensor locations across trials [8]. This grossly impedes generalization of CLE results. As remarked by Ke *et al.* [44], if a trained CLE model can only work well with data acquired for a certain task, visualization type and a set of users, then such an exercise would have no ecological validity. In this regard, Anderson *et al.* [3] compare visualizations for cognitive load, but in a rather simplistic way where the visualized data distributions are not matched with respect to interpretation difficulty.

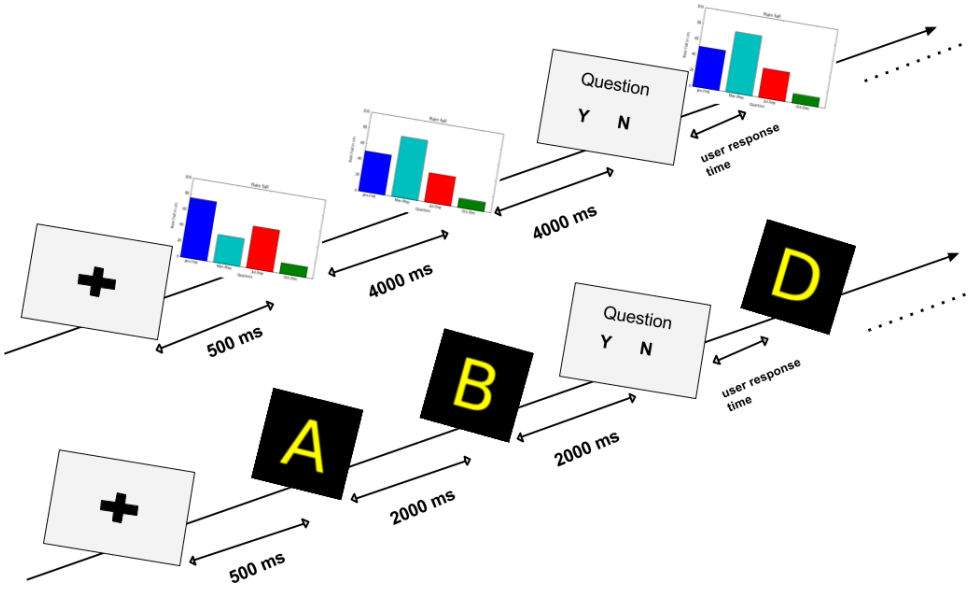


**Figure 5.2** Protocol timeline with 1-back exemplars.

**Contribution:** As a first step towards answering the above question, we examined *if there were similarities among the cognitive processes elicited by four different visualizations during n-back, based on signals captured by a wireless EEG headset*. Wireless neural sensors suffer from low signal fidelity, but nevertheless compare favorably with respect to lab devices for real-life applications as they are convenient to use, enabling a non-intrusive and naturalistic user experience. Our methodology is elaborated in the next sections.

## 5.2 Materials and Methods

The adopted experimental design and protocol are as follows.

### 5.2.1 Experimental Design

#### 5.2.1.1 Stimuli and users

We employed four visualization types, namely, character (*char*), spatial pattern or position-based (*pos*), bar graph (*bar*) and pie chart (*pie*) in our study. Exemplars for each Viz type are presented in Figure 5.1. These visualizations have been used in prior studies (*e.g.*, [8, 72, 101]), and are designed to utilize the visual sensory pathway and working memory. While some works [44] have employed verbal *n*-back for CLE, the validity of comparing or integrating visual and auditory memory performance is unclear as the two appear to have varying capabilities [18].

| Response | Predictor | Viz Type | N-back | Viz Type*N-back | Error | Total |
|---|---|---|---|---|---|---|
| RT | df | 3 | 3 | 9 | 304 | 316 |
| | F | 3.96 | 27.17 | 4.33 | | |
| | Sig | $p < 0.001$ | $p < 0.000001$ | $p < 0.000001$ | | |
| RA | F | 74.04 | 24.75 | 1.62 | | |
| | Sig | $p < 0.000001$ | $p < 0.000001$ | n.s. | | |

**Table 5.1** ANOVA summary for RTs and RAs. df and Sig respectively denote degrees of freedom and significance level.

Each *char* stimulus comprised one of sixteen characters (selected randomly) centered on the screen to ensure maximum readability, while each *pos* stimulus was a $3 \times 3$ spatial grid with one of the nine blocks highlighted. The *bar* and *pie* stimuli were generated from real-life rainfall data from January to April. The bar graphs depicted raw rainfall levels (marked on a cm scale), while the pie charts represented these values as percentage proportions. 20 graduate students (11 male, age 24.4±2.1) with normal or corrected vision took part in the study.

#### 5.2.1.2 Protocol Design

The standard *n*-back design (Figure 5.2) was used over 24 blocks that completed a user session. Within each block, users were presented with a series of slides, and needed to compare the current slide *s* with the $s - n^{th}$ slide, where *n* ranged between 0–3, to make a *yes* or *no* decision. The value of *n* was

specified at the beginning of each block. 0-back required comparison of the presented information with a pre-defined value or pattern. For the *char* and *pos* visualizations, the posed question was whether the displayed letter/highlighted block matched with what was shown *n* slides before. For *pie* and *bar*, we asked whether a particular measurement (*e.g.*, rainfall in March) was greater than its counterpart *n* slides ago.

Given that bar and pie charts are designed to make use of human visuospatial ability, and to ensure that the *n*-back task with *bar* and *pie* is not trivially reduced to a value comparison as with the *char* and *pos* conditions[2], the values to be compared had to be inferred by users for *pie* and *bar*. To eliminate any display-related differences, identical colors were used to present corresponding measures in the *bar* and *pie* charts. To minimize body movements, users recorded their responses via a mouse click to choose the *yes* or *no* radio button located near the screen center. To avoid fatigue and minimize recording errors, each user session was split into two 30 minute halves with a short break in-between. Each block contained 12 slides corresponding to one of the four Viz types, and each user session comprised 288 trials (2 halves/session $\times$ 12 blocks/half $\times$ 12 slides/block).

The timeline within each block is as shown in Figure 5.2. At the beginning of each block, following an instruction slide specifying the Viz type and *n* to the user (who could press the return key to proceed when ready), a fixation cross was displayed for 500 ms to orient user focus and to record resting state brain activity. For each successive slide, the display duration was set to 2s for *char*/*pos*, and 4s for *bar*/*pie* slides. Longer display time for *bar* and *pie* was motivated by the fact that users needed to infer the measure of interest via spatial and arithmetic deductions.

Users had to record their responses within a 10 second time limit and each instance involving a user response is denoted as a *trial*. Of the 24 blocks, 12 were designed to induce *low* cognitive load (0 or 1-back), while the other 12 induced *high* load (2 and 3-back) on the VWM. Our experimental design ensured that a *char*/*pos* block was always followed by a *bar*/*pie* block and vice-versa. The number of *low*/*high* blocks as well as the number of *char*/*pos*/*bar*/*pie* blocks were set to be identical across users. Also, the order of block appearance was randomized across users. Overall, our study employed a 4×4 within-subject design involving two factors– the *n*-back level (0,1,2,3 back) and the Viz type (*char*, *pos*, *bar*, *pie*).

All users availed one practice session involving eight blocks (one *low* and one *high* mental workload block for each of the four Viz types) prior to the actual experiment. As users performed the experiment, their neural activity was recorded via the 14-channel wireless Emotiv EEG device. We did not acquire subjective user opinions regarding workload (*e.g.*, NASA-TLX [64]) as with a number of related studies, as we were purely interested in predictive analytics using EEG and not in validation of a cognitive sensing mechanism against human impressions.

[2]Comparing characters essentially involves equating symbols. The highlighted block in the *pos* condition can be equivalently encoded by the numbers 1–9.

#### 5.2.1.3 Hypotheses

Based on the experimental design, our hypotheses were as follows:

1. ***N-back is more challenging with bar and pie:*** This is because users had to infer the measure of interest in the two slides via spatial and arithmetic inference before comparing for *bar* and *pie*, while *char* and *pos* required only a symbol/spatial pattern comparison. We posited that the challenge in *n*-back for bar and pie would reflect via response times and accuracies observed for the four Viz types.

2. ***User performance will decrease for higher n-back:*** In spite of *char* and *pos* comparisons being easier than *bar* and *pie*, we nevertheless expected that (a) user performance would decrease for all Viz types with higher cognitive load (2 and 3-back), and (b) this should reflect via cognitive sensing such that EEG-based categorization of low/high mental workload should be facile irrespective if the Viz type.

3. ***Cognitive processes for the char-pos and bar-pie Viz pairs should be similar:*** Following Hypothesis 1, even if the cognitive processes corresponding to the four Viz types are dissimilar, we still expected some compatibility between the CLE models for *char* and *pos*, and those for *pie* and *bar* given task similarity.

## 5.3 User Response Analysis

We firstly analyze data compiled from *explicit* user responses to examine the impact of Viz and *n*-back type on response times (RTs) and response accuracies (RAs).

### 5.3.1 Response times

Figure 5.3(left) presents RTs for the different Viz types and varying *n*. Supporting Hypothesis 1, user responses are much faster for *char* and *pos* in the 0 and 1-back conditions, while RTs for all four Viz types become very comparable from 2-back onwards. Mean RTs for 0, 1, 2 and 3-back across all Viz types are respectively 0.72±0.15, 0.88±0.14, 1.17±0.22, 1.26±0.23 seconds. Increasing RT with

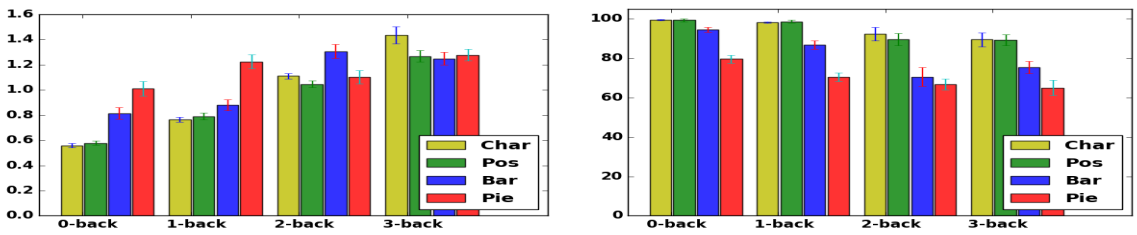


**Figure 5.3** (left) RTs (in seconds) and (right) RAs for different Viz and *n*-back types. Error bars denote unit standard error.

*n*-back type reveals that task difficulty increases with *n* due to greater load on working memory. A two-way ANOVA on RTs revealed the main effect of Viz type, *n*-back type and their interaction effect as shown in Table 5.1. A post-hoc Tukey test showed that mean reaction times differed significantly for *pos* ($\mu_{RT}$ = 0.94 s) and *pie* ($\mu_{RT}$ = 1.17 s) with $p < 0.01$, and marginally for *char* ($\mu_{RT}$ = 1 s) and *pie* with $p = 0.08$. RTs for *bar* ($\mu_{RT}$ = 1.10 s) were lower than for *pie*, but the difference was insignificant.

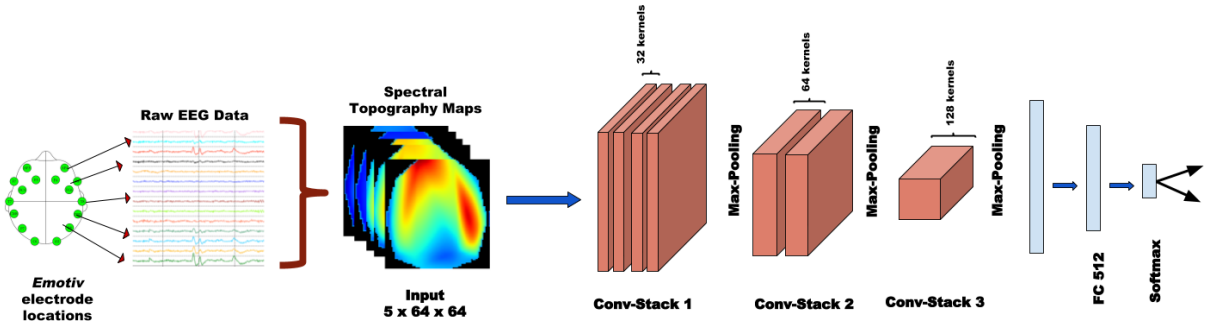


**Figure 5.4** Overview of the deep CNN architecture for cognitive load estimation.

### 5.3.2 Response accuracies

Consistent with Hypothesis 2(a), there is a steady decline in user performance with increasing *n*-back as evident from the RA plots in Figure 5.3(right). Close to ceiling performance is noted with *pos* and *char* for the *low* workload 0 and 1-back, whereas less than 80% accuracy is noted for *pie* even in these conditions. However, RAs considerably decrease across Viz types for the *high* workload 2 and 3-back, with a mean 3-back accuracy of 89.6, 89.2, 75.4 and 65.4 percent observed for *char*, *pos*, *bar* and *pie* respectively.

ANOVA on RAs revealed the main effect of Viz type and *n*-back level as shown in Table 5.1. A post-hoc Tukey test further revealed that the mean RAs for *char* ($\mu_{RA}$ = 94.9) and *pos* ($\mu_{RA}$ = 94.2) significantly differed from those for *bar* ($\mu_{RA}$ = 81.9) and *pie* ($\mu_{RA}$ = 70.6) at $p < 0.000001$. RAs for *bar* and *pie* also differed significantly with $p < 0.000001$.

### 5.3.3 Discussion of behavioral results

***User behavioral data clearly validate Hypotheses 1 and 2(a).*** The challenge posed by *n*-back with *bar* and *pie* is reflected via higher response times, and sharply lower accuracies for these Viz types. There is a steady increase in RTs, and conversely, a steady decrease in RAs for all Viz types as the mental workload increases from 0-back to 3-back. *Char* and *pos* visualizations appear to be comparable in their nature and difficulty level, and result in very similar RTs and RAs. However, bar graphs seem to be more easy to interpret than pie charts– while the response times for *bar* and *pie* only differ slightly, RAs for *bar* are significantly higher than for *pie*. Interestingly, our results agree with those of Cleveland

and Mcgill [16] who pose an estimation task, but differ with respect to the work of Peck *et al.* [72] who pose a comparison task similar to ours and find that *bar-pie* differences are not holistic but user-specific.

Given these user performance trends, we set out to examine how well they were replicated via cognitive sensing and machine learning models trained for CLE. The following sections describe two recently proposed CLE algorithms and how well they can predict mental workload across Viz types.

## 5.4 EEG-based Cognitive Load Estimation

### 5.4.1 EEG acquisition and preprocessing

We used the Emotiv EPOC headset to record EEG data as users performed the *n*-back task, and the Emotiv EEG signals are found to effectively capture memory workload in recent works [3, 13, 101]. The Emotiv EPOC is a 14-channel (plus CMS/DRL references) wireless, dry electrode device. The headset uses sequential sampling at 2048 Hz internally, which is down-sampled to 128 Hz for EEG recordings.

The recorded EEG data is contaminated by various noise sources that originate from outside the brain. For instance, one often encounters 50/60 Hz power-line noise as well as muscle or eye movement based artifacts in EEG data. For predicting cognitive load, upon removing corrupt EEG recordings we extracted *two second epochs* from the period immediately preceding user response from each trial. EEG preprocessing included (a) baseline power removal using the 0.5 second fixation samples (see Figure 5.2), (b) band-pass filtering of the EEG signal in 0.1-45 Hz frequency range, (c) independent component analysis (ICA) to remove artifacts relating to eye-blinks, eye and muscle movements. Muscle movement artifacts in EEG are mainly concentrated between 40-100 Hz. While most artifacts are removed upon EEG band-limiting, the remaining are removed via inspection of ICA components.

| Method | Raw EEG Input | Extracted Features | What is classified? | What is output? |
|---|---|---|---|---|
| **Deep CNN** | 14×2×128 samples from 2 sec pre-response epochs. | 14× 5 electrode-wise spectral measurements projected to form 5×64×64 spectral topography maps via interpolation and projection. | 512 dimensional FC layer output vector | One of two input labels as predicted by the softmax layer. |
| **pSVM** | | 14×47 electrode-wise spectral, statistical, morphological and entropy features. | Top 10 among 658 features as determined by feature selection. | One of two input labels as predicted by the pSVM classifier. |

**Table 5.2** A synopsis of various aspects pertaining to the Deep CNN and pSVM methods.

### 5.4.2 Deep CNN-based CLE

Bashivan et al [8] recently proposed cognitive load estimation using deep convolutional neural networks (deep CNNs). Deep CNNs have become very popular for solving computer vision problems like image classification [52], image segmentation, action recognition [83], *etc.* A critical factor that has contributed to the success of CNNs is that they *automatically* learn descriptors salient for the problem on hand, obviating the need for designing hand-coded features.

The key novelty of [8] is that it preserves the spatial and spectral EEG structure thereby leading to representations robust to variations and distortions within each dimension, unlike traditional approaches that vectorize EEG data. The CNN architecture employed in [8] is illustrated in Figure 5.4. The first step involves generating spectral topography maps (akin to 3D RGB color images) for five spectral bands from raw EEG data. Fast Fourier Transform (FFT) is performed on the two-second raw EEG epoch for each trial to estimate the signal power spectrum. Cortical activity relating to memory load is found to exist in the $\delta$ (0.5-4Hz), $\theta$ (4-7 Hz), $\alpha$ (8-13 Hz), $\beta$ (13-30 Hz) and $\gamma$ (30-45 Hz) bands [101]. Sum of squared-absolute values for these five frequency bands are computed and used as the electrode-specific measurement.

To generate multi-spectral maps as input to the CNN, scattered electrode power measurements are interpolated and projected onto a 2D surface via the distance-preserving Azimuthal Equidistant Projection (AEP). We generated maps of size 32×32, 64×64 and 128×128, and found that the 64×64 resolution worked best for our data. The deep CNN architecture (Figure 5.4) is very similar to the VGG network [84]. All convolutional layers use kernels of size $3\times3$, and stride of 1 pixel with ReLU (Rectified Linear Unit) activation function. The convolution layer inputs are padded with 1 pixel to preserve spatial resolution post convolution. Max-pooling is performed after every convolution over a $2\times2$ window with a stride of 2 pixels to achieve invariance to distortions. Overall, a three-fold stacking of convolutional layers with the first stack including four convolution layers with 32 kernels, second stack comprising two convolution layers with 64 kernels, and a third stack with one convolution layer of 128 kernels is used. A fully-connected (FC) layer with 512 nodes is added on top of all convolutional stacks followed by a terminal *softmax* layer, which outputs predicted test data labels.

### 5.4.3 CLE with pSVM

To demonstrate the feasibility of assessing mental workload using the Emotiv wireless headset, Wang *et al.* proposed a signal processing cum classification framework in [101]. Upon automatically removing EEG artifacts via an ICA-based algorithm[3], four types of features are extracted from the raw EEG data, namely, spectral power, statistical, morphological, and wavelet entropy. Upon extracting a total of 47 features on a per-electrode basis, all features are concatenated to form a $14\times47$ feature vector. A subset of features that have maximum relevance and minimum redundancy is identified from this set via an information theoretic approach. A proximal SVM (pSVM) algorithm that serves as a fast alternative to the standard SVM formulation is then used to predict the test class labels. A synopsis of various aspects concerning the deep CNN and pSVM algorithms is presented in Table 5.2.

### 5.4.4 Experiments and Results

A perfect benchmarking of the cognitive load induced by different visualizations would require a CLE algorithm to assign a real-value score for each Viz type– a process known as *regression* in

[3] which was not used in this work

| Type | 0-back | 1-back | 2-back | 3-back | Total |
|---|---|---|---|---|---|
| *Char* | 356 | 199 | 364 | 167 | 1086 |
| *Pos* | 377 | 175 | 165 | 376 | 1093 |
| *Bar* | 203 | 391 | 194 | 390 | 1178 |
| *Pie* | 193 | 401 | 393 | 192 | 1179 |
| **Total** | 1129 | 1166 | 1116 | 1125 | **4536** |

**Table 5.3** EEG epoch distribution based on Viz and *n*-back type.

machine learning. However, given the practical difficulties with cognitive sensing, most recent approaches [8, 13, 101] limit themselves to a coarse categorization of *low* and *high* workload levels by training a *classification* model. We also examined the efficacy of the deep CNN and pSVM models to effectively distinguish between each pair of workload levels induced by the *n*-back task on learning from the EEG epoch data. For deep CNN-based classification, we adopted the *single frame* approach where a single image (per spectral band) is used to encode the spectral power measurements over the entire epoch duration.

Specifically, we performed six (${}^4C_2$) disjoint pairwise classifications (*e.g.*, 0-back vs 1-back). Based on the behavioral results, categorizing *low* vs *high* workload (*e.g.*, 0 vs 3-back) should be easier than distinguishing between two *low* (0 vs 1-back) or *high* (2 vs 3-back) load conditions. To examine similarities among cognitive processes across Viz types, we trained a model with labeled epochs for one Viz type, and tested the same with epochs of another Viz type. Classification was performed in a user-independent setting, with epochs from all users utilized for model training. As performance metrics, we considered the classification *accuracy* and *F1-score*. F1-score denotes the harmonic mean of precision and recall, and is useful for evaluating classifier performance on unbalanced datasets given the varying EEG epoch counts (over all users) available per Viz and *n*-back condition upon removal of corrupted samples (Table 5.3).

#### 5.4.4.1 Correctness of results

We firstly remark on the validity of the classification results. Due to limited data typically available from cognitive studies, models trained thereof are prone to *overfitting*, where the trained model performs poorly on an independent test set derived identically to the training set. Overfitting commonly occurs when the number of model parameters far exceeds the training data size– we can nevertheless see from Tables 5.2 and 5.3 that the epoch samples available per *n*-back and Viz pair are fairly comparable to the number of classified features. We additionally repeated each classification experiment 10 times with randomly sampled 90:10 training-test splits of the epoch samples[4] and found that (a) the classifier performance on the training and test sets were comparable, and (b) the test accuracy/F1 standard deviation

[4]This process is known as *cross-validation*, and is commonly used for assessing the generalizability of a classifier's predictive power.

| | | char | | pos | | bar | | pie | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** |
| **char** | **0 vs 1** | 0.63 | 0.63 | 0.53 | 0.68 | 0.37 | 0.52 | 0.48 | 0.61 |
| | **0 vs 2** | 0.66 | 0.66 | 0.59 | 0.73 | 0.56 | **0.71** | 0.47 | 0.59 |
| | **0 vs 3** | 0.71 | 0.69 | 0.59 | 0.68 | 0.46 | 0.54 | 0.62 | 0.63 |
| | **1 vs 2** | 0.71 | 0.71 | 0.64 | **0.74** | 0.41 | 0.54 | 0.46 | 0.59 |
| | **1 vs 3** | 0.65 | 0.64 | 0.43 | 0.49 | 0.54 | 0.62 | 0.63 | **0.68** |
| | **2 vs 3** | 0.78 | **0.80** | 0.46 | 0.58 | 0.27 | 0.35 | 0.54 | 0.63 |
| **pos** | **0 vs 1** | 0.61 | **0.68** | 0.67 | 0.64 | 0.43 | 0.52 | 0.38 | 0.44 |
| | **0 vs 2** | 0.57 | 0.65 | 0.76 | **0.72** | 0.63 | **0.69** | 0.43 | 0.48 |
| | **0 vs 3** | 0.52 | 0.62 | 0.60 | 0.60 | 0.54 | **0.69** | 0.64 | **0.70** |
| | **1 vs 2** | 0.54 | 0.65 | 0.64 | 0.64 | 0.50 | 0.61 | 0.44 | 0.54 |
| | **1 vs 3** | 0.50 | 0.58 | 0.68 | 0.68 | 0.48 | 0.60 | 0.43 | 0.53 |
| | **2 vs 3** | 0.47 | 0.52 | 0.63 | 0.60 | 0.60 | 0.67 | 0.40 | 0.47 |
| **bar** | **0 vs 1** | 0.39 | 0.53 | 0.47 | 0.63 | 0.63 | 0.63 | 0.54 | 0.66 |
| | **0 vs 2** | 0.52 | 0.58 | 0.67 | **0.75** | 0.65 | 0.61 | 0.42 | 0.48 |
| | **0 vs 3** | 0.48 | 0.58 | 0.60 | 0.74 | 0.83 | **0.83** | 0.71 | **0.77** |
| | **1 vs 2** | 0.35 | 0.41 | 0.43 | 0.46 | 0.73 | 0.69 | 0.45 | 0.53 |
| | **1 vs 3** | 0.56 | **0.64** | 0.44 | 0.54 | 0.79 | 0.78 | 0.62 | 0.70 |
| | **2 vs 3** | 0.26 | 0.37 | 0.51 | 0.65 | 0.66 | 0.65 | 0.51 | 0.63 |
| **pie** | **0 vs 1** | 0.40 | 0.48 | 0.43 | 0.52 | 0.57 | **0.66** | 0.59 | 0.54 |
| | **0 vs 2** | 0.51 | **0.63** | 0.48 | 0.62 | 0.47 | 0.62 | 0.60 | 0.58 |
| | **0 vs 3** | 0.56 | 0.61 | 0.58 | **0.67** | 0.49 | 0.58 | 0.73 | 0.68 |
| | **1 vs 2** | 0.46 | 0.56 | 0.52 | 0.62 | 0.50 | 0.62 | 0.60 | 0.59 |
| | **1 vs 3** | 0.56 | 0.56 | 0.39 | 0.39 | 0.55 | 0.57 | 0.79 | **0.74** |
| | **2 vs 3** | 0.58 | 0.60 | 0.37 | 0.43 | 0.35 | 0.39 | 0.73 | 0.68 |

**Table 5.4** Cross-visualization CLE results achieved with the deep CNN [8]. Training data Viz-type is denote along the rows, while test data Viz type is shown along columns.

| | | char | | pos | | bar | | pie | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** |
| **char** | **0 vs 1** | 0.74 | 0.72 | 0.52 | 0.54 | 0.27 | 0.27 | 0.47 | 0.47 |
| | **0 vs 2** | 0.66 | 0.65 | 0.54 | 0.56 | 0.48 | 0.51 | 0.52 | 0.52 |
| | **0 vs 3** | 0.90 | 0.83 | 0.42 | 0.48 | 0.43 | 0.48 | 0.41 | 0.51 |
| | **1 vs 2** | 0.84 | 0.82 | 0.71 | **0.71** | 0.34 | 0.38 | 0.45 | 0.45 |
| | **1 vs 3** | 0.79 | 0.77 | 0.41 | 0.44 | 0.55 | **0.62** | 0.50 | **0.64** |
| | **2 vs 3** | 0.90 | **0.86** | 0.44 | 0.46 | 0.24 | 0.25 | 0.40 | 0.57 |
| **pos** | **0 vs 1** | 0.49 | 0.53 | 0.79 | 0.76 | 0.52 | 0.53 | 0.28 | 0.28 |
| | **0 vs 2** | 0.56 | 0.59 | 0.81 | 0.74 | 0.49 | **0.65** | 0.45 | 0.47 |
| | **0 vs 3** | 0.40 | 0.47 | 0.72 | 0.70 | 0.43 | 0.44 | 0.41 | 0.42 |
| | **1 vs 2** | 0.64 | **0.66** | 0.83 | **0.80** | 0.53 | 0.57 | 0.50 | **0.50** |
| | **1 vs 3** | 0.50 | 0.50 | 0.79 | 0.77 | 0.49 | 0.49 | 0.37 | 0.40 |
| | **2 vs 3** | 0.46 | 0.49 | 0.82 | 0.79 | 0.43 | 0.45 | 0.31 | 0.34 |
| **bar** | **0 vs 1** | 0.29 | 0.29 | 0.59 | **0.60** | 0.72 | 0.70 | 0.45 | 0.47 |
| | **0 vs 2** | 0.48 | 0.52 | 0.50 | 0.59 | 0.72 | 0.64 | 0.42 | 0.42 |
| | **0 vs 3** | 0.47 | 0.53 | 0.49 | 0.53 | 0.82 | **0.81** | 0.54 | **0.59** |
| | **1 vs 2** | 0.39 | 0.39 | 0.50 | 0.57 | 0.78 | 0.72 | 0.49 | 0.51 |
| | **1 vs 3** | 0.55 | **0.58** | 0.54 | 0.55 | 0.69 | 0.67 | 0.44 | 0.50 |
| | **2 vs 3** | 0.26 | 0.26 | 0.33 | 0.40 | 0.76 | 0.72 | 0.43 | 0.44 |
| **pie** | **0 vs 1** | 0.45 | 0.45 | 0.23 | 0.23 | 0.48 | 0.50 | 0.75 | 0.73 |
| | **0 vs 2** | 0.55 | 0.56 | 0.40 | 0.41 | 0.41 | 0.41 | 0.78 | 0.77 |
| | **0 vs 3** | 0.44 | **0.62** | 0.50 | **0.57** | 0.55 | **0.59** | 0.86 | 0.70 |
| | **1 vs 2** | 0.49 | 0.50 | 0.49 | 0.50 | 0.40 | 0.43 | 0.63 | 0.62 |
| | **1 vs 3** | 0.47 | 0.61 | 0.44 | 0.46 | 0.45 | 0.57 | 0.79 | 0.65 |
| | **2 vs 3** | 0.42 | 0.58 | 0.28 | 0.28 | 0.48 | 0.49 | 0.84 | **0.78** |

**Table 5.5** Cross-visualization CLE achieved with pSVM [101].

over the 10 repetitions was typically about 10% of the mean accuracy/F1 score, implying a fairly stable classifier performance.

#### 5.4.4.2 Discussion of results

Tables 5.4 and 5.5 respectively present cross-visualization CLE performance for a total of 16 (4×4) conditions with the deep CNN and pSVM algorithms. The mean accuracy and F1 scores achieved over 10 repetitions are tabulated. Within-visualization results are highlighted in blue along the table diagonal. Given the unbalanced number of samples per Viz and *n*-back pair, F1 is the more relevant performance metric and the highest F1 score obtained across all *n*-back categorizations for a given Viz pair is denoted in bold font. ***For both algorithms, the best F1-score in 13 of the 16 conditions corresponds to low vs high cognitive load categorization, validating Hypothesis 2(b)*** and implying that coarse-grained benchmarking of mental workload is more feasible than fine-grained differentiation with the two studied algorithms. Comparing the maximum within-Viz F1-scores achieved for deep CNN and pSVM, we find that pSVM outperforms deep CNN.

We also find that cross-Viz *n*-back categorization is inferior to within-Viz, and this is particularly true of the pSVM method. This result suggests (a) there are differences in the EEG signals, and consequently, EEG features obtained for the four Viz types, and (b) the pSVM method which performs classification from vectorized features cannot effectively deal with these differences. Cross-viz results achieved with deep CNN are relatively robust, and reflective of the fact that deep neutral networks are able to effectively learn adaptive and task-relevant features. To benchmark cross-viz CLE performance, we use the mean F1-score across all classifications, denoted as $\overline{\text{F1}}$, for a training-test Viz pair as the relevant metric. The *char*-based pSVM (within-Viz or *char-char* $\overline{\text{F1}}$ = 0.78) produces the best cross-Viz performance with *pos* (*char-pos* $\overline{\text{F1}}$ = 0.53), while the *pos*-based PSVM (*pos-pos* $\overline{\text{F1}}$ = 0.76) performs next best with *char* (*pos-char* $\overline{\text{F1}}$ = 0.54). The *bar* PSVM (*bar-bar* $\overline{\text{F1}}$ = 0.71) performs next best with *pos* (*bar-pos* $\overline{\text{F1}}$ = 0.54), while the *pie* model (*pie-pie* $\overline{\text{F1}}$ = 0.71) works best with *char* (*pie-char* $\overline{\text{F1}}$ = 0.55).

On the other hand, the deep CNN trained on *char* (*char-char* $\overline{\text{F1}}$ = 0.69) produces best cross-Viz results with *pos* (*char-pos* $\overline{\text{F1}}$ = 0.65), while the *pos* model (*pos-pos* $\overline{\text{F1}}$ = 0.65) achieves next best performance with *bar* (*pos-bar* $\overline{\text{F1}}$ = 0.63). The *bar* (*bar-bar* $\overline{\text{F1}}$ = 0.70) and *pie* (*pie-pie* $\overline{\text{F1}}$ = 0.64) models achieve best cross-viz performance with *pie* (*bar-pie* $\overline{\text{F1}}$ = 0.63) and *bar* (*pie-bar* $\overline{\text{F1}}$ = 0.57) respectively. These results ***only provide limited support to Hypothesis 3 that the cognitive processes for the char-pos and bar-pie pairs may be similar***. Nevertheless, a decrease in $\overline{\text{F1}}$ of up to 0.18 (for *bar-char*) and 0.36 (for *pos-pie*) is noted for cross-Viz CLE with the deep CNN and pSVM methods respectively, revealing the limitation of these models with respect to generalizing across Viz types. Overall, ***these results call for more research towards cross-Viz CLE benchmarking, even while conveying that available tools already hold some promise in this direction***.

## 5.5 Observations and Conclusions

The need to quantitatively evaluate visual interfaces, and the use of cognitive sensing as an effective tool to this end has captured the interest of the HCI community for quite some time. Many works have demonstrated the use of cognitive sensing for capturing mental workload corresponding to various task difficulty levels with a particular visualization. Nevertheless, as remarked in [34,44], the generalizability of these findings across users, tasks and Viz types remains a critical yet unsolved question.

In this context, this paper to our knowledge represents the first work to explore the generalizability of EEG-based CLE for the *n*-back task across Viz types. As against comparing mental workload across tasks [44], benchmarking cognitive load induced by different Viz types for the same task has profound implications for HCI and Viz-evaluation, as that would facilitate discovery of the more intuitive/interpretable Viz type(s) for the range of visual analytic tasks. Additionally, smart Viz UIs would be able to provide interventions and suggest potentially less demanding Viz alternatives on detecting a high user mental workload. Factors such as intra and inter-user variability in cognitive responses, and sensing noise nevertheless pose a huge challenge to cognition-based Viz evaluation.

Specifically, we examined the compatibility of EEG data acquired during *n*-back comparisons for the *char*, *pos*, *bar* and *pie* visualizations. Users had to perform first order comparisons ($x = y$ for *char* and *pos*, and $x > y$ for *bar* and *pie*) with all Viz types. While a straightforward comparison sufficed for *char* and *pos*, the compared values had to be inferred via spatial and arithmetic deductions for *bar* and *pie*. Behavioral results in the form of response times and accuracies revealed that the *n*-back comparisons were easier to perform with the *char* and *pos* Viz types, and became progressively difficult with *bar* and *pie*. Behavioral results revealing the ease of interpreting *bar* over *pie* charts are especially interesting given that we posed a *yes*/*no* question to users, which is presumably simpler than the size ratio estimation task posed by Cleveland and Mcgill [16] or slice size comparison to the nearest 10% required by Peck *et al.* [72].

We examined two recent CLE methods based on deep CNN [8] and pSVM [101] for their ability to generalize CLE across Viz types. Both methods showed good ability to distinguish between *high* (2 and 3-back) and *low* (0 and 1-back) mental workload levels when trained and tested with Viz-specific EEG data, with pSVM outperforming deep CNN. However, CLE ability drops for both methods in the cross-Viz condition with deep CNN achieving a more robust performance. The superiority of cross-Viz deep CNN performance can be attributed to the fact that (a) CNNs are able to efficiently learn task-specific and adaptable features in general, and (b) multi-spectral images are input to deep CNN retaining the spatial structure of EEG activations, while vectorized EEG features aggregated from all electrodes are input the pSVM.

Results from both approaches indicate the likelihood of similar cognitive processes for *char* and *pos* comparisons (for 3 out of 4 conditions, a pSVM/deep CNN model trained on one of the *char* or *pos* Viz types produces best cross-Viz performance on the other), while a possible cognitive similarity between *bar* and *pie* comparisons is indicated only by the deep CNN results. Even with data compiled from a limited set of 20 users, these results still hold promise. Another promising aspect of our study is

that our data was compiled with a portable consumer-grade wireless EEG device, which facilitates data collection on a large scale as with [81]– it is important to note here that deep CNNs tend to perform better with 'crowdsourced' data.

A limitation with EEG processing, especially with data recorded using wireless sensors with low signal-to-noise ratio, is that considerable manual supervision is required for data cleaning. As deep CNN architectures represent 'end-to-end' systems which can learn without any human intervention, a possible extension of this work involves employing deep CNNs for EEG noise removal. While the deep CNN architecture of Bashivan *et al.* [8] is inspired by VGGnet [84], more recent developments such as deep residual learning or Resnet [35] have raised the bar for computer vision tasks, and could also encode brain activations more efficiently and robustly. Other relevant network architectures for cognitive load benchmarking include hierarchical deep CNNs given the hierarchy involved in visual analytic tasks [30], and Siamese networks [48] which focus on differentiating between inputs and can learn with few training data that are generated by most brain-based studies. Our future work will explore these networks for cognitive load benchmarking.

*Chapter 6*

# Conclusions and Future Directions

In this thesis we explored several areas in the domain of cognitive vision. Each of the contribution has potential real world application. We have discussed the applications and future directions of our projects below.

In **Chapter 2** we proposed a CNN based deep neural network for eye contact detection with minimal data pre-processing. We also demonstrated how the transfer learning can improve results in case of insufficient labeled data.

- **Applications:** eye contact detection has very good applications in human-computer interactions, user-analytics , image filtering and gaze triggered photography. In for human-computer interactions, we can make smart systems who response to voice command only when the user is looking at it. We can use participants eye contact information to determine his/her interest for the contant being shown on a screen.
- **Future directions:** future work involves implementation of a seamless, real-time vision-voice system for assistive applications such as photo-capturing for the blind

In **Chapter 3** We presented AVEID video-based analytics system that was found to successfully capture various aspects of BOS employed for measuring engagement in people with dementia.

- **Future directions:** future work will focus on addressing the limitations in our current implementation, and employing additional modalities (such as verbal behavior) for measuring engagement among people with dementia.

In **chapter 4**, we propose an EEG-based fast image annotation system. Our annotation system exclusively relies on the P300 ERP signature, which is elicited upon the viewer detecting a pre-specified object class in the displayed image.Overall, our system achieves a peak F1-score of 0.88 with a 10 Hz annotation throughput.

- **Applications:** the proposed approach can be used to design various types of annotation tools which requires very less human efforts to complete the annotation task.

- **Future directions:** future work will focus on discovering and exploiting object-specific EEG signatures, and combining multiple human responses (e.g., EEG plus eye movements) for fine-grained object annotation and classification.

Finally in **chapter 5**, we investigated the feasibility of using wirelessly acquired EEG signals to assess memory workload in a well-controlled n-back task using a wireless EEG system with 14 signal channels. we explore the generalizability of EEG-based CLE for the *n*-back task across Viz types. Specifically, we examined the compatibility of EEG data acquired during *n*-back comparisons for the *char*, *pos*, *bar* and *pie* visualizations. We examined two recent CLE methods based on deep CNN [8] and pSVM [101] for their ability to generalize CLE across Viz types.

- **Applications:** our work would naturally facilitate usability evaluation of Viz UIs. We would then be able to automatically assess if Viz UI **B** is *more* or *less* usable than Viz UI **A** for a specific task when user cognitive data are available for **A** and **B**. Alternatively, a smart visual interface could replace/augment the current visualization with a potentially more interpretable one upon detecting high mental workload.

# Related Publications

1. Viral Parekh, Pin Sym Foong, Shendong Zhao and Ramanathan Subramanian. **AVEID: Automatic Video System for Measuring Engagement In Dementia**. the International Conference on Intelligent User Interfaces (IUI), 2018. (Acceptance rate: 23%)

2. Viral Parekh, Ramanathan Subramaian, Dipanjan Roy and C. V. Jawahar. **An EEG based image annotation System**.The Sixth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2017. (Best Poster Award)

3. Viral Parekh, Ramanathan Subramanian, and C. V. Jawahar. **Eye contact detection via deep neural networks**. In HCI International, pages 366-374. Springer, 2017.

---

**Under preparation**

1. Viral Parekh, Maneesh Bilalpur, C. V. Jawahar and Ramanathan Subramaian. **Investigating the generalizability of EEG-based Cognitive Load Estimation Across Visualizations**

# Bibliography


[1] Solving problems with visual analytics. *Procedia Computer Science*, 7:117 – 120, 2011.

[2] U. Ahlstrom and F. J. Friedman-Berg. Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, 36(7):623–636, 2006.

[3] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva. A user study of visualization effectiveness using eeg and cognitive load. In *Computer Graphics Forum*, volume 30, pages 791–800. Wiley Online Library, 2011.

[4] Anonymous. Ecvision: European research network for cognitive computer vision (network website).

[5] A. J. Astell, M. P. Ellis, L. Bernardi, N. Alm, R. Dye, G. Gowans, and J. Campbell. Using a touch screen computer to support relationships between people with dementia and caregivers. *Interacting with Computers*, 22(4):267–275, 2010.

[6] P. Bakliwal and C. Jawahar. Active learning based image annotation. In *NCVPRIPG*. IEEE, 2015.

[7] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[8] P. Bashivan, I. Rish, M. Yeasin, and N. Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[9] A. N. Belkacem, H. Hirose, N. Yoshimura, D. Shin, and Y. Koike. Classification of four eye directions from eeg signals for eye-movement-based communication systems. *Journal of Medical and Biological Engineering*, 34(6):581–588, 2014.

[10] D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. In *CVPR*, pages 451–458. IEEE, 2003.

[11] M. Bilalpur, S. M. Kia, M. Chawla, T. Chua, and R. Subramanian. Gender and emotion recognition with implicit user signals. In *International Conference on Multimodal Interaction*, 2017.

[12] T. S. Braver, J. D. Cohen, L. E. Nystrom, J. Jonides, E. E. Smith, and D. C. Noll. A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage*, 5(1):49 – 62, 1997.

[13] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld. Estimating workload using eeg spectral power and erps in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[14] P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. In L. E. Associates, editor, *Cognition and Instruction*, pages 292–332. Taylor & Francis, 1991.

[15] J. Chen and Q. Ji. 3d gaze estimation with a single camera without IR illumination. In *ICPR*, pages 1–4, 2008.

[16] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):pp. 531–554, 1984.

[17] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[18] M. A. Cohen, T. S. Horowitz, and J. M. Wolfe. Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, 106(14):6008–6010, 2009.

[19] J. Cohen-Mansfield, M. Dakheel-Ali, and M. S. Marx. Engagement in persons with dementia: the concept and its measurement. *The American journal of geriatric psychiatry*, 17(4):299–307, 2009.

[20] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[22] N. C. Ebner, M. Riediger, and U. Lindenberger. Faces-a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362, 2010.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[24] S. Favilla and S. Pedell. Touch screen ensemble music: collaborative interaction for older people with dementia. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, pages 481–484. ACM, 2013.

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

[26] M. Fölster, U. Hess, and K. Werheid. Facial age affects emotional expression decoding. *Frontiers in Psychology*, 5:30, 2014.

[27] P. S. Foong, S. Zhao, K. Carlson, and Z. Liu. Vita: Towards supporting volunteer interactions with long-term care residents with dementia. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6195–6207, New York, NY, USA, 2017. ACM.

[28] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *-ECCV 2012*.

[29] K. A. Funes Mora, F. Monay, and J.-M. Odobez. EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In *Eye Tracking Research and Applications*, pages 255–258, New York, NY, USA, 2014. ACM.

[30] F. L. Greitzer, C. F. Noonan, and L. Franklin. *Cognitive Foundations for Visual Analytics*. 2011.

[31] S. M. Hains and D. W. Muir. Infant sensitivity to adult eye direction. *Child development*, 67:1940–1951, 1996.

[32] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *T-PAMI*, 32(3):478–500, 2010.

[33] D. W. Hansen and A. E. Pece. Eye tracking in the wild. *CVIU*, 98(1):155–181, 2005.

[34] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[36] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *CoRR*, Mar. 2015.

[37] P. Hu and D. Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530. IEEE, 2017.

[38] W. Huang. Using eye tracking to investigate graph layout effects. In *Int'l Asia-Pacific Symposium on Visualization*, pages 97–100, 2007.

[39] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. Technical Report CMU-RI-TR-04-08, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2004.

[40] C. J.Camp, M. J. Skrajner, and G. J. Gorzelle. *Engagement in Dementia. In Assessment Scales for Advanced Dementia*. HPP, Baltimore.

[41] D. Kahneman and J. Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966.

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

[43] H. Katti, R. Subramanian, M. Kankanhalli, N. Sebe, T.-S. Chua, and K. R. Ramakrishnan. Making computers look the way we look: Exploiting visual attention for image understanding. In *ACM International Conference on Multimedia*, pages 667–670, 2010.

[44] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou, L. Zhang, and D. Ming. An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Frontiers in Human Neuroscience*, 8:703, 2014.

[45] C. Keysers, D. Xiao, P. Foldiak, and D. Perrett. The speed of sight. *J. Cognitive Neurosci.*, pages 90–101, 2001.

[46] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research. Brain research reviews*, 29(2-3):169–195, 1999.

[48] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Deep Learning Workshop (ICML)*, 2015.

[49] S. Koelstra, C. Mühl, and I. Patras. Eeg analysis for implicit tagging of video data. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[50] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. Thoughts on User Studies: Why, How, and When. *Computer Graphics and Applications*, 23(4):20–25, 2003.

[51] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *CVPR*, 2016.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105, 2012.

[53] S. R. Langton. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[54] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *arXiv preprint arXiv:1611.08024*, 2016.

[55] M. P. Lawton, K. Van Haitsma, M. Perkinson, and K. Ruckdeschel. Observed affect and quality of life in dementia: Further affirmations and problems. *Journal of mental Health and Aging*, 1999.

[56] D. Li, J. Babcock, and D. J. Parkhurst. Openeyes: a low-cost head-mounted eye-tracking solution. In *Eye tracking research & applications*, pages 95–100. ACM, 2006.

[57] D. E. Linden. The P300: where in the brain is it produced and what does it tell us? *The Neuroscientist*, 11(6):563–576, 2005.

[58] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3):169–179, 2014.

[59] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[60] E. L. Mahoney and D. F. Mahoney. Acceptance of wearable technology by people with alzheimerâĂŹs disease: Issues and accommodations. *American Journal of Alzheimer's Disease & Other Dementias®*, 25(6):527–531, 2010.

[61] A. Mihailidis, S. Blunsden, J. Boger, B. Richards, K. Zutis, L. Young, and J. Hoey. Towards the development of a technology for art therapy and dementia: Definition of needs and design constraints. *The Arts in Psychotherapy*, 37(4):293–300, 2010.

[62] E. Mohedano, G. Healy, K. McGuinness, X. Giró-i Nieto, N. E. OâĂŹConnor, and A. F. Smeaton. Improving object segmentation by using eeg signals and rapid serial visual presentation. *Multimedia tools and applications*, 74(22):10137–10159, 2015.

[63] C. H. Morimoto, A. Amir, and M. Flickner. Detecting eye position and gaze from a single camera and two light sources. In *ICPR*, pages 314–317. IEEE, 2002.

[64] W. F. Moroney, D. W. Biers, F. T. Eggemeier, and J. A. Mitchell. A comparison of two scoring procedures with the nasa task load index in a simulated flight task. In *Proceedings of the IEEE National Aerospace and Electronics Conference*, pages 734–740, 1992.

[65] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *ICML*, pages 807–814, 2010.

[66] A. Nakazawa and C. Nitschke. Point of gaze estimation through corneal surface reflection in an active illumination environment. In *ECCV*, pages 159–172. Springer, Berlin, Heidelberg, 2012.

[67] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 443–449, New York, NY, USA, 2015. ACM.

[68] A. Oliva. Gist of the scene. *Neurobiology of attention*, 696:251–256, 2005.

[69] V. Parekh, R. Subramanian, and C. V. Jawahar. *Eye Contact Detection via Deep Neural Networks*, pages 366–374. Springer International Publishing, Cham, 2017.

[70] A. Paszke, S. Chintala, R. Collobert, K. Kavukcuoglu, C. Farabet, S. Bengio, I. Melvin, J. Weston, and J. Mariethoz. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration, may 2017.

[71] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers. A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 9(1):21, Apr 2012.

[72] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[73] T. W. Picton et al. The p300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9:456–456, 1992.

[74] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S.-F. Chang, and P. Sajda. Closing the loop in cortically-coupled computer vision: a brain–computer interface for searching image databases. *Journal of neural engineering*, 8(3):036025, 2011.

[75] J. Polich. Updating P300: An integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

[76] M. Raschke, T. Blascheck, M. Richter, T. Agapkin, and T. Ertl. Visual analysis of perceptual and cognitive processes. In *Information Visualization Theory and Applications*, pages 284–291, 2014.

[77] A. Recasens*, A. Khosla*, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution.

[78] E. Ricci, J. Varadarajan, R. Subramanian, S. Rota Bulo, N. Ahuja, and O. Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *ICCV*, 2015.

[79] N. H. Riche. Beyond system logging: human logging for evaluating information visualization. In *BELIV*, 2010.

[80] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, aug 2017.

[81] P. Shenoy and D. Tan. Human-aided computing: Utilizing implicit human processing to classify images. In *SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[82] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Affect recognition in ads with application to computational advertising. In *ACM International Conference on Multimedia*, 2017.

[83] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[84] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[85] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *User Interface Software and Technology*, pages 271–280. ACM, 2013.

[86] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, M. Shah, and N. Souly. Deep learning human mind for automated visual classification. In *CVPR*, 2017.

[87] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of Vision*, 14(3):1–18, 2014.

[88] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: multimodal analysis of social attention in meetings. In *ACM international conference on Multimedia*, pages 659–662. ACM, 2010.

[89] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: Multimodal analysis of social attention in meetings. In *ACM Multimedia*, pages 659–662, 2010.

[90] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 2016.

[91] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 2017.

[92] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[93] G. Sychay, E. Chang, and K. Goh. Effective image annotation via active learning. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 209–212 vol.1, 2002.

[94] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Workshop on Applications of Computer Vision*, pages 191–195. IEEE, 2002.

[95] D. Torricelli, S. Conforto, M. Schmid, and T. DâĂŹAlessio. A neural-based remote eye gaze tracker under natural head motion. *Computer methods and programs in biomedicine*, 92(1):66–78, 2008.

[96] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[97] J. Veltman and A. Gaillard. Physiological indices of workload in a simulated flight task. *Biological psychology*, 42(3):323–342, 1996.

[98] Y. Verma and C. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*. 2012.

[99] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages I–511. IEEE, 2001.

[100] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *TIST*, 2011.

[101] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse. Using wireless eeg signals to assess memory workload in the *n*-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, 2016.

[102] V. Yashaswi and C. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *BMVC,2013*.

[103] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *CVIU*, 98(1):25–51, 2005.

[104] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[105] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. M. Automatic image annotation using group sparsity. In *CVPR*, 2010.

[106] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, pages 4511–4520. IEEE Computer Society, 2015.

[107] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *CVPR*, pages 918–923. IEEE, 2005.

[108] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *ICPR*, pages 1132–1135. IEEE, 2006.